SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Informatics, Systems and Communication

PhD program in Computer Science. Cycle XXXV

Robust Learning Methods for Imprecise Data and Cautious Inference

Surname: Campagner Name: Andrea

Registration number: 761976

Tutor: Prof. Gianluigi Ciocca

Supervisor: Prof. Davide Ciucci

Co-Supervisor: Prof. Federico Cabitza

Coordinator: Prof. Leonardo Mariani

ANNO ACCADEMICO / ACADEMIC YEAR 2021/2022

The problem with sending messages was that people responded to them, which meant one had to write more messages in reply.

Arkady Martine, *A Memory Called Empire*

Acknowledgements

In questa sezione, seppur ovviamente troppo breve per mettere giù per esteso tutti i pensieri e le sensazioni, ho cercato di esprimere sinteticamente la mia gratitudine verso le persone che mi hanno accompagnato e supportato lungo questo viaggio.

Innanzitutto, vorrei ringraziare i miei due supervisor, *Davide* e *Federico*, per avermi guidato nella mia carriera e nei miei studi di dottorato e per avermi fatto da mentori nella scoperta e nello studio dei miei interessi di ricerca e scientifici. Senza di loro, il mio dottorato non sarebbe stato così divertente ad appassionante.

Je voudrais aussi remercier *Thierry Denœux*, mon supervisor à l'Université de Technologie de Compiègne, et bien tous mes collegues dans le lab Heudiasyc qui m'ont entroduit à l'etude des fonctions des croyance et des probabilités imprécises. Gostaria também de agradecer a *Hugo Gamboa*, meu supervisor em Lisboa, e meus amigos e colegas do Instituto Frauhnofer que me ensinaram muito sobre análise de séries temporais e me acolheram na cidade que hoje considero minha segunda casa.

Arrivando alle note più personali, vorrei ringraziare tutta la mia famiglia, a partire da mio *papà* e mia *mamma*: sebbene con qualche difficoltà e discussione, se oggi penso alla singola ragione per cui, fin da piccolo, ho deciso di studiare Informatica, non posso non ricordare il tempo passato davanti ad un computer con mio papà. Similmente vorrei ringraziare i miei nonni, e soprattutto mio *nonno Franco*, per aver contribuito a far crescere in me l'amore per lo studio, la scienza e la tecnologia.

Per ultimo, ma prima per importanza, vorrei ringraziare la persona più speciale ed il mio amore, *Valentina*: durante tutti questi anni mi sei sempre stata a fianco, sei stata la prima (spesso l'unica) a sostenermi e con cui potevo confidarmi nei momenti di difficoltà e sei stata la mia principale fonte di ispirazione per la passione che metti in tutto quello che fai. Senza di te nulla di tutto questo sarebbe stato possibile: questa tesi è dedicata a te.

Abstract

The representation, quantification and proper management of uncertainty is one of the central problems in Artificial Intelligence, and particularly so in Machine Learning, in which uncertainty is intrinsically tied to the inductive nature of the learning problem. Among different forms of uncertainty, the modeling of *imprecision*, that is the problem of dealing with data or knowledge that are *imperfect* and *incomplete*, has recently attracted interest in the research community, for its theoretical and application-oriented implications on the practice and use of Machine Learning-based tools and methods.

This work focuses on the problem of dealing with imprecision in Machine Learning, from two different perspectives. On the one hand, when imprecision affects the input data to a Machine Learning pipeline, leading to the problem of *learning from imprecise data*. On the other hand, when imprecision is used a way to implement uncertainty quantification for Machine Learning methods, by allowing these latter to provide set-valued predictions, leading to so-called *cautious inference* methods. The aim of this work, then, will be to investigate theoretical as well as empirical issues related to the two above mentioned settings.

Within the context of learning from imprecise data, focus will be given on the investigation of the learning from fuzzy labels setting, both from a learning-theoretical and algorithmic point of view. Main contributions in this sense include: a learning-theoretical characterization of the hardness of learning from fuzzy labels problem; the proposal of a novel, pseudo labels-based, ensemble learning algorithm along with its theoretical study and empirical analysis, by which it is shown to provide promising results in comparison with the state-of-the-art; the application of this latter algorithm in three relevant real-world medical problems, in which imprecision occurs, respectively, due to the presence of conflicting expert opinions, the use of vague technical vocabulary, and the presence of individual variability in biochemical parameters; as well as the proposal of feature selection algorithms that may help in reducing the computational complexity of this task or limit the curse of dimensionality.

Within the context of cautious inference, focus will be given to the theoretical study of three popular cautious inference frameworks, as well as to the development of novel algorithms and approaches to further the application of cautious inference in relevant settings. Main contributions in this sense include the study of the theoretical properties of, and relationships among, decision-theoretic, selective prediction and conformal prediction methods; the proposal of novel cautious inference techniques drawing from the interaction between decision-theoretic and conformal predictions methods, and their evaluation in medical settings; as well as the study of ensemble of cautious inference models, both from an empirical point of view, as well as from a theoretical one, by which it is shown that such ensembles could be useful to improve robustness, generalization, as well as to facilitate application of cautious inference methods on multi-source and multi-modal data.

Contents

Chapter 1

Introduction

No sub-field of Computer Science has been more impactful in the recent years than Machine Learning (ML), both in the research community as well as in the industry. Indeed, nowadays, ML methods have been applied in a variety of settings, from medicine [197] to finance [95], from computer security [149] to physics [56], with an increasing number of reported success stories. From a conceptual point of view, ML can be understood as the discipline concerned with the design of algorithms enabling the (semi-)automatic extraction of *models* from data [247], as well as with the study of the computational and statistical properties of such algorithms [140, 247], in an attempt to capture implicit regularities and patterns in the data that could be used to extract actionable knowledge and gain some value.

Being grounded on inductive (or transductive) procedures, ML is intrinsically related to, and inseparable from, the notion of uncertainty [133]. Indeed, the data that is used to learn the model is usually not a complete representation of the setting of interest but rather a finite sample drawn from an unknown distribution: any ML algorithm, then, can use only such a finite sample to train a model that should ideally work well, not only on the data used to build it, but also on new data extracted from the above mentioned unknown distribution from which the data have been drawn. In this sense, the performance of a learning algorithm is inherently affected, and should be robust to, the stochastic nature of the *data generating process*. This form of uncertainty has been universally acknowledged in the traditional ML framework

since its inception [248], and most of the development in statistical and algorithmic learning theory has focused on the discovery, design and study of algorithms whose performance can be guaranteed a-priori to be *robust* to the selection of the underlying data generating distribution, i.e. on so-called distribution-free methods [7, 147, 247].

In real-world settings, however, other forms of uncertainty aside from the above mentioned stochasticity exist and can impact on the performance and robustness of any learning algorithm [42]. On the one hand, uncertainty can affect the data given as input to any learning algorithm. For example, the data can be affected by noise and errors [8, 182]; the available information could be incomplete, imprecise or otherwise reflect a lack of knowledge [131]; vagueness or ambiguity could be present in the definition of relevant characteristics of the data [41]; data can be collected and aggregated from different sources or annotators that may be in conflict with each other [27]. On the other hand, uncertainty can affect also the learning process itself as well as the output of any model obtained by means of this process. Indeed, not only uncertainty in the input data can be propagated through, and represented within, the predictions issued by a ML model [133], but also forms of uncertainty that are inherent in the learning process itself exist, such as *under-specification* [82], i.e. the inability to uniquely select a single best model, or *data shifts* [195], i.e. mismatches between the training and the deployment distributions.

Most relevantly, all of these forms of uncertainty can severely impact on the development of ML models as well as on their application and deployment in real-life decision making contexts [42, 123, 133, 141], especially in so-called critical domains such as the clinical one [4], making the need for robust approaches even more manifest than in the classical case [112]. In regard to uncertainty affecting the input to ML algorithm, even though some types of uncertainty can be handled through standard techniques such as regularization [160, 281], more generally either pre-processing uncertainty removal techniques [244] or specialized techniques and algorithms [131] are required to cope with these many faceted realizations of uncertainty that can affect the data. However, the underlying assumptions of these methods are rarely tenable [42], and they often do not offer any form of guarantee about their robustness

and soundness from a learning-theoretic perspective. On the other hand, uncertainty that is not properly and reliably accounted for can be unknowingly propagated in the output of ML models [141], undermining their deployment performance or otherwise causing such issues as automation bias [39] or detrimental algorithmic aversion [94] when they are embedded in decision support systems. Furthermore, in both cases, the presence of uncertainty can make the empirical validation of models in such situation more complex, thus highlighting the need for robust learning methods that can be guaranteed to work reliably even based on a biased or otherwise limited and partial estimate of empirical performance [21, 38, 199]. For these reasons, the issue of how to handle and communicate uncertainty in ML, as well as the development and evaluation of methods and algorithms to this purpose, has been widely investigated in the research community and has now become a blooming research field, as attested also by the increasing number of scholarly initiatives dedicated to this topic[1,2].

The aim of this work is to investigate issues and methods related to the handling of uncertainty in ML, focusing on a specific form of uncertainty called *imprecision* [53, 179]. Intuitively, imprecision refers to situations of uncertainty where the available data or knowledge are either *imperfect*, *incomplete* or *partially specified*. This form of uncertainty can arise in many natural settings and setups and can manifest itself in essentially two flavours. First, when it affects the input of the learning process (i.e. the training data) [131], in which case either the features or the target are not precisely known but are instead only partially specified. For example, these partially specified values can be expressed in terms of sets or distributions, in a very general sense, including probability distributions [11, 93, 164, 256] but also more general structures, such as possibility degrees [131, 151], belief functions [70, 193, 245] or imprecise probabilities [92, 151]. Such form of imprecision can arise due to many possible causes, namely due to a lack of knowledge about the domain of interest, as a way to reduce the annotation bottleneck problem [190], or also as a form of regularization to reduce noise-sensitiveness [151, 164]. Second, when it

[1]https://sites.google.com/view/udlworkshop2020/home

[2]https://sites.google.com/view/wuml2021/

affects the output of a model (i.e. the predictions issued by the ML model), as a form of uncertainty quantification [133] that allows to avoid issuing predictions that are at risk of being incorrect due to uncertainty in the learning process and instead resorting to *partial abstention*, i.e. allowing ML models that are able to predict sets of possible candidate labels so as to suggest a situation of indecision and thus nudge the decision makers using the ML model towards being more wary and cautious, requiring new information to arrive at a decision [141].

Thus, while in the first sense imprecision is seen as a problem of *learning from imprecise data* [92, 131], i.e. data that is not complete but only partially specified, with the aim of developing algorithms that can learn from such data despite their incompleteness as well as to study the properties of such algorithms; in the second sense imprecision is seen instead as a resource, as a way to implement *cautious inference* methods [133], i.e. ML methods that strike a trade-off between accuracy and precision [174], allowing models that are sometimes less precise and informative but, at the same time, more robust and accurate (and thus, hopefully, more useful).

The rest of this chapter will be devoted to describing in greater detail the two above mentioned problems, to reviewing the developments and state-of-the-art methodologies introduced in these settings as well as some of the most relevant gaps in the literature, and to describing the main contributions and outline of the rest of this work. More precisely, Section 1.1 will be devoted to a brief review of the basic setting of supervised ML, so as to outline how the issue of imprecision handling, in the two senses mentioned above, arises as a generalization of it. For each of the two settings, the existing state-of-the-art and some relevant research problems will be described, thus outlining the motivation and main contributions of this work. In regard to the research problems, in particular, these will appear numbered and highlighted in bold to denote that the contributions appearing in this thesis will be devoted at addressing them. Then, Section 1.2 will be devoted to describing the setting of learning from imprecise data, while Section 1.3 will instead focus on the setting of cautious inference. Then, Section 1.4 will outline the contents and contributions within the rest of this work.

1.1 Supervised Machine Learning

As mentioned in the introduction, supervised learning could be understood as the problem of constructing a model that is able to generalize to new data given a finite sample of data drawn from the same distribution [219]. Formally, one can assume the existence of a data space Z which can be represented as $Z = X \times Y$, with X being the *feature space*, i.e. the space of characteristics of instances that are relevant for their predictive value, and Y being the *target space*, i.e. the characteristics one is interested in predicting and which represent the supervision. Instances, then, are assumed to be drawn i.i.d (independent and identically distributed) from an unknown *data generating distribution* $\mathcal{D} \in \mathcal{P}(X \times Y)$, i.e. a probability measure over $X \times Y$. The learning task of supervised ML is then usually defined in relative terms, by referring to a predefined class of functions (or, hypotheses) $\mathcal{H}$, with $\forall h \in \mathcal{H}, h : X \mapsto Y$. Intuitively $\mathcal{H}$ represents the class of models from which a learning algorithm is allowed to select from: this latter can be understood as a function $A : Z^* \mapsto \mathcal{H}$, i.e. a function that accepts a finite sequence S of elements from Z (a *training set*) and returns an hypothesis $h \in \mathcal{H}$. In the most general setting where no assumption is made about A, such as when using a non-parametric learning method, $\mathcal{H}$ can be easily assumed to be the class of all measurable functions over Z.

The main goal in the supervised learning setting is then to find a model that fits well the data generating distribution. To formalize this idea, the notion of a *loss function* $l : Y \times Y \mapsto \mathbb{R}$ is introduced, where the value $l(y, h(x))$ represents the *cost* of predicting $h(x)$ when the correct target value is y. Thus, one defines the *true risk* of a function h as $L_{\mathcal{D}}(h) = \int_Z l(y, h(x))dz$, which represents the expected cost incurred by h when making predictions on instances sampled from the distribution $\mathcal{D}$. The main problem in supervised learning can then be formulated as:

Definition 1. *Find an algorithm A s.t., given any training set $S = ((x_1, y_1), ..., (x_m, y_m))$ and a class of functions $\mathcal{H}$, with probability greater than $1 - \delta$ returns an hypothesis $h \in \mathcal{H}$ s.t.*

$$|L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*)| \leq \epsilon(m, \delta), \tag{1.1}$$

where $h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ *is the function with minimal risk among* $\mathcal{H}$*, and* $\epsilon(m, \delta)$ *represents an approximation term which decreases monotonically with both* m *and* δ*.*

Thus, the above definition formalizes the intuition that the main goal in supervised learning setting is to find algorithms that can provide, based only on finite samples, models that are as good as those that would be obtained had one had access to the whole data generating distribution. Notably, however, the data generating distribution $\mathcal{D}$ is not known a priori which means that Problem 1 is an *ill-posed inverse problem* [247] since the *true risk* cannot be computed. The true risk can however be estimated as the *empirical risk*:

$$L_S(h) = \frac{1}{|S|} \sum_{(x,y) \in S} l(y, h(x)) \tag{1.2}$$

giving rise to the effective *empirical risk minimization* (ERM) [246] meta-algorithm, i.e. selecting the model that best fits the training data:

$$ERM(\mathcal{H}, S) = \arg\min_{h \in \mathcal{H}} L_S(h). \tag{1.3}$$

Surprisingly, despite its simplicity, seminal results in statistical learning theory (see e.g. [176, 248]) show that ERM is sufficient to solve Problem 1 as long as the chosen class of models $\mathcal{H}$ has low capacity, i.e. it is not able to fit arbitrarily well random data, thus providing a simple yet universal learning algorithm[3]. The above mentioned paradigm has a surprisingly broad applicability, indeed the ERM paradigm and its variations (such as regularized [234] and structural risk minimization [246]) can be used to describe common ML paradigms such as SVM [224] and deep learning [115] as well as statistical methods such as maximum likelihood [246] or Bayesian learning [230]. Furthermore variations on the ERM paradigm can also be used to handle specific forms of uncertainty, such as noisy data [8] or distribution shifts [199]. Despite this broad applicability, however, the handling of imprecision cannot

[3]Here the focus in only on the risk and sample complexity of the learning algorithm A. In general, if one also wants to take into account other computational resources, such as the time complexity, ERM may not be a satisfying solution as for many model classes ERM is NP-hard [139].

be directly realized within the standard supervised setting, as will be illustrated in the following two sections.

1.2 Learning from Imprecise Data

The handling of imprecision in the input of a ML algorithm arises as a generalization of the above mentioned supervised learning setting, as a way to model the incorporation of incomplete and imprecise data in the training process. In fact, one no longer assumes that the input is complete but rather both the features and the target supervision are allowed to be incomplete or otherwise be only partially specified [131]. More precisely, the input space is not assumed to be in the form $X \times Y$, but rather, in abstract terms, as $\mathcal{S}(X) \times \mathcal{S}(Y)$ where $\mathcal{S}(X), \mathcal{S}(Y)$ are collections of structures, over X and Y respectively, encoding partial information according to a knowledge representation formalism.

To make things more concrete and easier to understand, consider an example from the computer vision domain, in which the objective is to automatically tag images with an animal they depict. Each instance is represented as an image (i.e. an array of pixels), thus no imprecision is introduced in the feature space and therefore $\mathcal{S}(X) = X$. By contrast, since the main subject of an image could be partially occluded, the image could have been taken in bad lighting conditions or could otherwise be affected by noise, one may assume that the annotator is not always able to precisely describe which animal is depicted in the image, but rather may be partially undecided about the correct labeling. Thus, for example, an image could be tagged with the set {horse, pony, zebra}, and thus $\mathcal{S}(Y) = 2^Y$, suggesting that the animal shown on the picture is either an horse, a pony, or a zebra and, though it is not exactly known which of them, it is known that other animals (e.g., an elephant) are excluded.

More in general, a variety of formalisms for representing imprecise information about the input space have been considered in the literature [75, 89, 131]. The simplest and more restricted form of imprecision is represented by the tasks of learning with missing data [202] and semi-supervised learning [63], in which imprecision is

all-or-nothing: the available features and targets are either precisely defined and correspond to a single value, or they are completely unknown and then all relevant values are deemed equally possible and a-priori equally plausible. Both of these settings have been widely studied from the theoretical [30, 162, 216] as well as the algorithmic and empirical point of view. In the case of missing data, imputation [244], i.e. techniques to fill in missing values, and latent variables likelihood-based methods [156] are routinely applied in practical problems. In the case of semi-supervised learning, several algorithms and methods have been proposed [63]: these include theoretically justified generalizations of standard ML methods such as SVM [25] or manifold regularization [22], as well as heuristic methods such as pseudo-label learning [146] and self-supervised learning [239].

Missing data and semi-supervised learning, however, do not exhaust the scope of all possible forms of imprecision and, therefore, other more general formalisms and settings have been studied in the specialized literature. A very general framework in this sense is the case of fuzzy data [131] which encompasses, as special cases, the above mentioned missing data and semi-supervised learning settings as well as other commonly studied forms of imprecision, such as the case of set-valued data [69, 75]. The task of learning from fuzzy data and its sub-problems, in particular, have attracted interest for their potential use as a form of weakly supervised learning to avoid the annotation bottleneck of supervised learning [190], as well as for their wide-ranging occurrence in practical settings: indeed, fuzzy data can be used to model subjective information [41] or data from conflicting sources [20, 280], they can arise as a result of anonymization techniques [207, 221], they can be used to model uncertain or gradual information [42, 129] and can even be used to model noise, errors and outliers in data [152] thus in principle allowing to frame other forms of uncertainty under the perspective of learning from imprecise data.

Formally, in the general case of learning from fuzzy data, imprecision is represented by setting $\mathcal{S}(X) = [0,1]^X, \mathcal{S}(Y) = [0,1]^Y$, that is each instance is represented as a *fuzzy set* over the instance space. These fuzzy sets have an epistemic semantics and represent possibility distributions [76, 97]: only one of the possible

instantiations, i.e. the precise datasets compatible with the original imprecise one, is the correct one and the fuzzy membership degrees, then, describe their possibility, i.e. they provide an indication of the relative plausibility, according to the sensors or agents who produced or annotated the data, they indeed represent the correct instantiation. Restricted forms of fuzzy data can then be obtained by constraining the above mentioned possibility distributions: for example, the problem of learning from set-valued data can be obtained by requiring that all possibility distributions are boolean (i.e. $\mathcal{S}(X) = \{0,1\}^X, \mathcal{S}(Y) = \{0,1\}^Y$), while the cases of missing data and semi-supervised learning can be obtained by requiring, respectively, that $\mathcal{S}(X) = X \cup \{\perp\}$ and $\mathcal{S}(Y) = Y \cup \{\perp\}$, where, in both cases, $\perp$ represents that any value in the corresponding domain is a-priori possible. Going back to the imaging example introduced previously, to provide intuition about the use of fuzzy data as a way to model imprecision in ML tasks, an image could be tagged with $\{\text{horse} : 1, \text{pony} : 0.8, \text{zebra} : 0.5, \text{dog} : 0.0\}$, suggesting that the animal shown on the picture is one among $\{\text{horse}, \text{pony}, \text{zebra}\}$ and certainly not a *dog*: though it is not exactly known which of them, *horse* is deemed more plausible than *pony*, which in turn is deemed more plausible than *zebra*.

Aside from definitional and conceptual issues, the use of fuzzy data to model imprecision in the input of a ML task has remarkable implications for the learning task of learning from imprecise data. In this setting, the data is assumed to be generated i.i.d. from a *random fuzzy set* [81] $\tilde{\mathcal{D}}$ defined over $\mathcal{P}(\mathcal{S}(X) \times \mathcal{S}(Y) \times X \times Y)$, i.e. a distribution over tuples (π^X, π^Y, x, y), where π^X, π^Y are possibility distributions over X and Y respectively, (x, y) is a corresponding precise instantiation, satisfying $\pi^X(x) > 0, \pi^Y(y) > 0$. However, any learning algorithm is not given access to the above mentioned complete distribution, from which the ERM algorithm could be applied on the precise instantiations, but rather only to the imprecise instances sampled from the marginal $\tilde{\mathcal{D}} \downarrow (\mathcal{S}(X) \times \mathcal{S}(Y))$. The aim of the learning from fuzzy data problem can then be formulated as:

Definition 2. *Find A s.t., given any* imprecise *training set $\tilde{S} = ((\pi_1^X, \pi_1^Y), ..., (\pi_m^X, \pi_m^Y))$ and a class of function $\mathcal{H}$, with probability greater than $1 - \delta$ returns an hypothesis*

$h \in \mathcal{H}$ *s.t.*

$$|L_{\tilde{\mathcal{D}}}(h) - L_{\tilde{\mathcal{D}}}(h^*)| \leq \epsilon(m, \delta), \tag{1.4}$$

where $\forall h \in \mathcal{H}, L_{\tilde{\mathcal{D}}}(h) = \int l(y, h(\pi^X)) d\tilde{\mathcal{D}}$ *and* $h^* = \arg\min_{h \in \mathcal{H}} L_{\tilde{\mathcal{D}}}(h)$.

Notably, even though the learning algorithm A is given only an *imprecise* training set $\tilde{S}$, the model h given as output by A is evaluated in terms of its *true risk*, rather than a possible definition of risk over imprecise instances. A practical approach to address the task of learning from imprecise data is to understand this latter as a combination of two different tasks [131], that are, *disambiguation* of the imprecise data, i.e. finding a suitable *instantiation* of the imprecise data, a precise dataset that should be as close as possible to the real but unknown one; second, *learning*, i.e. finding a model that fits as well as possible the true disambiguated data and generalize well on new, previously unobserved, data. While the learning sub-problem is essentially equivalent, mutatis mutandis, to the one described in Definition 1 for the case of fully supervised learning, nonetheless it is easy to observe that the general problem of learning from fuzzy data is inherently more difficult than supervised learning. Indeed, not only the learner can access just a finite sample from the data generating distribution, but also the true representation of the instances in the training set is unknown to the learner. As a consequence of the former observations, the ERM approach, as well as other standard ML approaches, cannot be directly applied to solve this joint learning problem. Thus, several methods have been proposed or adapted to this setting, including:

- Instance-based methods, such as generalized nearest neighbors [26, 145] and related approaches [16, 259, 276]. These methods adapt transductive instance-based learning algorithms for supervised learning to the case of imprecise data, either by defining appropriate generalizations of the notion of distance between instances, or by adopting some variation of weight-based voting;

- Heuristic methods, such as pseudo-label learning [146] or label purification [165, 262]. The core idea underlying these methods is to train iteratively a standard supervised learning model: first, the model is trained based on a subset of

precise data, then the predictions of the model are used to augment the precise dataset and re-train the model in subsequent steps;

- Generalized risk minimization approaches, such as optimistic risk minimization [131], pessimistic risk minimization [121] and variants thereof [70, 75, 89, 132, 137]. These models start from the definition of a generalized loss function that can then be used to directly extend the ERM principle to imprecise data.

Despite the abundance of learning algorithms and models to address the problem of learning from fuzzy data, several issues and questions remain open in the specialized literature, both from the the theoretical and empirical points of view.

From the theoretical standpoint, even the very question of the *learnability* of learning from imprecise data, i.e. whether Problem 2 can be solved at all and with which resource constraints, has yet to receive a satisfying answer in all but the most basic cases. Drawing from the statistical literature concerned with censored data [44, 109] and robust inference [59, 60], previous works [38, 75, 158] have studied the learnability of the optimistic risk minimization [131] approach in the superset learning setting, i.e. the case where imprecision only affects the target labels and is given in the form of boolean possibility distributions (i.e. sets). In this setting, general, albeit distribution conditional[4], risk and sample complexity bounds have been provided. By contrast **so far no work has considered the learnability of more general types of imprecise data, as well as of other learning algorithms (P1.1)**. This gap is further exacerbated by the fact that the above mentioned results assume that the optimistic risk minimization problem can actually be solved efficiently. However, the optimization problems underlying the optimistic risk minimization approach are non-convex and non-smooth and are therefore not computationally feasible in real-world settings, thus limiting the applicability of the above mentioned results [38].

[4]Namely, the risk bound in Eq. (1.4) generally depends not only on m and δ but also on the data generating distribution $\mathcal{D}$. This property is remarkable, as in the standard supervised learning setting analogous results are usually distribution-free. Notably, as shown in [38], such a property, even though in general undesirable, is inevitable in learning from imprecise data.

Thus, results for other learning paradigms would be of interest, however such results, have not been provided in previous work.

An additional limitation in current state of the art regards the fact that most work in the learning from imprecise data setting has focused on classification (or, to a smaller degree, regression), while other tasks have so far been overlooked. A rather striking instance of this problem regards the **lack of study related to the tasks of feature selection and dimensionality reduction (P1.2)**. The importance of these tasks in the learning from imprecise data settings stems from a feature of the above mentioned theoretical results, namely their *dimensionality- dependence*: intuitively, this means that, without further assumptions, the risk of overfitting increases at least linearly with the dimensionality of the input space. While such a property by itself is not remarkable, indeed dimensionality-dependence also occurs in supervised learning unless stronger assumptions (e.g. margin assumption) are made, a consequence of it is that the availability of effective dimensionality reduction methods is of critical importance for reducing model complexity, improving generalization and hence control the so-called *curse of dimensionality*. However, limited work [16, 259, 276] has focused on this topic and the existing methods rely on strong parametric and distributional assumptions, limiting their applicability and real-world performance.

In parallel to the above mentioned gaps in the theoretical knowledge about the learning from imprecise data setting, under the empirical point of view, even though several algorithms have been proposed in the literature, few works have compared their empirical performance. Indeed, existing experimental comparisons [38, 102, 165, 262] have mainly considered prototypical evaluations of proposed learning methods against naive baselines or a limited set of competitive approaches and mostly employed relatively small and toy benchmark datasets while **no comprehensive evaluation of data analysis and learning algorithms for these tasks has been performed (P1.3)**. Also, the evaluation of the above mentioned algorithms in practical, real-world problems has rarely been considered [64, 209, 256].

1.3 Cautious Inference

In contrast to the case of imprecision handling in the input of a ML algorithm, in the output imprecision is used as a way to implement *cautious inference* [84], that is a generalization of supervised learning in which the Machine Learning (ML) models are allowed to express set-valued imprecise predictions, which can be understood as being affected by imprecision in the sense that they do not precisely refer to a single decision or prediction. The imprecise predictions allow the ML models to highlight a possible state of uncertainty, suggesting that the prediction should be discarded, that it should require further intervention from a human decision maker [46]. Therefore, such techniques have been advocated as a promising approach in the uncertainty quantification setting [133], to develop reliable ML-based decision support in so-called *decision-critical domains*, e.g. medicine [141]. Indeed, in all these settings, errors induced by ML models could have high-impact consequences. Therefore, the decision makers could accept imprecise but more reliable predictions, which could then be used either to take a decision, if the risk of doing so is deemed acceptable, or to prompt the need to collect more information in order to reduce the imprecision and foster human-in-the-loop decision-making [12, 32, 127, 157, 180].

As an example of this setting, consider a medical diagnosis scenario. Here instances are represented by medical cases which may be described by vectors in the input space X that characterize each such case with the presence or absence of certain symptoms. While the aim is to associate each symptomatic manifestation $x \in X$ with a given diagnosis $y \in Y$ to enable appropriate treatment, some manifestations could be characteristic of multiple diseases $y_1, \ldots, y_k$. If these diseases correspond to different and potentially contradictory treatment plans which may have negative consequences for the involved patients then issuing any single, precise prediction could have too large a cost. By contrast, additional medical tests could be used to perform a differential diagnosis and consequently arrive at the correct decision. Hence, a cautious inference method would issue an imprecise prediction in the form $\{y_1, \ldots, y_k\}$ to denote the above mentioned state of uncertainty.

Formally speaking, imprecision in the output of a ML algorithm can be formalized as relaxing the assumption that $\forall h \in \mathcal{H}, h : X \to Y$, and instead allowing models that are able to *partially abstain* and provide imprecise predictions, i.e. $h : X \to 2^Y$, with the aim of minimizing some generalized definition of loss $l : Y \times 2^Y \to \mathbb{R}$. Similarly to the case of learning from imprecise data, and even more markedly so, the definition of such a loss function depends on the considered cautious inference paradigm and many different cautious learning techniques have been proposed. These include:

- Decision-theoretic models [173, 181, 179], including models based on imprecise probabilities [263, 271] or belief functions [161, 166], as well as three-way decision theory [267, 47]. Decision-theoretic models directly generalize the expected utility criterion [84] by assigning an utility value to set-valued imprecise predictions [264], so that any given loss function $l : Y \times Y \to \mathbb{R}$ is extended to a generalized loss function $\tilde{l} : 2^Y \times Y \to \mathbb{R}$ which can then be used to select the decision-theoretically optimal imprecise prediction for any given instance;

- Selective prediction [110, 192], which encompasses approaches based on a generalization of supervised loss-based learning obtained by combining a standard supervised model with a model that controls when to issue an imprecise prediction (more specifically, an *abstain* prediction [261]) and jointly learning both models by means of learning rules based on version space learning [171] or other theoretically-grounded learning paradigms [74];

- Conformal prediction [7, 13, 251], a general post-hoc approach to obtain calibrated cautious inference methods starting from any supervised model, based on evaluating the similarity of any given instance with the training set and then applying ideas from non-parametric frequentist statistics [217] to obtain confidence sets or intervals around the prediction issued by the former model.

Clearly, by changing the type of expected prediction, from single-valued precise to set-valued imprecise ones, cautious inference methods entail a trade-off between different quality dimensions, that should be properly evaluated so as take into account different desirable properties:

- *Cost-sentitiveness* [99]: that is, whether an imprecise model properly takes into account information about the utilities and costs of the alternative decisions;

- *Validity* [251]: that is, whether the reduction in risk, or increased robustness, offered by the imprecise model can be analytically characterized or bounded;

- *Efficiency* [252]: that is, whether the imprecise predictions provided by the model are as *informative* as possible, i.e. its set-valued predictions are as small as possible while still preserving the two above mentioned properties.

In recent years, compared to the learning from imprecise data setting, cautious inference and thus the handling of imprecision in the output of ML model has become a more established and mature field within the uncertainty quantification literature. Indeed, all the mentioned models have been successfully employed in practical applications: ranging from drug discovery [6, 33] and protein function classification [241] to prediction of financial trends [187] and natural language processing [261]. Additionally, strong emphasis has been placed on reducing the computational complexity of these methods, which is often larger than that of standard supervised learning[5], to enable their use also in large-scale problems: such approaches include general heuristics or utility-specific algorithms to reduce the complexity of decision-theoretic methods [45, 173], inductive conformal prediction [184], as well as the generalization of standard approaches adopted in statistical learning theory to reduce the computational complexity of learning, e.g. boosting [74].

On the other hand, several gaps still remain open in regard to the theoretical study of cautious inference approaches. The first such limitation regards the fact that the study of the theoretical properties of different cautious inference methods has been mostly isolated, with scarce communication and translation of results among different approaches and research on different methods mainly focusing on different properties among the ones mentioned above. Indeed, while work centered on decision-theoretic methods has emphasized the balance among cost-sensitiveness

[5]Indeed, e.g. the worst-case computational complexity of decision-theoretic methods is exponential in the number of classes, while selective prediction is in general NP-hard.

and validity, research related to conformal prediction and selective prediction has instead mainly focused on the trade-off between validity and efficiency. Thus, an important missing step regards the definition of a **general picture drawing relationships among different approaches and their theoretical properties, establishing conditions for equivalence or comparability among different cautious inference approaches (P2.1)**.

A related theoretical, but also empirical, gap, stems from one of the main focuses in the cautious inference literature, which is the study of the so-called *validity-efficiency* trade-off, i.e. the strive in cautious inference between more precise and more accurate predictions while preserving desirable computational and data efficiency. Drawing from analogy with the bias-variance trade-off [177] in standard supervised learning and, especially, from the theoretical and empirical effectiveness of ensemble methods in addressing the bias-variance trade-off [274], the use of ensembles and combination methods in cautious inference has been recently investigated as a promising way to address the above mentioned trade-off [236], to improve the generalization of standard ensemble models by reducing the overfitting of the base models [18], as well as to increase accuracy of cautious classifiers without an excessive impact on computational complexity and data efficiency [57, 250]. Indeed, this idea has long been studied in the fields of information fusion, to enable the combination of imprecise probabilities and similar uncertainty quantification structures [14, 72, 194], and statistical inference, as a way to combine and aggregate results from multiple hypothesis tests or confidence intervals [163] in meta-analysis studies [119], as well as more recently in the study of regularization mechanisms for standard ensemble learning, for reducing overfitting and improving uncertainty quantification and robustness by using cautious inference models as the base classifiers [18, 188]. Nonetheless, compared to the study of ensemble methods in supervised ML setting, which is an established field from both the theoretical and empirical point of views [204, 213, 274], the **advantages and limitations of ensemble methods for imprecise classifiers have not yet been clarified save in idealized settings, neither from a theoretical nor from an empirical point of view (P2.2)**. In

regard to the theoretical properties of such ensemble methods, most studies so far assumed independence of the combined imprecise classifiers [236], relied on hard-to-verify or hard-to-realize assumptions about the data generating distribution or the ensemble mechanism [57], and were in general shown to fail at preserving the properties (e.g. validity) enjoyed by the cautious classifiers to be combined [153]. Similarly, in regard to the empirical point of view, limited work has been performed in regard to analyzing the effective usefulness of ensembles of cautious classifiers, with most existing work in this sense limited to evaluation on small collections of simple benchmark datasets [14, 18, 153].

A final, but practically very relevant, gap regards the ecological utility of cautious inference methods. Indeed, even though the practical application of cautious inference methods has been seriously investigated in the literature, most works have focused on the utility of such methods as techniques to improve the accuracy and robustness in solving the desired problems as compared with standard ML techniques [141, 235, 266, 272]. Nonetheless, as has been previously discussed the study and use of imprecise classifiers has been motivated in the literature as a way to enable more-informed and less risky AI-supported human decision-making [141]: however, while these methods has been improve the decision-making accuracy of the humans who interact with them [12, 32], the practical usefulness to human decision-makers equipped with this kind of support, and more in general the user-oriented impact on the socio-technical system that embeds these types of support, has rarely been evaluated in the real world [220] and the ability of imprecise classifiers to reduce or mitigate negative biases due to the use of AI, e.g. automation bias or deskilling, has not yet been evaluated.

1.4 Outline and Main Contributions

As previously mentioned, the aim of this thesis is to present a collection of contributions to the problem of handling imprecision in ML, focusing both on the problem of learning from imprecise data as well as the problem of cautious inference so as

to address the above mentioned gaps in these two research fields. In particular, as emphasized in the title of the thesis, the focus will be on *robust* learning methods, i.e. methods and algorithms which can be proved to work reliably and satisfy statistical and computational guarantees under reasonable but sufficiently general assumptions about the data. In regard to the problem of learning from imprecise data, which will be considered in Part I, the focus is mainly on the *learning from fuzzy label* setting, i.e. the setting where imprecision affects the target features and is represented in the form of a fuzzy set or possibility distribution. The decision to focus on this setting has been motivated by the fact that even though, obviously, it is more restrictive than the full learning from imprecise data one, it occurs commonly in several application scenarios, providing a generalization of common settings [190] such as partial label learning [75], semi-supervised learning [63] and learning from multi-rater observations [225], and still retains a significant degree of complexity, both from a conceptual and computational point of view. Nonetheless, despite this focus, the later chapters of this part will generalize the application of methods for learning from fuzzy labels to more general learning from imprecise data settings. In detail, the major contributions in this context are as follows:

- Chapter 2 focuses on the study of the learning from fuzzy label setting from a theoretical as well as empirical point of view, in order to address research problems **P1.1** and **P1.3**. Theoretically, the main contributions are a characterization of the learnability of this problem by analyzing two of the main learning paradigms proposed in this setting, namely the generalized risk minimization approach and instance-based methods, which will be given in Section 2.1 as well as the proposal and study of a novel pseudo-label ensemble-based method called RRL (random resampling-based learning). Empirically, the main contribution regards an extensive evaluation of learning algorithms on both synthetic as well as real datasets, showing the empirical effectiveness of RRL compared with other state-of-the-art algorithm for learning from fuzzy labels;

- Chapter 3 focuses on the problem of feature selection with imprecise data, so

as to address research problem **P1.2**. To this end, the main theoretical contribution will be the proposal of a novel feature selection approach based on the combination of Rough Set theory [185] and the generalized risk minimization paradigm, the systematic study of its computational properties in both the superset learning, in Section 3.1, and learning from fuzzy label settings, in Section 3.2, as well as the proposal of an efficient implementation of such an algorithm based on generic algorithms. From the empirical point of view, the main contribution will be an extensive benchmark analysis of its effectiveness, both in terms of accuracy and computational complexity, in comparison with the state-of-the-art method, in Section 3.3.

- Finally, Chapter 4 describes real-world applications of methods for learning from imprecise data in medical settings, focusing in particular on the application of the developed RRL algorithm and related variations, so as to address research problem **P1.3**. Section 4.1 focuses on the application of learning from fuzzy label methods in the problem of multi-rater ground truthing, that is the task of obtaining a ground truth supervision from a set of potentially conflicting labels provided by multiple annotators. By contrast, Sections 4.2 and 4.3 depart from the learning from fuzzy label setting and instead consider the more general setting of learning from fuzzy data setting, in which imprecision can also affect the features: more specifically, Section 4.2 focuses on the application of methods for learning from fuzzy data to model and manage imprecise data arising from vague medical terminology, while Section 4.3 focuses on the application of such methods to problem of modeling and managing *individual variation* in biomedical data, i.e. the intrinsic and characteristic patterns of variation pertaining to a given instance or the measurement process.

In regard to the setting of cautious inference and imprecision in the output of ML algorithms, which will be considered in Part II, the focus will be mainly on the study of *three-way decision* and *conformal prediction* methods, their relationship with other cautious inference methods and their ensembling. The focus on these two

cautious inference paradigms has been motivated by their growing popularity in the ML and data science communities, as well as due to their generality and easeness of application: indeed, both methods can be applied as general-purpose, post-hoc uncertainty quantification mechanisms [7, 45, 47]. In detail, the major contributions in this context are as follows:

- Chapter 5 focuses on the study of the foundations of cautious inference methods, studying the theoretical relationships among three-way decision, selective prediction and conformal prediction, in order to address research problem **P2.1**. Theoretically, the main contributions are characterizations of the conditions under which the three above mentioned paradigms can be considered equivalent, by studying separately the correspondence between three-way decision and selective prediction, on the one hand in Section 5.1, and three-way decision and conformal prediction, on the other hand in Section 5.2. This latter analysis, in particular, will lead to the definition of a novel non-conformity measure for conformal prediction based on three-way decision as well as the generalization of conformal prediction methods to the weakly supervised setting. These latter methods will then be applied to evaluate the ecological utility of cautious inference methods in a pilot user study based on a real-world medical problem, as an initial step toward addressing the above mentioned research problem related to the socio-technical impact of cautious inference methods;

- Chapter 6 focuses on the study of ensembling methods for cautious inference algorithms, so as to address research problem **P2.2**. First, Section 6.1 will be devoted to a large-scale experimental comparison of ensembling methods, with the aim of considering the benefits and limitations offered by ensembles of cautious inference methods. In particular, the main contribution will regard the study of the performance of such ensembles and their robustness to noise and the curse of dimensionality in comparison with standard, state-of-the-art, ensemble methods. The following section, Section 6.2, will instead address the theoretical properties of ensembles of cautious inference methods:

in particular, focusing on the conformal prediction framework, the aim will be to study conditions for validity and efficiency of such ensembles under very general conditions by adopting an information fusion perspective on ensembling where the validity and efficiency of different combination methods will be studied by means of a copula-based approach. These theoretical contributions will be complemented by an empirical contribution, where the studied ensemble methods will be evaluated in the setting of multi-variate time series classification, showing the efficacy of such ensemble methods in comparison with state-of-the-art supervised learning as well as cautious inference methods.

Finally, the Conclusion, in Chapter 7, will provide a summary on the main contributions in this thesis, as well as delineate some relevant directions for future work. A summative, graphical description of the main concepts studied in this section, as well as the of the related contributions, is given in Figure 1.1.

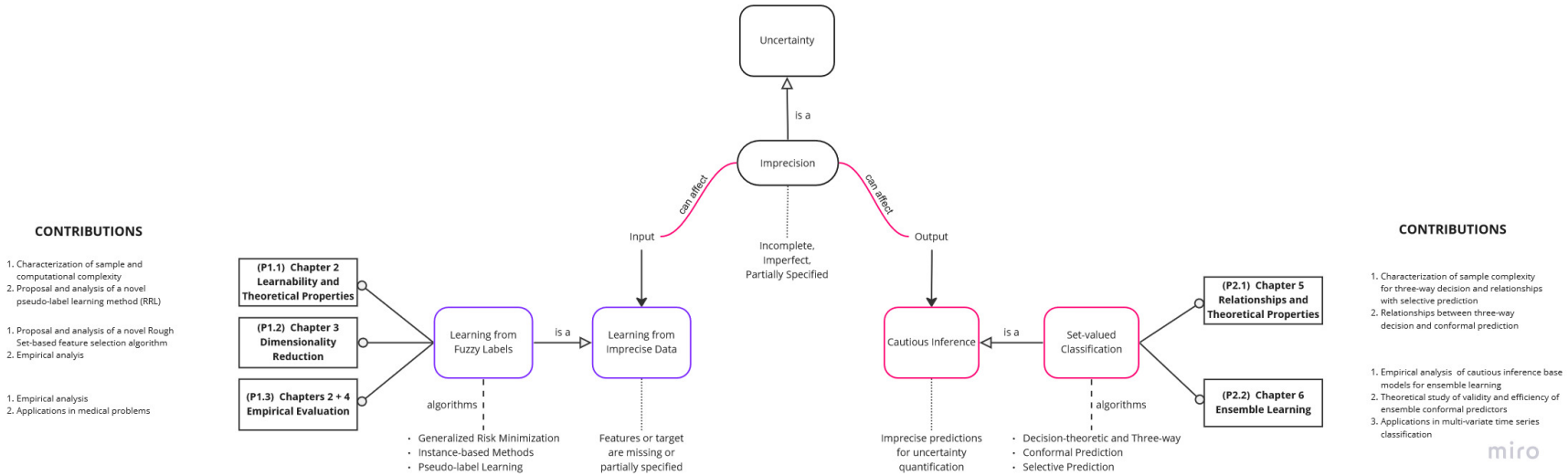


Figure 1.1: A mind map of the concepts and contributions appearing in this thesis.

1.5 List of Included Articles

To present the above mentioned main contributions this thesis will mainly take the form of a compilation thesis, by reporting the collection of published articles, which will be included in extenso, in which these contributions have originally appeared. These latter will also be complemented by novel contributions and developments that have not yet appeared in the literature, as well as by an extensive introduction and discussion of the investigated topics. The following list enumerates the included articles, in order of appearance in the thesis, along with the chapter in which they appear and the specific research problem addressed:

- Campagner, A. (2021). Learnability in "Learning from Fuzzy Labels". In 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-6). IEEE. doi: 10.1109/FUZZ45933.2021.9494534 **(Chapter 2, P1.1)**

- Campagner, A., Ciucci, D., Hüllermeier, E. (2021). Rough set-based feature selection for weakly labeled data. International Journal of Approximate Reasoning, 136, 150-167. doi: 10.1016/j.ijar.2021.06.005 **(Chapter 3, P1.2)**

- Campagner, A., Ciucci, D. (2021). Feature selection and disambiguation in learning from fuzzy labels using rough sets. In LNCS 12872, International Joint Conference on Rough Sets (pp. 164-179). Springer, Cham. doi: 10.1007/978-3-030-87334-9_14 **(Chapter 3, P1.2)**

- Campagner, A., Ciucci, D. (2022). Rough-set Based Genetic Algorithms for Weakly Supervised Feature Selection. In CCIS 1602, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 761-773). Springer, Cham. doi: 10.1007/978-3-031-08974-9_60 **(Chapter 3, P1.2)**

- Campagner, A., Ciucci, D., Svensson, C. M., Figge, M. T., Cabitza, F. (2021). Ground truthing from multi-rater labeling with three-way decision and possibility theory. Information Sciences, 545, 771-790. doi: 10.1016/j.ins.2020.09.049

(Chapter 4, P1.3)

- Seveso, A., Campagner, A., Ciucci, D., Cabitza, F. (2020). Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings. BMC Medical Informatics and Decision Making, 20(5), 1-14. doi: 10.1186/s12911-020-01152-8 **(Chapter 4, P1.3)**

- Campagner, A., Famiglini, L., Carobene, A., Cabitza, F. (2022). Everything is Varied: The Surprising Impact of Individual Variation on ML Reliability in Medicine. Under Review. **(Chapter 4, P1.3)**

- Campagner, A., Ciucci, D. (2022). Three-way Learnability: A Learning Theoretic Perspective on Three-way Decision. In Proceedings of the 17th Conference on Computer Science and Intelligence Systems (pp. 243–246). IEEE. doi: 10.15439/2022F18 **(Chapter 5, P2.1)**

- Campagner, A., Cabitza, F., Berjano, P., Ciucci, D. (2021). Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches. Information Sciences, 579, 347-367. doi: 10.1016/j.ins.2021.08.009 **(Chapter 5, P2.1)**

- Campagner, A., Cabitza, F., Ciucci, D. (2022). Aggregation Models in Ensemble Learning: a Large-Scale Comparison. Information Fusion, 90, 241-252. doi: 10.1016/j.inffus.2022.09.015 **(Chapter 6, P2.2)**

- Campagner, A., Barandas, M., Folgado, D., Gamboa, H., Cabitza, F. (2022). Evidential Predictors: Evidential Combination of Conformal Predictors for Multivariate Time Series Classification. Under Review. **(Chapter 6, P2.2)**

- Campagner, A., Lienen, J. Hüllermeier, E., Ciucci, D. (2022). scikit-weak, A Python Library for Weakly Supervised Machine Learning. In International Joint Conference on Rough Sets. To appear. **(Appendix)**

Part I

Dealing with Imprecision in the Input: Learning from Imprecise Data

The focus of the first part of this work will be on the handling of imprecision in the input of a ML task, that is the problem of learning a ML model based on an imprecise and incomplete ground truth dataset. In particular, the focus will mostly be on the problem of *learning from fuzzy label*, where only the target supervision is affected by imprecision and this latter is represented in the form of a fuzzy set encoding a possibility distribution which describes the uncertainty of the annotating agent(s).

As will be described in the following chapters, the rationale to focus on this problem is twofold: first, the problem of learning from fuzzy label represents one of the most commonly occurring settings in learning from imprecise data [131, 151]; second, compared to more general forms of imprecision (and, in particular, the general problem of learning from fuzzy data), the learning from fuzzy label problem allows to retain and focus on the main features underlying the handling of imprecise data, while limiting the computational complexity [131]. Nonetheless, the final chapters of this part will illustrate the application of methods developed in the learning from fuzzy label setting in the more general setting of fuzzy data.

In the first chapter, the main objective will be to study a theoretical characterization of the statistical and computational properties of learning from fuzzy label setting, with the aim of addressing research question **P1.1** concerning the learnability of this setting. The aim of this first chapter will then be to understand whether the learning from fuzzy label problem is feasible from a statistical point of view (i.e. is it really possible to learn from data affected by this form of imprecision?) using an approach building on PAC learning theory, one of the most popular theoretical frameworks in ML theory. In particular, results will be derived for both GRM-like algorithms, providing a generalization of previous results about the application of this methodology in the superset learning setting, as well as for instance-based methods. Aside from the theoretical study of the two above mentioned approaches, the major contribution in this chapter will be the proposal and formal analysis of a novel ensemble-based approach, called RRL (random resampling-based learning) and based on the pseudo-label learning paradigm, which will be shown to exhibit statis-

tical and computational properties that strike a balance between the two above mentioned methodologies. Building on these theoretical development, the first chapter also provides an empirical comparison of the performance of several existing learning algorithms for learning from fuzzy label, with the aim of addressing research question **P1.3**. The considered empirical analysis encompasses several state-of-the-art methods, including different implementations of the GRM and instance-based approaches as well as the above mentioned RRL algorithm, which are evaluated on a large collection of benchmark datasets, including both synthetic as well as real-world examples of learning from fuzzy label datasets.

In the second chapter, on the other hand, the focus will be on the problem of feature selection and dimensionality reduction in the learning from fuzzy label setting, with the aim of addressing research question **P1.2**. As mentioned in the introduction, this setting has been much less studied than classification, and only a single state-of-the-art algorithm, called DELIN, has so far been proposed in the literature. The main contribution will be the development of a novel feature selection method, based on Rough Set theory and the GRM paradigm, along with the study of the proposed methodology from a computational point of view, analyzing the computational complexity of the associated problems as well as proposing different algorithms based both on standard greedy heuristics as well as meta-heuristics based on evolutionary computing. Finally, the effectiveness of the proposed approach will be evaluated in comparison with the state-of-the-art both in the superset learning as well as in the learning form fuzzy label settings.

Finally, the last three chapters of the first part will be devoted at exploring applications of learning from fuzzy label in real-world problem arising from the clinical setting, thus contributing to addressing research question **P1.3**.

Chapter 2

Learning from Fuzzy Label

The *Learning from fuzzy label* problem [131] refers to a generalization of the standard supervised Machine Learning problem, subsuming also other *weakly supervised* [282] learning paradigms such as *semi-supervised* and *superset learning* [159]. In this setting, instead of assuming *precise* observations (x, y) in the input space $X \times Y$ (where X is the set of predictive features, and Y is the set of possible classification target labels), the target information is allowed to be only weakly specified as a *fuzzy set*. Thus, in the general formalism described in the Introduction, the input space can be described as $X \times \mathcal{S}(Y)$, where $\mathcal{S}(Y) = [0, 1]^Y$. These imprecise observations (x, π_Y), where π_Y is a fuzzy set over the set of labels Y, represent the uncertainty of the agent that produced the label annotations and have an associated *epistemic* semantics [79, 81]: there is an underlying true label associated with the observation which is not precisely known, but the uncertainty with respect to its true value can be represented through a *possibility distribution* [87].

In recent years, the learning from fuzzy label problem has attracted increasing interest [2, 77], both because of the relative ease of acquiring data of this type compared to other types of imprecise data [129] (data for learning from fuzzy label problem can be easily acquired in multi-rater annotation settings [50, 214], or through the use of self-labeling techniques [131, 273, 151]), to its flexibility (indeed, both semi-supervised learning and partial label learning can be understood as special cases of learning from fuzzy label) and also due to the availability of effective learning

algorithms, such as generalized risk minimization (GRM) [131, 179, 214], generalized nearest neighbor and instance-based methods (GNN) [90, 179, 232, 254] or pseudo-label methods (PL) [146].

The first aim of this chapter will be to study a theoretical characterization of this setting, from the perspective of Statistical Learning Theory (SLT), thus providing an answer to research question **P1.1**, concerning the learnability of learning from fuzzy label. To recall this theoretical question, following Definition 2, this amounts to asking whether it is in general possible to find an algorithm whose true risk can be made arbitrarily close to the optimal classifier (either relatively to a pre-defined comparison class $\mathcal{H}$, or with respect to the Bayes classifier, i.e. the measurable function with lowest absolute risk). Obviously, this is a problem of general interest in SLT. Such interest derives primarily from a characterization of the standard supervised learning, in which Vapnik and Chervonenkis [248], Valiant [243] and later Natarajan [176], provided a positive answer to the analogous question described by Definition 1 by deriving two remarkable results. First, that the generalization gap $\epsilon(m, \delta)$ (i.e. the right-side of inequality in Definition 1) for any function class $\mathcal{H}$ can be characterized as a polynomial function of the complexity of $\mathcal{H}$ itself, measured through e.g. its Natarajan dimension d [83, 176]:

$$\begin{aligned} d(\mathcal{H}) = \arg\max\{|C|, C \subseteq X : \\ \exists f_0, f_1 : C \to Y \text{ s.t. } \forall x \in C, f_0(x) \neq f_1(x) \\ \wedge (\forall B \subseteq C, \exists h \in \mathcal{H} \text{ s.t. } \forall x \in B, h(x) = f_0(x) \wedge \forall x \notin B, h(x) = f_1(x)\} \end{aligned} \tag{2.1}$$

which intuitively measures the ability of $\mathcal{H}$ to arbitrarily discriminate between pairs of classes from Y, or its Rademacher complexity [19]:

$$R(\mathcal{H}, S) = \frac{1}{|S|}\mathbb{E}_{\sigma \sim \{\pm 1\}^m}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i l(y_i, h(x_i))\right] \tag{2.2}$$

which measure the ability of $\mathcal{H}$ to fit random noise over any given training set S. Second, and most remarkable, that for any learning class $\mathcal{H}$ with bounded complexity, the above mentioned generalization gap can be achieved by simply applying a ERM

learning rule, i.e. (one of) the algorithms that minimize the empirical risk[1]

The generalization of this result to the setting of learning from fuzzy label is however non-trivial, indeed as described in the Introduction the problem of learning from fuzzy label is strictly harder than supervised learning, for two main reasons. On the one hand, any learning algorithm obviously has access to less information that can be used for optimization purposes, which usually implies that one can no longer achieve distribution-independent guarantees on the form of $\epsilon(m, \delta)$ [38]. On the other hand, the question of how to properly generalize the ERM paradigm to this setting does not have a single and straight-forward answer [132]. The most natural generalization of ERM in this sense is the generalized risk minimization (GRM) method [131]. This latter is based on the extension of a standard loss function $l : Y \times Y \to \mathbb{R}$ to a surrogate loss function over fuzzy labels $\tilde{l} : [0, 1]^Y \times Y \to \mathbb{R}$ as follows:

$$\tilde{l}(\pi, y) = \int_0^1 A(\{l(y', y) : y' \in \pi^\alpha\})d\alpha,$$

where $\pi^\alpha = \{y' \in Y : \pi(y) \geq \alpha\}$ is the α-cut of the fuzzy label π and A is an *aggregation function* specifying how to aggregate different loss function values. Then GRM is implemented by simply applying the ERM rule for $\mathcal{H}$ and the imprecise training set $\tilde{S}$ considering the loss function $\tilde{l}$. Several versions of GRM have been proposed in the literature [92], based on the selection of an appropriate aggregation function, including the average [75, 89, 137], the maximum [121, 120], the minimum [38, 131] or variants thereof [132]. While different choices of aggregation function correspond to different properties of the derived GRM rule [80, 78], the case of the minimum (usually called *optimistic risk minimization* [131] or minimin optimization [179]) has attracted particular interest in the SLT literature, due to its correspondence with the maximax estimator in statistical decision theory [78] and its appealing theoretical

[1]Notably, the actual form of this second result depends on whether one considers binary or multi-class learning tasks. In the binary case, every ERM rule achieves the above mentioned risk bounds [247]. In the multi-class case, by contrast, the result only guarantees that the above mentioned bounds can be achieved for at least one ERM rule: indeed, there exist multi-class learning problems for which some ERM rules fail [83].

properties. Indeed, different authors [158, 38] studied the application of optimistic risk minimization to the setting of superset learning (i.e. the case where the possibility distributions π are all Boolean, i.e. $\forall y \in Y, \pi(y) \in \{0, 1\}$), deriving guarantees similar to the ones for supervised learning. However, the generalization of these results to the more general setting of learning from fuzzy label has not been studied previously.

Even leaving aside the above mentioned knowledge gap, these results still do not provide a complete characterization of the theoretical landscape for the problem of learning from fuzzy label. Indeed, due to some unattractive features of optimistic risk minimization (primarily, its high computational complexity), several other methods have been proposed to address this problem, as already mentioned in the Introduction: these include, among others, instance-based methods [26, 88, 90, 91, 277, 279] and pseudo-label learning methods [101, 102, 146, 165, 262]. The first family of algorithms arises from a simple generalization of the nearest-neighbors learning rule to the setting of learning from fuzzy label:

$$GNN(S, x) = \arg\max_{y \in Y} \left\{ \sum_{(x_i, \pi_i) \in N(x)} \pi_i(y) : N(x) \text{ is a set of neighbors of } x \text{ in } S \right\} \tag{2.3}$$

Intuitively, such methods use the possibility degree of each possible class y as a weight, thus favoring classes with higher possibility degree assigned to them.

By contrast, pseudo-label methods are based on an iterative training procedure similarly to the one summarized in Algorithm 1. Intuitively, pseudo-label learning operates by iteratively selecting a precise training set from the given imprecise one $\tilde{S}$ by means of a given selection criterion (e.g. by randomly sampling precise labels, or by using only a subset of data which is already precisely labeled). Any such precise training set is then used to train a standard supervised ML algorithm by ERM, whose predictions are then used to refine the selection criterion so that new instances will be added as training data for the model at the next iteration.

Despite the practical popularity of these techniques, however, their theoretical properties have not been studied in the literature, save in strongly restricted settings

Algorithm 1 The meta-procedure for pseudo-label learning.

procedure PSEUDO_LABEL_LEARNING(h: ML model, $\tilde{S}$: imprecise dataset, C: inclusion criterion)

$S_0 \leftarrow$ select precise instances (x, y) from $\tilde{S}$

$T \leftarrow \{(x, \pi_x) \in \tilde{S} : C((x, \pi_x)) = True\}$

while $T \neq S$ **do**

Train h on S_i

$S_{i+1} \leftarrow$ refine the instances in $S_i, \tilde{S}$ based on h

$T \leftarrow \{(x, h(x)) \in S : C((x, h(x))) = True\}$

end while

return h

end procedure

[278]. Thus, as mentioned at the beginning of this section, the main aim of this chapter will be to address research problem **P1.1**. To this end, in Section 2.1, a formalization (in terms of fuzzy random sets [81, 86]) of the learning from fuzzy label setting will be provided, by which a characterization of problem instances in terms of hardness parameters is derived. These hardness parameters will be used to derive a characterization of GRM (in particular, optimistic risk minimization), in terms of sample complexity (that is, a bound on the sufficient number of instances to guarantee a required level of accuracy), and of instance-based methods (in particular, generalized nearest neighbors), in terms of expected classification error. The derived results provide a generalization of the above mentioned results for the superset learning setting. A relevant consequence of these results regards the limitations of both GRM and instance-based methods in practical applications: indeed, it is shown that instance-based methods generally exhibit exponential dependency on the dimensionality of the input space, limiting their applicability in large-scale problem, while the optimization problem underlying GRM (specifically, optimistic risk minimization) is not computationally feasible in the general case, having exponential computational complexity in the dimensionality of the feature space. To address these limita-

tions, in Section 2.2, the second aim of this chapter will be the proposal of a novel learning algorithm, called Random Resampling-based Learning (RRL) and based on the pseudo-label paradigm, whose study will provide the first learning-theoretical analysis of pseudo-label approaches in the setting of learning from imprecise data. Finally, to complement the above mentioned theoretical results, in Section 2.3, an experimental analysis of state-of-the-art learning from fuzzy label algorithms on a large benchmark, encompassing both synthetic and real-world datasets, will be discussed, providing the first large scale analysis of such techniques in terms of accuracy as well as computational complexity and thus addressing research problem **P1.3** and showing, in particular, the efficacy of the proposed RRL algorithm in comparison with the state-of-the-art.

Learnability in "Learning from Fuzzy Labels"

Andrea Campagner
DISCo, University of Milano-Bicocca
Viale Sarca 336, Milano, Italy
Email: a.campagner@campus.unimib.it

2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) | 978-1-6654-4407-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/FUZZ45933.2021.9494534

***Abstract*—Learning from fuzzy labels (LFL) refers to a generalization of supervised learning in which supervision is represented as a (epistemic) fuzzy set over the collection of possible classification labels: this represents the uncertainty of the annotating agent with respect to the true class label to be assigned to the data instances, using a possibility distribution. The two most popular LFL algorithmic approaches are either based on generalized risk minimization (GRM) or nearest-neighbor (NN) methods: while both methods have been applied successfully with promising empirical results, theoretical characterizations of these approaches, in the framework of learning theory, have been lacking. In this article we address this gap and study the LFL problem from the perspective of statistical learning theory, providing a theoretical analysis of both GRM and NN, in terms of sample complexity and risk bounds.**

***Index Terms*—Learning Theory, Fuzzy Labels, Risk Minimization, Nearest Neighbors**

I. Introduction

The *Learning from Fuzzy Labels* (LFL) problem [1] refers to a generalization of the standard supervised Machine Learning problem, subsuming also other *weakly supervised* [2] learning paradigms such as *semi-supervised* and *superset learning* [3]: in this setting, instead of assuming *precise* observations (x, y) in the input space $X \times Y$ (where X is the set of predictive features, and Y is the set of possible classification target labels), the target information is allowed to be only weakly specified as a *fuzzy set*. These fuzzy observations (x, π_Y) (where π_Y is a fuzzy set over the set of labels Y) represent the uncertainty of the agent that produced the label annotations, be it computational or human, and usually have an *epistemic* semantics [4], [5]: the true label of the observation is not precisely known, but we can represent our uncertainty with respect to its true value through a *possibility distribution*.

Thus, for example, an image could be tagged with the fuzzy set $\{\text{horse} : 1, \text{pony} : 0.8, \text{zebra} : 0.5, \text{dog} : 0.0\}$, suggesting that the animal shown on the picture is one among $\{\text{horse}, \text{pony}, \text{zebra}\}$ and, though it is not exactly known which of them, it is known that $horse$ is deemed more plausible than $pony$, which in turn is deemed more plausible than $zebra$.

In recent years, the LFL problem has attracted increasing interest [6], [7], both because of the relative ease of acquiring data of this type compared to other types of fuzzy data [8] (data for LFL problem can be easily acquired in multi-rater annotation settings [9], [10], or through the use of self-labeling techniques [1], [11]) and also due to the availability of effective learning algorithms, such as generalized risk minimization (GRM) [1], [10], nearest neighbor methods (GNN) [12]–[14] or generalized maximum likelihood (GML) [15].

In this article we study the theoretical properties of the former two classes of algorithms, from the perspective of Statistical Learning Theory (SLT): we first provide a formalization (in terms of fuzzy random sets [5], [16]) of the LFL setting and we define important parameters that can be used to characterize the learnability of LFL problems. In particular, we use these parameters to derive a characterization of GRM, in terms of sample complexity (that is, a bound on the sufficient number of instances to guarantee a required level of accuracy), and of G-NN, in terms of expected classification error. Our results provides a generalization of a similar recent result, due to Liu et al. [17], obtained in the superset learning (SSL) setting, in three directions: first, in terms of generality of the learning setting (indeed, SSL can be seen as a restricted form of LFL); second, while the results in [17] apply only to the Realizable setting our results hold also in the general agnostic case (that is, we do not assume that a zero-error classifier exists); third, we present the first analysis, to our knowledge, of nearest neighbors methods for weakly supervised learning.

II. Background

A. Possibility and Belief Function Theory

A fuzzy set (or, equivalently a possibility distribution [18]) is a function $\pi : X \mapsto [0, 1]$. In this article, we focus on *normalized* possibility distributions, that is distributions s.t. $\exists x \in X.\pi(x) = 1$. We denote with $\mathcal{F}(X)$ the collection of normalized possibility distributions over X. Given $\alpha \in [0, 1]$ we denote with $\pi^{\alpha} = \{x : \pi(x) \geq \alpha\}$ the α-cut of π, and with $\pi^{\alpha+} = \{x : \pi(x) > \alpha\}$ the strong α-cut.

We recall some basic notions from Belief Function theory (BFT): a mass function (also, random set) is defined as $m : 2^X \mapsto [0, 1]$ s.t. $\sum_{A \subseteq X} m(A) = 1$.

From a mass function, one can define three set function: namely, a *Belief function* $Bel(A) = \sum_{B:B \subseteq A} m(B)$; a *Plausibility function* $Pl(A) = \sum_{B:A \cap B \neq \emptyset} m(B)$; and a *Commonality function* $Q(A) = \sum_{B:A \subseteq B} m(B)$.

The *contour function* of m is $pl : X \mapsto [0, 1]$, defined as $pl(x) = Pl(\{x\}) = Q(\{x\})$. When the support of m is nested, then the contour function pl is a possibility distribution.

Finally we recall some basic notions about the generalization of BFT to the case of fuzzy events [5], [16]. Given X, a fuzzy random set is defined as $\tilde{m} : \mathcal{F}(X) \mapsto [0, 1]$, s.t. $\sum_{\pi \in \mathcal{F}(X)} \tilde{m}(\pi) = 1$. Belief, Plausibility and Commonality measures can be generalized to this setting by using notions

978-1-6654-4407-1/21/$31.00 © 2021 IEEE

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

from generalized measure theory: however, in the context of this paper we only provide definitions for singleton events, that is $\tilde{x} \in \mathcal{F}(X)$ s.t. $\exists! x \in X \forall x' \neq x,\ \tilde{x}(x) = 1 \wedge \tilde{x}(x') = 0$. In this case we have that

$$bel(\tilde{x}) = \sum_{\pi \in \mathcal{F}(X): \pi(x)=1 \wedge \forall x' \neq x, \pi(x')=0} \tilde{m}(\pi), \quad (1)$$

$$pl(\tilde{x}) = \sum_{\pi \in \mathcal{F}(X): \pi(x)>0} \tilde{m}(\pi), \quad (2)$$

$$q(\tilde{x}) = \sum_{\pi \in \mathcal{F}(X): \pi(x)=1} \tilde{m}(\pi). \quad (3)$$

We note that in the case of fuzzy random sets $bel(\tilde{x}) \leq q(\tilde{x}) = Q(\tilde{x}) \leq pl(\tilde{x})$ (in standard BFT the latter inequality actually holds with equality).

B. Supervised and Superset Learning

In the framework of SLT [19], data is usually assumed to be sampled i.i.d. (identically and indipendently distributed) from and unknown distribution $\mathcal{D}$ over $X \times Y$, where X is the instance space and Y is the label space. The distribution $\mathcal{D}$, and in particular its conditional $\mathcal{D}(y|x)$, encodes a functional (but in general non-deterministic) dependency between input features and target labels: our goal is to find a mapping f (or, more in general, a conditional density function) that provides us with a good approximation of $\mathcal{D}(y|x)$. Thus, we call a set of functions $\mathcal{H}$ hypothesis space, where $h \in \mathcal{H}$ is a function $h : X \mapsto Y$. The *true risk* of an hypothesis over $\mathcal{D}$ is:

$$L_{\mathcal{D}}(h) = \int_{X \times Y} l(h(x), y) d\mathcal{D}(x, y),$$

where l is a loss function, which is usually assumed to be the 0-1 loss function $l_{0-1}(y', y) = \begin{cases} 0 & y = y' \\ 1 & y \neq y' \end{cases}$

Since $\mathcal{D}$ is unknown, usually we can only access a finite sample of data, that is a training set, $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$, sampled i.i.d. from $\mathcal{D}$. The *empirical risk* of h over S is:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h(x_i), y_i).$$

Empirical Risk Minimization is any algorithm $ERM_{\mathcal{H}} : (X \times Y)^{\omega} \mapsto \mathcal{H}$ s.t. $ERM_{\mathcal{H}}(S) \in argmin_{h \in \mathcal{H}} L_S(h)$.

The fundamental theorem of multi-class learning [20], [21] tells us that we can characterize the true risk of an ERM classifier through the following result:

Theorem 1 ([20])**.** *Let $\mathcal{H}$ be an hypothesis class with Natarajan dimension d [21]. For each $\epsilon, \delta \in (0, 1)$ and distribution $\mathcal{D}$, then if $ERM_{\mathcal{H}}$ is given a dataset S of size $m \geq n_0$ with*

$$n_0 = O(\frac{d \cdot ln(|Y|) + ln(\frac{1}{\delta})}{\epsilon^2}),$$

with probability greater than $1 - \delta$, it holds that $|L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) - L_S(ERM_{\mathcal{H}}(S))| \leq \epsilon$.

Here we recall that the Natarajan dimension d of an hypothesis class $\mathcal{H}$ is a measure of the complexity of $\mathcal{H}$ [20]. Similar bounds can also be derived for non-parametric approaches such as k-Nearest Neighbors methods:

Theorem 2 ([19])**.** *Let $X = [0, 1]^d$, $Y = \{0, 1\}$, $\eta_y(x) = \mathcal{D}(y = 1|x)$ and assume that $\forall y, \eta_y$ is c-Lipschitz. Then it holds that:*

$$\mathbf{E}(L_D(k - NN(S))) = (1 + \sqrt{\frac{8}{k}}) L_{\mathcal{D}}^{Bayes} + (6c\sqrt{d} + k) m^{\frac{-1}{d+1}},$$

where the expectation is w.r.t. the sampling of a training set S of size m from $\mathcal{D}$.

In the SSL setting, the distribution $\mathcal{D}$ is defined over $X \times Y \times 2^Y$, where for an instance (x_i, y_i, C_i) the learning algorithm is not given access to the true label y_i, but only to set of candidate labels C_i. The joint distribution $\mathcal{D}$ can be decomposed as $\mathcal{D}^{x,y}$ on $X \times Y$ and the conditional $\mathcal{D}^s(x, y)$ on 2^Y given (x, y). We note that $\mathcal{D}^s(x, y)$, for each x, y, can also be seen as a mass function $m_{x,y}$.

The *superset condition* assumes that the correct label y_i is guaranteed to be in the set of possible labels: that is, we assume $Pr(y \in C) = m_{x,y}\{B \in 2^Y : y \in B\} = q_{x,y}(y) = pl_{x,y}(y) = 1$, where C is sampled from $\mathcal{D}^s(x, y)$.

In superset learning (and weakly supervised learning more in general) the empirical risk cannot be evaluated, as the learning algorithm cannot access the true labels y_i in S. The *superset risk* is defined as:

$$L_S^s(h) = \frac{1}{n} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \notin C_i}.$$

Generalized Risk Minimization is the algorithm GRM that returns (one of) the $h \in \mathcal{H}$ with minimal superset risk.

The learnability of the superset learning problem is characterized by the so-called *ambiguity degree* [22], defined as:

$$\alpha^* = sup_{(x,y) \in X \times Y} \{pl_{x,y}(l) : p(x, y) > 0, l \neq y)\}.$$

Then the true risk of the GRM algorithm can be bounded through the following result:

Theorem 3 ([17])**.** *Let $\mathcal{H}$ be an hypothesis class with Natarajan dimension d. Let $\mathcal{D}$ be a distribution, α the corresponding ambiguity degree and define $\theta = ln\frac{2}{1+\alpha^*}$. Assume that $\exists h \in \mathcal{D}. L_{\mathcal{D}}(h) = 0$ (*Realizability Assumption*). For each $\epsilon, \delta \in (0, 1)$ if $GRM_{\mathcal{H}}$ is given a dataset S of size $m \geq n_0$ with*

$$n_0 = O\left(\frac{1}{\theta\epsilon}(d \cdot ln(\frac{d|Y|^2}{\theta\epsilon}) + ln\frac{1}{\delta})\right),$$

then with probability greater than $1 - \delta$, it holds that $L_{\mathcal{D}}(GRM_{\mathcal{H}}(S)) \leq \epsilon$.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

III. Learnability of the Learning from Fuzzy Labels Problem

In the LFL setting, the probability distribution $\mathcal{D}$, from which the instances are sampled, is defined over $X \times Y \times \mathcal{F}(Y)$: each instance is a triple (x_i, y_i, π_i), where π_i is a normalized possibility distribution over Y. The interpretation of such a (generalized) instance is that all labels in π_i^{0+} are regarded as fully plausible for instance x_i and, in particular, if $\pi_i(l) \geq \pi_i(l')$, then l is considered more plausible than l'. As usual, we can decompose $\mathcal{D}$ as a probability distribution $\mathcal{D}^{x,y}$ over $X \times Y$ and a (conditional, over (x, y)) fuzzy random set $\tilde{m}_{x,y}$. The Belief, Plausibility and Commonality measures induced by $\tilde{m}_{x,y}$, for each x, y, can thus be defined as in Section II-A.

In order to evaluate the error of any hypothesis $h \in \mathcal{H}$, we first need to define the appropriate generalization of *superset risk*. We define the *fuzzy (empirical) risk* of an hypothesis h over training set S as:

$$L_S^f(h) = \frac{1}{m} \sum_{i=1}^{m} 1 - \pi_i(h(x_i)).$$

We will assume that, $\forall (x, y) \in X \times Y, p(x, y) > 0$, it holds that $pl_{x,y}(\tilde{y}) = 1$: that is, the correct label is never considered impossible. This, in turn, implies $q_{x,y}(\tilde{y}) > 0$. We call this the *weak superset assumption* and contrast it with the *strong superset assumption*: $\forall (x, y) \in X \times Y, q_{x,y}(\tilde{y}) = 1$. In this article we will not assume this stronger version as, arguably, it makes the LFL problem trivial, in the following sense:

Theorem 4. *Under the strong superset assumption and the Realizability Assumption, the learning from fuzzy labels problem is equivalent to the superset learning problem.*

Proof. Under Realizability we know that $\exists h \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h) = 0$: let h^* be one such hypothesis. Under the strong superset assumption we know that the correct label y_i is such that $\pi_i(y_i) = 1$. This implies that we can restrict $\mathcal{H}$ to $\mathcal{H}|S = \{h \in \mathcal{H} : \forall x_i \in S, h(x_i) \in \pi_i^1\}$. The result easily follows. □

Given any instance (x_i, y_i, π_i) sampled from $\mathcal{D}$, we can bound the probability that any wrong label $l \neq y_i$ would be s.t. $\pi_i(l) = 1$ (i.e. the wrong label l is considered fully possible for instance x_i):

Definition 1. *The (lower, upper)* Ambiguity Degree *is defined as:*

$$\alpha_* = inf_{(x,y)\in X\times Y}\{q_{x,y}(\tilde{l}) : p(x, y) > 0, l \neq y\},$$
$$\alpha^* = sup_{(x,y)\in X\times Y}\{q_{x,y}(\tilde{l}) : p(x, y) > 0, l \neq y\}.$$

Thus, α_* (resp. α^*) represents a lower (resp. upper) bound on the probability that an incorrect label l is such that $\pi_i(l) = 1$: if $\alpha_* = 1$ (hence, there exists l s.t. $q_{x,y}(\tilde{l}) = 1$), then we would never be able to recognize a classification error when $h(x_i) = l$. And similarly, for the correct label y_i:

Definition 2. *The (lower, upper)* Knowledge Degree *is defined as:*

$$k_* = inf_{(x,y)\in X\times Y}\{q_{x,y}(\tilde{y}) : p(x, y) > 0\},$$
$$k^* = sup_{(x,y)\in X\times Y}\{q_{x,y}(\tilde{y}) : p(x, y) > 0\}.$$

Thus, the *strong superset* condition would be equivalent to $k_* = k^* = 1$: in this case, the correct label y_i would always be included in the 1-cut of π_i. In general, however, we only have that $0 < k_* \leq k^*$ due to the weak superset assumption. We note that, for any instance (x_i, y_i, π_i), under the weak superset condition, we would be able to detect a classification error if were to observe that $h(x) = l \neq y_i$ and $\pi_i(l) = 0$. In particular, we can bound the probability of this event through what we call the *Unfalsifiability Degree*:

$$\phi_* = inf_{(x,y)\in X\times Y}\{pl_{x,y}(\tilde{l}) : p(x, y) > 0, l \neq y\},$$
$$\phi^* = sup_{(x,y)\in X\times Y}\{pl_{x,y}(\tilde{l}) : p(x, y) > 0, l \neq y\}.$$

Indeed, when $\phi_* = 1$, then we will be never be able to detect a classification error under the weak superset condition. By contrast, when $\phi^* = 0$ then we are in the standard classification setting: note that in this case we always know that for a triple (x_i, y_i, π_i) it holds that $\forall y \neq y_i \in Y, \pi_i(y) = 0$ and thus $\pi_i(y_i) = 1$ (since we require that π_i is normalized).

We also note that in the superset learning setting it holds that $\alpha_* = \phi_*, \alpha^* = \phi^*$, as in a standard random set it always hold that $q_{x,y}(l) = Q_{x,y}(\{l\}) = pl_{x,y}(l)$ (see Sections II-A,II-B). Further, due to the restriction to normalized possibility distributions, it holds that $1 - \alpha_* \leq k_*$: thus, in particular, it holds that $1 \leq \alpha^* + k^* \leq 2$.

As we will see in the following Section, the Ambiguity, Knowledge and Unfalsifiability degrees are important parameters of the distribution $\mathcal{D}$ and they characterize the learnability of a LFL problem.

A. Generalized Risk Minimization

The Generalized Risk Minimization (GRM) algorithm for the *learning from fuzzy labels* setting can be easily defined as:

$$GRM_{\mathcal{H}}^f(S) \in argmin_{h\in\mathcal{H}} L_S^f(h),$$

that is, the GRM^f algorithms always returns an hypothesis h with minimal fuzzy empirical risk. In this article we do not assume Realizability: this makes our result more general, as it also applies in the so-called *agnostic setting*, in which the data-generating distribution $\mathcal{D}$ is not necessarily deterministic. Furthermore we do not make the assumption that, for any given S, it always exists $h \in \mathcal{H}$ s.t. $L_S^f(h) = 0$: indeed, this assumption requires restrictive conditions on the generating distribution $\mathcal{D}$ to avoid the No-Free Lunch Theorem.

Thus, our goal in this Section will be to find a bound, similar to those presented in Section II-B, that applies to the LFL problem in the general agnostic setting: that is, we want to bound $Pr[|L_{\mathcal{D}}(GRM_{\mathcal{H}}^f) - L_S^f(GRM_{\mathcal{H}}^f)| > \epsilon]$.

Theorem 5. *Let $\mathcal{H}$ be an hypothesis class with Natarajan dimension d. Let $\mathcal{D}$ be the data generating distribution and*

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

$\theta_{\mathcal{D}} = log_2(\frac{2}{1+max\{\phi^*,1-k_*\}})$*, where* ϕ_*, k_* *are the respective (upper) Falsifiability and (lower) Knowledge Degrees. For each* $\epsilon, \delta \in (0,1)$*, if* $GRM^f_{\mathcal{H}}$ *is given a dataset* S *of size* $m \geq n_0$ *with*

$$n_0 = O\left(\frac{1}{(\epsilon\theta_{\mathcal{D}})^2}(d \cdot ln(\frac{d \cdot |Y|^2}{(\epsilon\theta_{\mathcal{D}})^2}) + ln(\frac{1}{\delta}))\right),$$

with probability greater than $1-\delta$*, it holds that* $|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L^f_S(GRM^f_{\mathcal{H}}(S))| \leq \epsilon$.

Proof. We start by finding a bound for

$$\mathbf{E}[|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}) - L^f_S(GRM^f_{\mathcal{H}})|]. \tag{4}$$

From this, we can obtain a bound for the quantity of interest through Markov's inequality.

First, we note that the following inequality holds:

$$\begin{aligned}&\mathbf{E}[|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}) - L^f_S(GRM^f_{\mathcal{H}})|] \leq \\ &\mathbf{E}[sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L^f_S(h)|] = \\ &\mathbf{E}[sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h) + L_S(h) - L^f_S(h)|]\end{aligned}$$

By the triangle inequality, the last expression can be upper bounded by:

$$\begin{aligned}&\mathbf{E}[sup_{h\in\mathcal{H}}|L_d(h) - L_S(h)|] + \\ &\mathbf{E}[sup_{h\in\mathcal{H}}|L_S(h) - L^f_S(h)|]\end{aligned} \tag{5}$$

The upper summand can be upper bounded as $2R(\mathcal{H}, S)$, where R is the *empirical Rademacher complexity* [19] of $\mathcal{H}$ on S, which, in turn through Massart Lemma [19] and Natarajan's Lemma [21] can be upper bounded as:

$$2\sqrt{\frac{2d(ln(m) + 2ln(|Y|))}{m}},$$

where d is the Natarajan dimension of $\mathcal{H}$.

As regards the lower summand, we note that

$$\begin{aligned}&\mathbf{E}[sup_{h\in\mathcal{H}}|L_S(h) - L^f_S(h)|] \leq \\ &\mathbf{E}[sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}|l(h(x_i), y_i) - l^f(h(x_i), \pi_i)|]\end{aligned}$$

The value of $|l(h(x_i), y_i) - l^f(h(x_i), \pi_i)|$ is bounded between 0 (when $(h(x_i) = y_i \wedge \pi_i(y_i) = 1) \vee (h(x_i) \neq y_i \wedge \pi_i(h(x_i)) = 0)$) and 1 (when $(h(x_i) = y_i \wedge \pi_i(h(x_i)) = 0) \vee (h(x_i) \neq y_i \wedge \pi_i(h(x_i)) = 1)$.

Thus the lower summand of Eq. 5 can be bounded as:

$$\begin{aligned}&\mathbf{E}[sup_{h\in H}\mathbf{1}_{h(x_i)\neq y_i}1_{\pi_i(h(x_i))>0}] + \\ &\mathbf{E}[sup_{h\in H}\mathbf{1}_{h(x_i)=y_i}1_{\pi_i(y_i)<1}] = \\ &Pr[h(x) = y \wedge \pi(y) < 1] + \\ &Pr[h(x) \neq y \wedge \pi(h(x)) > 0] \leq \\ &max\{\phi*, 1-k_*\}\end{aligned}$$

Thus, Eq. 4 can upper bounded as:

$$2\sqrt{\frac{2d(ln(m) + 2ln(|Y|))}{m}} + max\{\phi*, 1-k_*\} \leq 2\frac{\sqrt{\frac{2d(ln(m)+2ln(|Y|))}{m}}}{\theta_{\mathcal{D}}},$$

where $\theta_{\mathcal{D}} = log_2(\frac{2}{1+max\{\phi*,1-k_*\}})$. By Markov inequality, for each $\epsilon \in (0,1)$, it holds that

$$Pr[|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}) - L^f_S(GRM^f_{\mathcal{H}})| > \epsilon] \leq \frac{2\sqrt{\frac{2d(ln(m)+2ln(|Y|))}{m}}}{\epsilon\theta_{\mathcal{D}}}$$

Let $\delta \in (0,1)$, we obtain

$$\begin{aligned}&2\sqrt{\frac{2d(ln(m) + 2ln(|Y|))}{m}} \leq \epsilon\delta\theta_{\mathcal{D}} \\ &m \geq \frac{8d \cdot ln(m) + 8d \cdot ln(|Y|^2)}{(\epsilon\delta\theta_{\mathcal{D}})^2}\end{aligned}$$

A sufficient condition for this to hold is the following:

$$m \geq \frac{32d \cdot ln(\frac{64d\cdot|Y|^2}{(\epsilon\theta_{\mathcal{D}})^2}) + ln(\frac{1}{\delta})}{(\epsilon\theta_{\mathcal{D}})^2}.$$

Hence, the statement of the theorem follows. □

Thus, Theorem 5 provide an upper bound for the number of samples needed to guarantee (with high probability) that the GRM^f algorithm returns an hypothesis with a small generalization gap (i.e. an hypothesis that does not overfit). This bound, crucially, depends on two parameters that encode the uncertainty within the data generating distribution: the probability of observing a possibility degree $\pi(y') > 0$ for a wrong label (i.e. ϕ^*) and on the probability that the correct label has a possibility degree $\pi(y) = 1$ (i.e. k_*).

Comparing the bound obtained in Theorem 5 to the result by Liu et al. for superset learning (see Theorem 3), we can make the following observations:

- In our bound, there is a quadratic (inverse) dependence w.r.t. $\epsilon\theta_{\mathcal{D}}$, while in Theorem 3 this dependence is only (inverse) linear: this difference stems from the fact that our result holds for the general agnostic setting, while Theorem 3 only applies to the Realizable setting (which, as we argued in Section III, is a very restrictive assumption in the LFL setting);
- In our bound $\theta_{\mathcal{D}} = ln(\frac{2}{1+max\{\phi^*,1-k_*\}})$, while in Theorem 3 it holds that $\theta = ln\frac{2}{1+\alpha^*}$. It is easy to observe that in the superset learning setting it holds that $k_* = 1$ and $\phi^* = \alpha^*$, thus in the limit we recover the same distribution parameter used in Theorem 3. Remarkably, if we assume the strong superset assumption, it holds that $k_* = 1$ but, in general, $\alpha^* \leq \phi^*$. This implies that our bound would be more conservative than the one provided in Theorem 3: as a result of Theorem 4,

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

this difference stems from the fact that under these assumptions the arguments used through our proof are too conservative. Indeed, one can obtain a different derivation by considering only the labels $y \in Y$ s.t. $\pi(y) = 1$: in this case, in the derivation of Theorem 5 we could substitute ϕ^* with α^* and we would obtain that $\theta_{\mathcal{D}} = \theta$.

Example 1. *To provide an illustration of Theorem 5 we show a numerical example showing the dependence between the sample complexity, the parameters of the data distribution and the difference between the agnostic and Realizable settings.*

Consider a learning problem where $|Y| = 10$, the data distribution $\mathcal{D}$ is s.t. $k_* = 0.95$ (i.e., the true label is almost always s.t. $\pi(y) = 1$), while $\phi^* = 0.1$ (i.e., in one case out of ten an incorrect label is s.t. $\pi(y') > 0$) and $\alpha^* = 0.05$ (i.e., in one case out of twenty an incorrect label is s.t. $\pi(y') = 1$).

Assume that $X = \mathbb{R}^{10}$, and $\mathcal{H}$ is the class of linear multiclass predictors (i.e. $\mathcal{H} = \{x \mapsto argmax_{i \in [|Y|]}(Wx)_i : W \in R^{|Y|,10}\}$), whose Natarajan dimension is $d \leq 10$.

Assume we want to guarantee, with probability $\geq 95\%$ w.r.t. the sampling of a training set S, that the generalization gap $|L_{\mathcal{D}}(h) - L_S^f(h)| \leq 0.1$. Then, according to Theorem 5, if we apply the GRM algorithm, we should require that the training set S has sample size $m \geq 16380$. By contrast, if it were the case that $k_* = 1$ (hence, the strong superset assumption holds), it would suffice that $m \geq 13826$. If we further assume that Realizability holds (i.e. there is a linear predictor with zero risk) then, by Theorem 3, it would suffice that $m \geq 1031$.

B. Nearest Neighbors Methods

In this Section we study generalization bounds, in the form of bounds over the expected error, for Generalized Nearest Neighbor (GNN) [12], [13]: that is, techniques that generalize the standard k-NN algorithm such that for each new instance x, its classification is obtained by taking a weighted vote (where the votes are usually represented by the possibility degrees π_i of the different classes) for the k instances closest to x in S. In this article, we focus on a generalization of the 1-NN algorithm that we denote as F1-NN (Fuzzy 1-NN), and we leave the generalization to k-NN as open problem. Given an instance x and a training set S, the F1-NN algorithm finds the instance $S_1(x)$, the nearest neighbor of x in S, and returns a label chosen uniformly at random among $\pi^1_{S_1(x)}$ (i.e., the 1-cut of the possibility distribution associated with $S_1(x)$).

We will study the expected error, as a function of the size of the training set S, the data dimensionality d and parameters (see Section III) of the data generating distribution $\mathcal{D}$. In particular we show the following:

Theorem 6. *Let $X = [0,1]^d$, $Y = \{0, 1, \cdots, |Y| - 1\}$, $\forall y \in Y$, $\eta_y(x) = D(\hat{y} = y|x)$ and assume that $\forall y$ η_y is c-Lipschitz. Then:*

$$\begin{aligned}\mathbf{E}[L_{\mathcal{D}}(F1 - NN(S))] \leq (\alpha^* + 2k_* - k^*\alpha^*)|Y|L_{\mathbf{D}}^{2-Bayes} \\ + (1 + k^*\alpha^* - k_*) \\ + (1 + \alpha^* + k^*\alpha^*)4c\sqrt{d}m^{\frac{-1}{d+1}}\end{aligned} \tag{6}$$

where the expectation is w.r.t. $\mathcal{D}$ and $L_{\mathbf{D}}^{2-Bayes}$ denotes $max_y L_{\mathcal{D}|y}^{Bayes}$ with $\mathcal{D}|y$ the distribution, obtained from $\mathcal{D}$, through the standard One-vs-Rest reduction.

In order to prove this result, we first recall the following Lemma (see [19]):

Lemma 1. *Let S be a dataset of size m, and $x \in [0,1]^d$. Let $S_1(x) = argmin_{x' \in S}||x - x'||$, where $||x - x'||$ is the standard Euclidean metric. Then $E_{S,x}[||x - S_1(x)||] \leq 4\sqrt{d}m^{\frac{-1}{d+1}}$.*

Proof of Theorem 6. Let x be a new instance to be classified with true label y, S a training set and $x' = S_1(x)$ with associated possibility distribution $\pi_{x'}$ and true label y'. Then, we note that if $\pi_{x'}(y) < 1$ the F1-NN algorithm makes a classification error on x: this happens with probability no greater than $1 - \alpha_*$, or $1 - k_*$, depending on whether $y \neq y'$. Furthermore, even if $\pi_{x'}(y) = 1$, the F1-NN algorithm can make a classification error, if y is not chosen (uniformly at random) from $\pi^1_{x'}$: this happens with probability $\frac{|\pi^1_{x'}|-1}{|\pi^1_{x'}|} \leq 1$. Thus:

$$\begin{aligned}&\mathbf{E}[L_{\mathcal{D}}(F1 - NN(S))] \leq \\ &\mathbf{E}[\sum_{y \in Y} \eta_y(x)(1 - \eta_y(x'))(1 - \alpha_*) + \eta_y(x)(1 - \eta_y(x'))\alpha^* \\ &+ \eta_y(x)\eta_y(x')k^*\alpha^* + \eta_y(x)\eta_y(x')(1 - k_*)] \\ &= \mathbf{E}[\sum_{y \in Y} \eta_y(x)(1 - \eta_y(x'))(1 + \alpha^* - \alpha_*) \\ &+ \eta_y(x)\eta_y(x')(1 + k^*\alpha^* - k_*)]\end{aligned}$$

Then, it holds that:

$$\begin{aligned}\eta_y(x)\eta_y(x') &= \eta_y(x)(\eta_y(x') + \eta_y(x) - \eta_y(x)) \\ &\leq \eta_y(x)(\eta_y(x) + c||x - x'||) \\ &= \eta_y(x)^2 + \eta_y(x)c||x - x'||\end{aligned}$$

And, similarly:

$$\begin{aligned}\eta_y(x)(1 - \eta_y(x')) &= \eta_y(x)(1 - \eta_y(x') + \eta_y(x) - \eta_y(x)) \\ &\leq \eta_y(x)(1 + c||x - x'|| - \eta_y(x)) \\ &= \eta_y(x) + \eta_y(x)c||x - x'|| - \eta_y(x)^2\end{aligned}$$

Henceforth, and noting that $1 - \alpha_* \leq k_*$:

$$\begin{aligned}&\mathbf{E}[L_{\mathcal{D}}(F1 - NN(S))] \leq \\ &(k_* + \alpha^*)(1 + c\mathbf{E}[||x - x'||] - \mathbf{E}[\sum_{y \in Y} \eta_y(x)^2]) + \\ &(1 + k^*\alpha^* - k_*)(c\mathbf{E}[||x - x'|| + \mathbf{E}[\sum_{y \in Y} \eta_y(x)^2]]\end{aligned}$$

Then, we have that:

$$\begin{aligned}\mathbf{E}[\sum_{y \in Y} \eta_y(x)^2] &= \mathbf{E}[1 - \sum_{y \in Y} \eta_y(x)(1 - \eta_y(x))] \\ &\leq 1 - \sum_{y \in Y} \mathbf{E}[min\{\eta_y(x), 1 - \eta_y(x)\}] \\ &= 1 - |Y|L_{\mathbf{D}}^{2-Bayes}\end{aligned}$$

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

Therefore:

$$\begin{aligned}\mathbf{E}[L_{\mathcal{D}}(F1-NN(S))] \leq (1+\alpha^*+k^*\alpha^*)c\mathbf{E}[||x-x'||] \\ + (\alpha^*+2k_*-k^*\alpha^*)|Y|L_{\mathbf{D}}^{2-Bayes} \\ + (1+k^*\alpha^*-k_*)\end{aligned}$$

The Theorem then follows from Lemma 1.

□

Thus, the generalization error of F1-NN is proportional with the (upper) Ambiguity Degree α^* and inverse proportional with the Knowledge Degrees k_*, k^* of the generating distribution $\mathcal{D}$. In the worst case, where $\alpha^* = 1$ and $k_* = k^* \sim 0$, the bound given in Theorem 6 becomes vacuous: the second term in Equation 6 is ~ 1. On the other hand, if $k_* = k^* = 1$ we obtain a bound for the superset learning setting:

$$\mathbf{E}[L_{\mathcal{D}}(\text{F1-NN})] \leq 2|Y|L_{\mathcal{D}}^{2-Bayes}+(1+2\alpha^*)4c\sqrt{d}m^{\frac{-1}{d+1}}+\alpha^*.$$

From this latter bound, in the best case, we have that $\alpha^* = 0$ and Equation 6 simplifies to:

$$\mathbf{E}[L_{\mathcal{D}}(\text{F1-NN})] \leq 2|Y|L_{\mathcal{D}}^{2-Bayes} + 4c\sqrt{d}m^{\frac{-1}{d+1}},$$

which can be seen as a generalization of Theorem 2 to the multi-class case.

We remark that, the bound in Theorem 6 is exponential in the number of features d, thus the greater the number of features, the larger the expected risk of the F1-NN algorithm. This observation highlights the necessity of developing feature selection and dimensionality reduction techniques that could effectively be applied in the superset and LFL settings.

IV. Conclusion

In this article we provide the first learning-theoretic study, to our knowledge, of the *learning from fuzzy labels* problem: we studied two general-purpose LFL algorithm schemes (that is, GRM and GNN methods) from the perspective of SLT, and provided sufficient conditions for their effective learnability, expressed as, respectively, a sample complexity bound and an expected risk bound. These results, in turn, could be useful also for real-world applications: to inform model selection [23], to design effective learning algorithms (e.g. boosting), or to perform regularization through label smoothing [24]. The application of our results in these settings should be investigated, to analyze their applicability in real-world problems. Furthermore, in light of our results, the following open problems could be worthy of further research:

- Our results depend on the data generating distribution $\mathcal{D}$. While this distribution is unknown, it can be estimated conditional on some assumptions: e.g., $\mathcal{D}$ has a specific *parametric form* (in which case the GRM problem is equivalent to GML [15]); or the fuzzy labels are generated through self- [11] or multi-rater [6], [9] labeling. It would be thus interesting to consider instantiations of Theorems 5, 6 for such specific cases;
- Theorem 5 provides an upper bound to the *sample complexity* of any hypothesis class $\mathcal{H}$ for the LFL problem: it would be interesting to find a matching lower bound;
- The expected risk bound for G-NN in Theorem 6 applies specifically to the case where $k = 1$: it would be interesting to generalize this bound also to the case $k > 1$;

References

[1] E. Hüllermeier, "Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization," *Int J Approx Reason*, vol. 55, no. 7, pp. 1519–1534, 2014.

[2] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Natl Sci Rev*, vol. 5, no. 1, pp. 44–53, 2018.

[3] L. Liu and T. G. Dietterich, "A conditional multinomial mixture model for superset label learning," in *Adv Neural Inform Process Syst*, 2012, pp. 548–556.

[4] I. Couso and D. Dubois, "Statistical reasoning with set-valued information: Ontic vs. epistemic views," *Int J Approx Reason*, vol. 55, no. 7, pp. 1502–1518, 2014.

[5] I. Couso, D. Dubois, and L. Sánchez, "Random sets and random fuzzy sets as ill-perceived random variables," *SpringerBriefs Computat Intell*, 2014.

[6] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowl Based Syst*, p. 106771, 2021.

[7] I. Couso, C. Borgelt, E. Hullermeier, and R. Kruse, "Fuzzy sets in data analysis: From statistical foundations to machine learning," *IEEE Comput Intell Mag*, vol. 14, no. 1, pp. 31–44, 2019.

[8] E. Hüllermeier, "Does machine learning need fuzzy logic?" *Fuzzy Sets Syst*, vol. 281, pp. 292–299, 2015.

[9] A. Campagner, D. Ciucci, C.-M. Svensson, M. T. Figge, and F. Cabitza, "Ground truthing from multi-rater labeling with three-way decision and possibility theory," *Inf Sci*, vol. 545, pp. 771–790, 2020.

[10] L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, and R. Koch, "Beyond cats and dogs: Semi-supervised classification of fuzzy labels with overclustering," *arXiv preprint arXiv:2012.01768*, 2020.

[11] M. M. El-Zahhar and N. F. El-Gayar, "A semi-supervised learning approach for soft labeled data," in *Proceedings of ISDA 2010*. IEEE, 2010, pp. 1136–1141.

[12] J. Derrac, S. García, and F. Herrera, "Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects," *Inf Sci*, vol. 260, pp. 98–119, 2014.

[13] C. Thiel, "Classification on soft labels is robust against label noise," in *Proceedings of KES-2008*. Springer, 2008, pp. 65–73.

[14] N. Wagner, V. Antoine, J. Koko, and R. Lardy, "Fuzzy k-nn based classifiers for time series with soft labels," in *Proceedings of IPMU 2020*. Springer, 2020, pp. 578–589.

[15] T. Denoeux, "Maximum likelihood estimation from fuzzy data using the em algorithm," *Fuzzy Sets Syst*, vol. 183, no. 1, pp. 72–91, 2011.

[16] T. Denœux, "Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence," *Fuzzy Sets Syst*, 2020.

[17] L. Liu and T. Dietterich, "Learnability of the superset label learning problem," in *Proceedings of ICML 2014*, 2014, pp. 1629–1637.

[18] D. Dubois and H. Prade, "Possibility theory: qualitative and quantitative aspects," in *Quantified representation of uncertainty and imprecision*. Springer, 1998, pp. 169–226.

[19] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, "Multiclass learnability and the erm principle," in *Proceedings of COLT 2011*. JMLR Workshop and Conference Proceedings, 2011, pp. 207–232.

[21] B. K. Natarajan, "On learning sets and functions," *Mach Learn*, vol. 4, no. 1, pp. 67–97, 1989.

[22] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *The Journal of Machine Learning Research*, vol. 12, pp. 1501–1536, 2011.

[23] V. Cherkassky, "Model complexity control and statistical learning theory," *Nat Comput*, vol. 1, no. 1, pp. 109–133, 2002.

[24] J. Lienen and E. Hüllermeier, "Instance weighting through data imprecisiation," *International Journal of Approximate Reasoning*, 2021.

Authorized licensed use limited to: UNIVERSITA BICOCCA MILANO. Downloaded on June 28,2022 at 08:17:15 UTC from IEEE Xplore. Restrictions apply.

2.2 Pseudo-label Learning

In the previous section, the theoretical properties of optimistic risk minimization and instance-based methods have been studied, providing a partially positive answer for the learnability problem for learning from fuzzy label. Indeed, for the case of instance-based methods, this derives directly from Theorem 6 which shows that, conditional on the ambiguity and knowledge parameters being not too large for the instance problem at hand, the expected risk for such methods converges to that of the Bayes classifier. Similarly, for the case of optimistic risk minimization this result follows from the following corollary:

Corollary 1. *Let $\mathcal{H}$ be a hypothesis class with Natarajan dimension d. Let $\mathcal{D}$ be the data generating distribution and $\theta_{\mathcal{D}} = log_2(\frac{2}{1+max\{\phi^*,1-k_*\}})$, where ϕ_*, k_* are the respective (upper) Falsifiability and (lower) Knowledge Degrees. Let h^* be the classifier with minimal risk in $\mathcal{H}$, then with probability greater than $1-\delta$, when given a training set S of size m, it holds that:*

$$|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*)| \leq \epsilon(m, \delta, \theta_{\mathcal{D}}), \tag{2.4}$$

where $\epsilon(m, \delta, \theta_{\mathcal{D}})$ is poly-logarithmically decreasing in m, δ and $\theta_{\mathcal{D}}$.

Proof. By Theorem 5 in Section 2.1:

$$\begin{aligned}
|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*)| = \\
|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L^f_S(GRM^f_{\mathcal{H}}(S)) + L^f_S(GRM^f_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*)| \leq \\
|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L^f_S(GRM^f_{\mathcal{H}}(S)) + L^f_S(h^*) - L_{\mathcal{D}}(h^*)| \leq \\
|L_{\mathcal{D}}(GRM^f_{\mathcal{H}}(S)) - L^f_S(GRM^f_{\mathcal{H}}(S))| + |L_{\mathcal{D}}(h^*) - L^f_S(h^*)| \leq \\
\epsilon(m, \delta, \theta_{\mathcal{D}})
\end{aligned}$$

where, following Theorem 5 in Section 2.1, $\epsilon(m, \delta, \theta_{\mathcal{D}}) = O(\sqrt{\frac{d\ln(\frac{d|Y|^2}{m\theta^2_{\mathcal{D}}})+\ln(\frac{1}{\delta})}{m\theta^2_{\mathcal{D}}}})$ □

Thus, also for the case of optimistic risk minimization the risk can be made arbitrarily close to that of the optimal classifier in $\mathcal{H}$, conditional on the falsifiability

and knowledge parameters being not too large for the instance problem at hand. If, furthermore, the considered task is realizable (i.e. the optimal classifier in $\mathcal{H}$ has true risk equal to 0), then the expected risk for such methods converges to that of the Bayes classifier.

Despite these positive results, however, it is easy to observe that both instance-based methods and GRM present some limitations that may hinder their application in real-world problems. For the case of instance-based methods this limitation is already evident from Theorem 6 in Section 2.1: indeed, it is easy to observe that the expected risk of generalized nearest neighbors is exponential in the dimensionality of the feature space, thus showing that as the number of features grows also the tendency of this learning algorithm to over-fitting similarly increases. This problem will be addressed in Chapter 3 by means of the introduction of effective feature selection algorithms for learning from fuzzy label. For the case of optimistic risk minimization, on the other hand, the above mentioned limitations do not regard so much its sample complexity but rather its computational complexity. The next result shows that already for a very simple class of learning problems (which admit a computationally efficient solution in the supervised learning setting), optimistic risk minimization does not admit any polynomial-time algorithm (unless P = NP):

Proposition 1. *Let $\tilde{S}$ be an imprecise training set obtained sampling i.i.d. from $\tilde{\mathcal{D}}$, where the feature space $X = \mathbb{R}^d$ and $Y = \{-1, 1\}$. Let $\mathcal{H}$ be the class of half-spaces on X (i.e. $\mathcal{H} \cong \mathbb{R}^d$). Let $l : Y \times \mathbb{R} \to \mathbb{R}$ be a loss function satisfying the following properties:*

1. *For each $(x, y) \in Z$, l can be expressed as $l(y, \langle w, x \rangle)$ and is convex in w;*
2. *$sign(y) = sign(\langle w, x \rangle) \implies l(y, \langle w, x \rangle) < l(-y, \langle w, x \rangle)$;*
3. *If $sign(y - \langle w, x \rangle) = sign(y)$ then $l(y, \cdot)$ is monotonically increasing in $|y - \langle w, x \rangle|$.*

Let $\tilde{l}$ be the fuzzy loss obtained from l. Then, if l does not also satisfy the following property:

$$\forall t \in [-1, 1], l(y, t) = l(-y, t) \tag{2.5}$$

it holds that, for any polynomial-time randomized learning algorithm A, the probability that $|L_S^f(A_{\mathcal{H}}(S)) - L_S^f(GRM_{\mathcal{H}}^f(S))| \geq \epsilon$ is greater than $1 - O(e^{-\epsilon d})$;

Proof. It is easy show that for l satisfying conditions 1-3 in the theorem statement, it holds that, when $t \in [-1, 1]$, $l(1, t)$ is monotonic non-decreasing in $-t$ while $l(-1, t)$ is monotonic non-decreasing in t. Thus, unless $l(1, t) = l(-1, t)$ for any t in the same range, $\tilde{l}$ is not convex. In, particular, there either is at least a value $t \in [-1, 1]$ where $l(1, t) = l(-1, t)$ and $\tilde{l}$ is non-smooth or $\tilde{l}$ is unbounded in $[-1, 1]$. Then, the result follows from [142], Theorem 1, by noting that L_S^f is non-convex and non-smooth. □

As a consequence of the above result, it can be noted that the hinge loss, the log-loss and quadratic loss (as well as most other commonly adopted loss functions) all satisfy conditions 1-3, but do not satisfy Eq. (2.5): as a consequence, it is easy to show that popular learning algorithms such as least squares linear regression, SVM or logistic regression (which are polynomial-time in supervised learning) do not admit a polynomial-time extension to the optimistic risk minimization setting.

Pseudo-label methods could then be an alternative to both instance-based and GRM-based methods, that aim to obtain a good trade-off between the properties of these different approaches. Indeed, in experimental comparisons [201] pseudo-label based methods reported good empirical performance, comparable with or better than other general learning from imprecise labels methods. Nonetheless, the theoretical properties of pseudo-label based methods have been investigated only in the semi-supervised learning setting [15], while the most general setting of learning from fuzzy label has not been studied: a possible reason for this gap may regard the complexity of studying the dynamics of sequential iterative re-training of common pseudo-label learning methods, as shown in Algorithm 1.

The aim of this Section will be to study a novel pseudo-label learning algorithm called Random Resampling-based Learning (RRL), originally proposed in [50], which is based on the *parallel* composition of base classifiers trained using a precise learning algorithm, following the principles of ensemble learning. The pseudo-code formulation of RRL is shown in Algorithm 2.

Algorithm 2 The RRL algorithm.

procedure RRL(S: dataset, n: ensemble size, $\mathcal{H}$: base function class)

$Ensemble \leftarrow \emptyset$

for all iterations $i = 1$ to n **do**

Draw a boostrap sample S' from S

$Tr_i \leftarrow \emptyset$

for all $(x, \pi) \in S'$ **do**

Sample $\alpha \sim Uniform[0, 1]$

Add (x, y') to Tr_i, where $y' \sim Uniform[\pi^\alpha]$

end for

Add base model $h_i \in \mathcal{H}$ trained on Tr_i to $Ensemble$

end for

return $Ensemble$

end procedure

The RRL algorithm is an extension of Random Forest to the setting of learning from fuzzy label. The precise pseudo-label to be associated with the imprecise instances (x, π) in each of the bootstrap samples are drawn from a probability distribution compatible with π: in particular, pseudo-label are drawn from the distribution $\hat{Pr}_\pi(y) = \int_0^{\pi(y)} \frac{d\alpha}{|\pi^\alpha|}$, where π^α is the α-cut of π, i.e. $\pi^\alpha = \{y \in Y : \pi(y) \geq \alpha\}$. The distribution $\hat{Pr}$ is obtained by means of the possibility-probability transform [98] and implemented by means of a two-stage sampling procedure: first, a α-cut is selected uniformly at random, then, one element of the α-cut is selected uniformly at random. Intuitively, this sampling procedure favors class labels associated with higher possibility degrees. The above mentioned procedure is applied to obtain n bootstrap samples which are used to train a corresponding number of h base models. Finally the base models are aggregated by simple majority voting or averaging.

It is easy to observe that the computational complexity of the RRL algorithm is $O(Tn+|S||Y|n)$, where T is the cost required to train a base model h: thus, if h can be trained in polynomial time, also RRL can be trained in polynomial time, in contrast

with optimistic risk minimization studied in the previous sections. In regard to the generalization properties of RRL, it can be noted first that the sampling scheme for the pseudo-label can be given a formal justification, under weak assumptions about the data generating fuzzy random set $\tilde{D}$, by means of the following results which show that the sampling distribution of RRL corresponds to the distribution over labels for the *imprecise Bayes classifier*:

$$f^* = \arg\min_{f \text{ measurable w.r.t. } \tilde{\mathcal{D}}\downarrow(X\times\mathcal{S}(Y))} L_{\tilde{\mathcal{D}}}(f), \tag{2.6}$$

that is, the classifier with optimal performance among those that do not have access to the true labels.

Theorem 1. *Assume that $\tilde{\mathcal{D}}$ satisfies the following calibration property: with probability 1 over $(x_i, y_i, \pi_i) \sim \tilde{\mathcal{D}}$, it holds that $\tilde{\mathcal{D}} \downarrow (X \times Y)(y_i|x_i) \leq \pi_i(y_i)$. Then, f^* given by $Pr(f^*(x) = y) = \hat{Pr}_\pi(y)$ is the Bayes classifier w.r.t. the l_2 loss among probability distributions and the uniform prior.*

Proof. The calibration property assumed in the statement of the theorem guarantees that, for each $x \in X$, the true probability distribution over Y lies in the credal set $\mathbb{P}_{x,\pi} = \{P \in \mathcal{P}(Y) | P(y) \leq \pi(y)\}$. From [98], Theorem 1, it follows that $\hat{Pr}_\pi \in \mathbb{P}_{x,\pi}$ and $\hat{Pr}_\pi = \arg\min_{P\in\mathbb{P}_{x,\pi}} \mathbb{E}_{P'}[(P' - P)^2]$, where P' is selected uniformly from $\mathbb{P}_{x,\pi}$. Thus, among all possible distributions over Y, $\hat{Pr}_\pi$ is the one having minimal expected l_2 loss and the result follows. □

Corollary 2. *If the base class $\mathcal{H}$ is consistent, then RRL is consistent, that is, RRL converges to the imprecise Bayes classifier.*

Proof. The result follows from Theorem 1, consistency of $\mathcal{H}$ and the definition of RRL. □

Thus, in case RRL would have access to the whole data generating distribution $\tilde{D}$, the previous results would provide intuitive justification for the sampling scheme adopted in the corresponding algorithm, showing that in this case it would be equivalent to the imprecise Bayes classifier. Nonetheless, it is easy to see that, in general,

the ensemble classifier returned by RRL is not guaranteed to be the imprecise Bayes classifier, since the underlying data generating distribution $\tilde{D}$ is unknown and in general cannot be estimated from finite samples. To address this shortcoming, the following two results then study the generalization properties of RRL, under two different assumptions about the base function class $\mathcal{H}$. The first results assumes that the base function class is a bounded convex set with finite Natarajan dimension d and that the loss function which is used to measure the accuracy of the classifiers is Lipschitz, allowing to derive a finite sample bound based on the rich literature on randomized kernel methods [196]. First, recall that a set $\mathcal{H}$ is convex if $\forall h_1, h_2 \in \mathcal{H}$ and $\alpha \in [0,1]$ it holds that $\alpha h_1 + (1-\alpha)h_2 \in \mathcal{H}$, similarly a function l is convex if $l(\alpha h_1 + (1-\alpha)h_2) \leq \alpha l(h_1) + (1-\alpha)l(h_2)$ and it is L-Lipschitz if $|l(h_1) - l(h_2)| \leq L|h_1 - h_2|$. Then, the following theorem holds:

Theorem 2. *Assume the base hypothesis class $\mathcal{H}$ is a bounded convex set in a Hilbert space of functions $X \to \mathbb{R}^Y$, with $sup_{x,h}|h(x)| \leq 1$ and Natarajan dimension d. Let p be the probability density over $\mathcal{H}$ determined by RRL and let $C = min_{h\in\mathcal{H}} p(h) > 0$. Let $l : Y \times Y \to [0,1]$ be a loss function which is L-Lipschitz w.r.t. its second argument. Then, when the RRL algorithm is executed on a training set S, with $|S| = m$, sampled i.i.d. from $\tilde{\mathcal{D}}$, it returns a function $\hat{h} = \frac{1}{n}\sum_i h_i$ s.t. $|\min_{h\in\mathcal{H}} L_D(h) - L_S^f(\hat{h})|$ can be upper bounded by*

$$(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})\frac{|Y|L}{C}\sqrt{log\frac{2}{\delta}} + \sqrt{\frac{r \cdot ln(\frac{r|Y|^2}{\theta_{\mathcal{D}}^2}) + ln\frac{1}{\delta}}{m\theta_{\mathcal{D}}}} + \sqrt{\frac{K_n + ln\frac{m}{\delta}}{2(m-1)}} \tag{2.7}$$

with probability greater than $1-\delta$ over the sampling of the S and the randomized execution of RRL, where K_n depends only on $\tilde{\mathcal{D}}$, m and $r = \max\{n, d\}$.

Proof. Since $\mathcal{H}$ is a class satisfying the assumptions given in the statement, each $h \in \mathcal{H}$ can be expressed as $h = \int_{\mathcal{H}} \alpha(f) f df$, with $\int_{\mathcal{H}} \alpha(h) dh = 1$ and $\forall h \in \mathcal{H}, \alpha(h) \geq 0$. Let $h^* = \arg\min_{h\in\mathcal{H}} L_D(h)$. Assume the learning algorithm A for $\mathcal{H}$ is deterministic, and let $S_1, \ldots, S_n$ be the bootstrap samples randomly selected in any randomized execution of RRL. Denote with $h_i = A(S_i)$ and let $h^+ = \arg\min_{h\in\mathcal{H}}\{L_S^f(h) : h = \sum_i \alpha_i h_i \wedge \sum_i \alpha_i = 1 \ \wedge \forall_i \alpha_i \geq 0\}$. Then, the generalization gap $|L_D(h^*) - L_S^f(\hat{h})|$

can be decomposed and upper bounded as:

$$|L_D(h^*) - L_D(h^+)| + |L_D(h^+) - L_S^f(h^+)| + |L_S^f(h^+) - L_S^f(\hat{h})|,$$

thus the risk of RRL can be estimated by bounding the three terms above separately. By [196], Theorem 1, and noting that since l is L-Lipschitz $\tilde{l}$ is $L|Y|$-Lipschitz, the first term can be upper bounded by

$$(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})\frac{|Y|L}{C}\sqrt{log\frac{2}{\delta}}.$$

For the second term, note that function h^+ can be expressed as a linear classifier defined over a n-dimensional feature space A, where A is the space obtained by convex combinations of functions in the ensemble returned by the RRL algorithm. Since $\mathcal{H}$ has Natarajan dimension d and is convex, the Natarajan dimension of the above mentioned linear classifier is $r = O(\max\{n, d\})$. Thus, the second term can be bounded, by Theorem 5 in Section 2.1, as:

$$\sqrt{\frac{r \cdot ln(\frac{r|Y|^2}{\theta_{\mathcal{D}}^2}) + ln\frac{1}{\delta}}{m\theta_{\mathcal{D}}}}.$$

Finally, noting that $L_S^f(h^+) \leq L_S^f(\hat{h})$ and $L_S^f = \mathbb{E}_{S_i^f} L_{S_i^f}$, where S_i^f is sampled i.i.d. from S, and h^+ can be written in the form $h^+ = \sum_i \alpha_i h_i$, the third term can be upper bounded by a simple argument based on PAC-Bayes learning (see [238], Theorem 1) as:

$$\sqrt{\frac{KL(\alpha||u) + ln\frac{m}{\delta}}{2(m-1)}},$$

where α is the probabilitity distribution s.t. $P(h_i) = \alpha_i$, u is the probability distribution s.t. $P(h_i) = \frac{1}{n}$, and KL is the Kullback-Leibler divergence. Letting $K_n = sup_{S\sim\tilde{D}}\mathbb{E}_{h_1\sim p,\ldots,h_n\sim p}KL(\alpha||u)$, the result follows. □

Thus, the result above shows that RRL's generalization error, as the training set size m and the number of ensembled models n grows to infinity, converges to the generalization error of optimistic risk minimization. Indeed, the first and last terms of Eq. (2.7) converge to 0 with a rate that is equivalent to the square root of the above mentioned parameters. It can be noted, however, that while the previous

theorem could be applied to obtain generalization bounds for RRL when using linear or kernel methods as base classifiers, the same does not hold for the common case of tree-based classifiers, which nonetheless are among the most popular methods as base classifiers for ensemble methods due to their computational efficiency and good performance [204]. Indeed, such classes of classifiers do not usually satisfy the assumption in the previous theorem.

The following result, then, focuses on a setting which is more similar to that of the standard Random Forest algorithm: assuming the classifiers in the ensemble to be independent of each other, a tail bound on the probability of error is directly obtained by an application of either Chernoff's or Slud's inequalities [34]:

Theorem 3. *Let l_{0-1} be the 0-1 loss. Let $\mathcal{H}$ be a class of hypotheses whose Natarajan dimension is d. Let $\hat{h}$ be the function returned by Algorithm 2 and $\mathcal{H}_A \subseteq \mathcal{H}$ be the set of hypothesis whose averaging is $\hat{h}$. Let $\gamma_T = \max_{h \in \mathcal{H}_A} L_S^f(h) + \epsilon \leq \frac{1}{2}$ and $\gamma_V = \max_{h \in \mathcal{H}_A} L_v^f(h) + \sqrt{\frac{\log(2/\delta)}{2m_v^h}} \leq \frac{1}{2}$. Then, assuming the $h \in \mathcal{H}_A$ err independently, the following inequalities hold jointly with probability greater than $1 - 2\delta$:*

$$1 - L_D(h) \geq \frac{1}{2}(1 - \sqrt{1 - e^{\frac{-K\gamma_T^2}{1-\gamma_T^2}}}) \tag{2.8}$$

$$1 - L_D(h) \geq \frac{1}{2}(1 - \sqrt{1 - e^{\frac{-K\gamma_V^2}{1-\gamma_V^2}}}) \tag{2.9}$$

$$L_D(h) \leq e^{-n \cdot KL(\frac{1}{2}||\gamma_V)} \tag{2.10}$$

where $\epsilon = 2\sqrt{\frac{2d(\ln(m_T)+\ln(|Y|))}{m_T}}$, m_T is the size of the training set, m_v^h is the size of the out-of-bag validation set for base classifier h, $L_v^f(h)$ is out-of-bag error for base classifier h and $KL(a||b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$ is the Kullback-Leibler divergence.

Proof. Inequality (2.8) follows by applying Slud's inequality [34, 219] to $\mathcal{H}_A$, by noting that $L_D(h)$ is distributed as a Bernoulli random variable whose parameter p is upper bounded by γ_T and $\hat{h}$ errs on an instance x iff at least $K/2$ hypotheses in $\mathcal{H}_A$ also err. Inequality (2.9) similarly follows by Slud's inequality, bounding $L_D(h)$ with the validation error derived by direct application of Hoeffding's inequality. Finally, inequality (2.10) follows from Chernoff's bound for binomial distributions [9]. □

It is easy to notice, that under the condition of independence among the base classifiers, the previous theorem, along with Theorem 1, implies that as the number of ensembled base classifiers n grows to infinity, the performance of RRL converges to that of the imprecise Bayes classifier, a result which is analogous to the consistency of Random Forest in the standard supervised learning setting [28]. Nonetheless, even though widely assumed in the literature on ensemble methods [28], the assumption of independence of the base classifier is rather strong and in general cannot be guaranteed to hold as n grows, in which case RRL may have a rate of convergence much smaller than exponential or may even fail to be consistent [104]. In any case, two differences can be noted between Theorems 2 and 3. On the one hand, the two theorems apply to different classes of base function classes: indeed, while Theorem 2 applies to convex base classes it cannot be applied to tree-based models, as mentioned above, while Theorem 3 cannot be directly applied to convex base classes as the corresponding learning problems usually satisfy stability assumptions [219] that would make the independence assumption not applicable. On the other hand, Theorem 3 directly bounds the l_{0-1} loss generalization error of RRL, while Theorem 2 only provides a bound in terms of a surrogate convex loss l for which, generally, it holds that $l_{0-1} \leq l$. Thus, Theorem 2 provides a less informative bound than Theorem 3 whenever accuracy is the real target performance metric.

Concluding this section, it is not hard to observe that the RRL algorithm provides a trade-off among the positive characteristics of instance-based methods as well as optimistic risk minimization. Similarly to instance-based methods, the time complexity of RRL is polynomial as long as the time required to train the base classifiers is also polynomial. This is in contrast with optimistic risk minimization, for which the associated learning problem was shown to be in general computationally hard. On the other hand, RRL shares favourable risk bounds with optimistic risk minimization. Indeed, in general the generalization risk of RRL increases only polynomially with the dimensionality of the input space, in contrast with instance-based methods where in general the growth in generalization risk is exponential in the dimensionality. Furthermore, it can easily be seen that, under certain conditions,

the generalization error of RRL asymptotically tends, when the sample size m and the number of ensembled models n grow, to the bound shown in Theorem 5 for optimistic risk minimization. Nonetheless, it can be noted that the above mentioned error bounds suffer from the same limitations that were previously mentioned also for the other considered learning from fuzzy label ML algorithms. In particular, the obtained bounds depend on hardness parameters of the data generating distribution, which in general are unknown. Thus, application of these bounds in real-world settings can be difficult when no information about such parameters is available or when they cannot be estimated. For this reason, even more so than for standard supervised ML, experimental evaluation is of paramount importance in the validation of learning from fuzzy label algorithm: the following section, then, will be devoted to the assessment of common state-of-the-art methods for learning from fuzzy label.

2.3 Experimental Analysis

As a complementary focus to the above theoretical analysis, the aim of this section will be to discuss the empirical validation and experimental comparison of state-of-the-art learning from fuzzy label algorithms, based on a large benchmark suite, encompassing both synthetic and real-world datasets. The following algorithms, in particular, were considered:

- Two pseudo-label learning algorithms, namely the RRL (denoted as RRL) algorithm described in the previous section and the state-of-the art POP algorithm (denoted as PLC) introduced in [262] (itself, a modification of the progressive identification learning algorithm proposed in [103, 165]), using a multi-layer perceptron as base model;

- Two variants of instance-based methods, namely *generalized nearest neighbors* (denoted as GNN), i.e. the instantiation of learning rule (2.3) where $N(x)$ are the k nearest neighbors of x, and *generalized radius neighbors* (denoted as GRN), i.e. the instantiation of learning rule (2.3) where $N(x)$ are all instances

at distance smaller than ϵ from x, for ϵ a threshold hyper-parameter. For the case of GNN, the hyper-parameter k was set to 5[2], while for the case of GRN the hyper-parameter ϵ was optimized during training;

- A hybrid pseudo-label and instance-based learning method, called DELIN [16, 259, 276] (denoted as DELIN). DELIN combines a pseudo-label learning approach for dimensionality reduction based on linear discriminant analysis with an instance-based classification method based on generalized nearest neighbors. The two algorithms are iteratively and alternatively executed to improve the classification performance of standard instance-based methods by addressing the curse of dimensionality. Since the number of reduced dimension is a hyper-parameter, this was optimized during training and validation. For the generalized nearest neighbors classifier, as before, the number of neighbors was set to 5;

- Two implementations of generalized risk minimization, namely a version generalizing linear SVM learning using hinge loss as base loss (denoted as GRMSVM), and a version generalizing a single hidden layer multi-layer perceptron using the cross-entropy loss as base loss (denoted as GRMNN).

For all of the above mentioned algorithms, in particular, their scikit-weak implementation (see Appendix A) was considered.

Algorithms were evaluated on both contaminated version of standard precise benchmark datasets from the UCI collection [96], as well as on real imprecise datasets. The full list of datasets is reported in Table 2.1. For the precise benchmark datasets two different contamination models were considered:

- Fully random contamination: for each training instance x, each of the wrong labels y' is randomly assigned possibility degree $\pi(y')$ with probability

$$\binom{n}{\lceil \pi(y') \cdot n \rceil} \epsilon^{\lceil \pi(y') \cdot n \rceil} (1-\epsilon)^{n - \lceil \pi(y') \cdot n \rceil},$$

[2]This value was selected as default in analogy with the default recommended value in the scikit-learn library (see `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`).

where $\epsilon \in \{0.1, 0.25, 0.5, 0.7, 0.9\}$, $n = 100$, while the correct label is assigned possibility degree $\pi(y) = 1$. This contamination model represents a generalization of the random contamination model for superset learning adopted in [179] to the learning from fuzzy label setting and corresponds to drawing n random samples from a Bernoulli random variable (with parameter ϵ) and selecting for y' the possibility degree matching the observed number of successes;

- Label relaxation contamination [152]: in particular, a k-nearest neighbors model was used for realizing the label relaxation. For each training instance x, the $k \in \{3, 5, 7\}$ nearest neighbors of x (including x itself) are selected and each label y is given possibility degree $\pi(y) = \frac{|\{x' \in N(x):(x',y) \in S\}|}{max_{y' \in Y} |\{x' \in N(x):(x',y') \in S\}|}$. Notice that for this contamination model the possibility degree of the correct class label y is always $\pi(y) > 0$ but, in general, it may happen that $\pi(y) \neq 1$.

For the real-world imprecise datasets, 5 different medical tasks were considered:

- Circulating Tumor Cells detection [50, 227, 226] from fluorescence microscopy, where fuzzy labels are obtained by consensus among 11 raters. In particular, each rater provided a label y and $\pi(y) = \frac{\text{num. of raters who proposed label } y}{\max_{y'} \text{num. of raters who proposed label } y'}$;

- COVID-19 diagnosis from routine laboratory exams [43], where fuzzy labels are obtained by weighted consensus (based on sensitivity and specificity of the medical tests) among the results of a RT-PCR swab test and computer imaging;

- Knee lesion detection [40] from magnetic resonance imaging, where fuzzy labels are obtained by confidence-weighted consensus among 12 raters;

- Spine surgery invasiness prediction [51], as an example of semi-supervised task, in which a single rater labeled all instances as either non-invasive, invasive or uncertain;

- Sagittal misalignment assessment [52], as an example of superset learning task, where sets of labels are obtained by selecting all labels provided by two medical specialists who annotated the dataset.

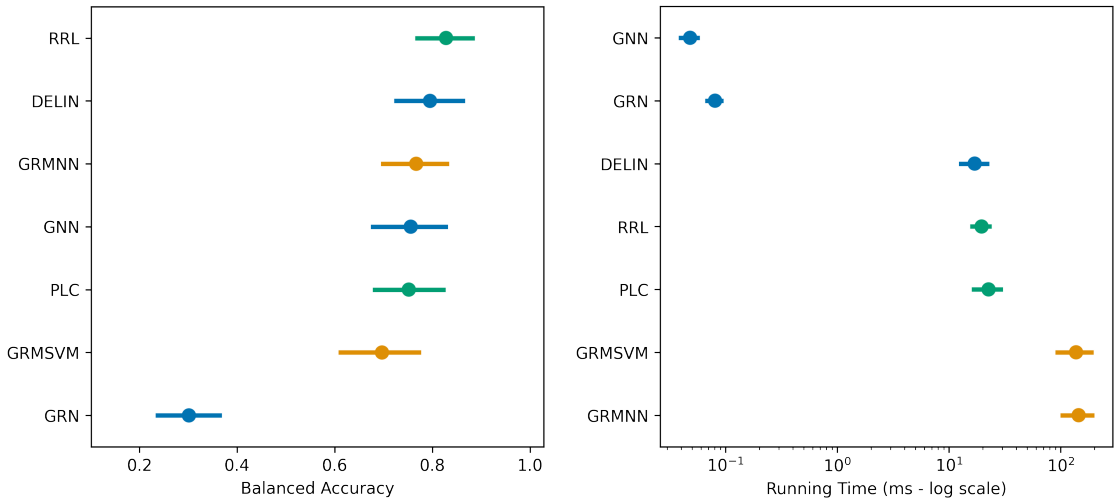
Table 2.1: List of datasets considered in the experimental comparison of learning from fuzzy label algorithms.

	Classes	Features	Instances
UCI datasets			
avila	10	10	20768
banknote	2	4	1372
cancerwisconsin	2	9	683
car	4	16	864
credit	2	61	1000
crowd	6	28	10845
diabetes	2	8	768
digits	10	62	5620
frog-family	4	22	7195
frog-genus	8	22	7195
frog-species	10	22	7195
hcv	4	12	582
htru	2	8	17898
ionosfera	2	33	351
iranian	2	45	7032
iris	3	4	150
mice	8	78	972
mushroom	6	99	5644
myocardial	2	111	1700
obesity	7	31	2111
occupancy	2	5	20560
pen	10	16	10992
robot	4	24	5456
sensorless	11	48	20000
shill	2	9	6321
sonar	2	60	208
vowel	11	9	990
wifi	4	7	2000
wine	3	13	178
Imprecise Datasets			
ctc	2	2500	617
covid	2	69	1624
mri	2	100	427
invasiveness	3	186	72
spine	7	14	120

All algorithms were evaluated in a 10-repeated 5-fold cross-validation experimental setting, to take into account sensitivity to initialization and randomization. In particular, all models were evaluated in terms of balanced accuracy, in order to measure the models' error rate also under conditions of label imbalance, and running time (in ms), as a measure of computational efficiency. For the synthetically contaminated datasets, balanced accuracy was evaluated by comparison with the known ground truth labeling (which was not available to the learning algorithms). For the real-world imprecise datasets, instead, balanced accuracy was evaluated on a subset

of the data whose labels were precise, that is: for the ctc, mri and spine datasets, the test sets encompassed only instances on which all raters proposed the same label; for the covid dataset, the test set encompassed only instances on which the two diagnostic tests provided the same diagnosis, while for the invasiveness dataset the test set encompassed only instances rated as invasive or non-invasive. Statistical analysis of the results was performed by means of a ranking comparison, using Friedman test with Nemenyi post-hoc procedure [24, 85].

Results of the experimental analysis are reported in Figures 2.1a and 2.2, in terms of balanced accuracy, and 2.1b and 2.3, in terms of running time.



(a) (b)

Figure 2.1: Results of the experiments. Left: mean balanced accuracy scores of the models under study (higher is better), Error bars denote 95% C.I. Mean running times (ms) of the models under study (lower is better). Error bars denote 95% C.I. Legend, okra: generalized risk minimization based, green: pseudo-label learning based, blue: instance-based methods.

In terms of balanced accuracy, the three best models were RRL, DELIN and multi-layer perception GRMNN. In particular, RRL was the best algorithm in terms of both raw balanced accuracy as well as average ranks: even though the performance of RRL and DELIN was not statistically significant, RRL still reported better

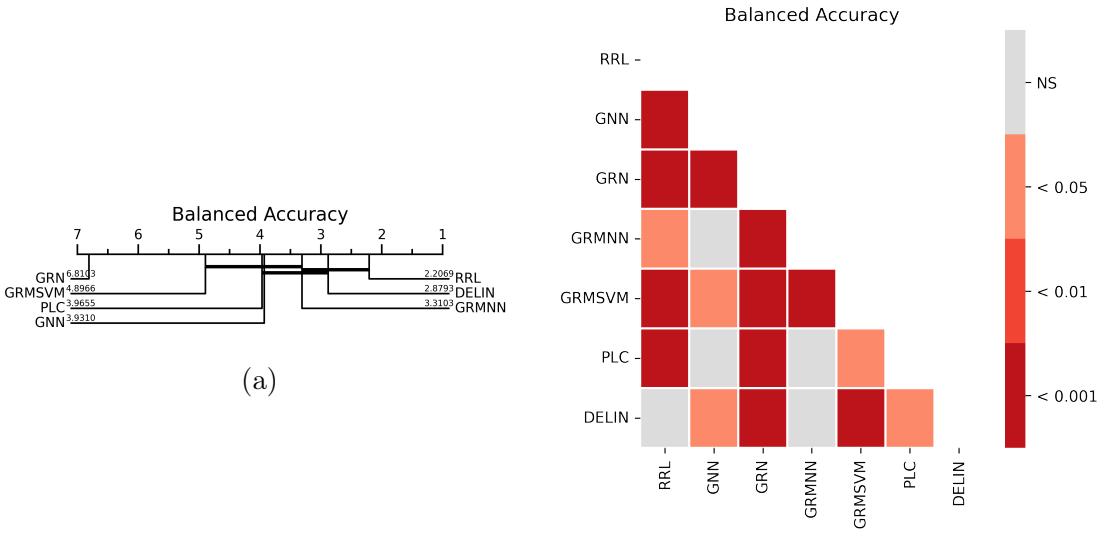
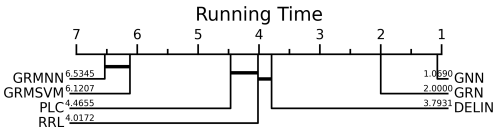
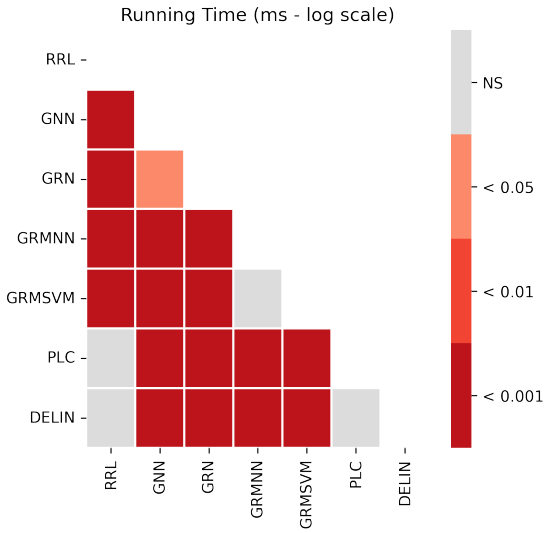


Figure 2.2: Comparison of the the models under study in terms of balanced accuracy. Left: critical difference diagram of the mean ranks (lower is better), bars denote significance cliques at 95% confidence level. Right: heatmap of p-values obtained with the post-hoc Friedman-Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank.

performance on average and was further significantly better than all other considered algorithms. Similarly, no significant difference was detected among DELIN and GRMNN, as well as between GRMNN, GNN and PLC. These results confirm the good performance of RRL, which can then be related with the theoretical results demonstrated in the previous section. Indeed, the performance of RRL was comparable with those of GRMNN and DELIN, respectively an optimistic risk minimization and a (dimensionality reduced) instance-based method: interestingly, however, the proposed RRL algorithm reported better on average performance than the other two. Also this difference could be explained by referring to the theoretical results of the previous sections. For the case of GRMNN, the hardness of solving the optimistic risk minimization could lead to premature convergence to local minima or



(a)



(b)

Figure 2.3: Comparison of the the models under study in terms of running time.Left: critical difference diagram of the mean ranks (lower is better), bars denote significance cliques at 95% confidence level. Right: heatmap of p-values obtained with the post-hoc Friedman-Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank.

saddle points, and consequently to suboptimal generalization error. For the case of DELIN, by contrast, even though this algorithm performs a data dimensionality pre-processing step to reduce the risk of over-fitting of GNN, the number of reduced dimension may still be too large to avoid the curse of dimensionality: indeed, in the experiments the number of reduced dimension was dynamically optimized during cross-validation, thus leading to possible over-fitting the hyper-parameter. By contrast the worst performing algorithm was GRN, which was reported significantly lower performance than all the other considered methods. Interestingly, GRMNN reported significantly better performance than GRMSVM, likely due to the fact that most of the considered datasets did not satisfy the linear separability assumed by the linear SVM model underlying GRMSVM.

In terms of running time, the best performing algorithm was GNN, which was significantly more computationally efficient than all other considered algorithms excluding GRN. This result is expected, since the training time of lazy instance-based methods such as GNN and GRN is typically constant or linear in the size of the training set. By contrast, the two worst performing algorithms were both generalized risk minimization methods, namely GRMNN and GRMSVM, which were significantly less computationally efficient than all other consider algorithms. This result, on the one hand, confirms the general hardness of these learning algorithm (see Theorem 1; on the other hand, it can be remarked that memory transfer bottlenecks in the scikit-weak implementation of these algorithms, which employ GPU for tensor processing optimization, could also have a role in the observed performance gap. Further research should devoted at decomposing these two computing costs, and possibly optimizing memory usage. The proposed RRL algorithm reported a running time which was intermediate between those of instance-based methods and generalized risk minimization ones, having in particular an average running time comparable with (i.e. not significantly different from) that of DELIN.

Thus, the experimental results show the effectiveness of the proposed RRL algorithm: indeed, the proposed approach reported a running time which was comparable or better than other state-of-the-art methods for learning from fuzzy label, while at the same time exhibiting the best generalization accuracy among the compared methods. These results, furthermore, are complemented by the robust generalization guarantees for RRL which were proved in the previous section and which provide an interpolation between the properties of instance-based methods and generalized risk minimization ones.

2.4 Conclusion

The aim of this chapter was to study the problem of learnability in the setting of learning from fuzzy label. To this aim, through two theoretical results, the first contributions consisted in providing robust generalization guarantees, as well as an

analysis from a computational complexity perspective, for two of the main learning paradigms in this setting, namely instance-based methods and generalized risk minimization. Furthermore, the second main contribution consists in the proposal of a novel pseudo-label learning algorithm, called RRL, and the study of its statistical and generalization properties, which marks the first theoretical investigation of such paradigm in the setting of learning from imprecise data. These theoretical contributions have then been complemented by a third, experimental, contribution through which the performance (in terms of generalization accuracy and running time) of several state-of-the-art methods for learning from fuzzy label has been compared, showing, in particular, the effectiveness of the proposed RRL algorithm, which confirms and reinforces the presented theoretical analysis. In light of these results and contributions, the following open problems could be worthy of further research:

- The focus of this chapter has been on the investigation of a specific instance of the learning from imprecise data, namely the learning from fuzzy label problem. While the following chapters (see Chapter 4), will investigate the application of the RRL algorithm to the setting of learning from fuzzy data, future work should be devoted to the investigation of the theoretical characterization as well as of practical algorithms for more general forms of imprecise data;

- Several theoretical characterization of the main learning from fuzzy label paradigms, namely generalized risk minimization, instance-based methods and pseudo-label learning, have been considered, focusing on the establishment of upper bounds on the learnability of this setting: further work should be devoted at exploring tighter bounds, especially under constraining assumptions on the problem instances, as well as matching lower bounds;

- From an empirical perspective, the performance gap reported by generalized risk minimization algorithms, despite being consistent with the hardness of the associated optimization problems, could partially be attributed to costs related to GPU usage. Further work should be devoted at assessing this hypothesis and optimizing resource usage to improve the efficiency of these algorithms.

Chapter 3

Feature Selection in Learning from Imprecise Data

The aim of the previous chapter has been to study the learnability of the learning from fuzzy label problem: as a main result it has been shown that for three widely adopted learning paradigms, namely generalized risk minimization, instance-based methods and pseudo-label learning, learnability is indeed possible. Nonetheless, it has also been shown that the dimensionality of the input feature space X (i.e. the number of features) may have a significant impact on either the generalization or the computational complexity of the above mentioned models. On the hand, for instance-based methods it has been shown that even though the complexity grows linearly with the dimensionality, the generalization error instead grows exponentially, exhibiting a so-called *curse of dimensionality* phenomenon. Similarly, for the generalized risk minimization and pseudo-label learning it is not hard to show that their computational complexity and generalization error grows at least linearly with the dimensionality of the feature space. These results are not unexpected: indeed, such a dependence directly translates from the standard supervised setting [219] where, without further assumptions (e.g. large margin or sparsity assumptions), dimensionality directly affects both computational complexity [215] (e.g. algorithms for computing norms or inner products have usually complexity at least linear in the

dimensionality of the feature space[1]) as well model complexity [186, 249] (e.g. for linear models the Natarajan dimension introduced in Chapter 2 is equal to the dimensionality of the feature space [219]) and has thus long been studied. Indeed, several feature selection, regularization and dimensionality reduction have been developed in the supervised learning setting to address this problem: these include unsupervised approaches (e.g. principal component analysis, autoencoders [115] and topological methods [169]) as well as supervised ones (e.g. linear discriminant analysis, filter methods such as Relief [242], embedded methods such as LASSO [211, 233], or general wrapper methods).

By contrast, limited work has focused on feature selection or dimensionality reduction in learning from imprecise data [259]. Research in this sense has mostly focused on the development of dimensionality reduction algorithms for superset learning, with the state-of-the-art algorithms for such a task being the DELIN algorithm and its variations [16, 17, 259, 276]. From an algorithmic perspective, DELIN can be understood as a classification and dimensionality reduction method based on pseudo-label learning relying on linear discriminant analysis (LDA) and generalized nearest neighbors as sub-routines, as described in the pseudo-code in Algorithm 3. Thus, in its most basic implementation, DELIN is an iterative algorithm that, starting from a uniform distribution over labels in the corresponding supersets, subsequently alternates a dimensionality reduction step based on LDA with a classification step on the reduced data using GNN, whose predicted label distributions are then used in the successive iteration. The two-step procedure is repeated until convergence, or after a fixed number of iterations has been performed.

In previous research, DELIN has been shown to outperform unsupervised dimen-

[1]In some cases, for example in kernel methods, one usually considers two different features spaces, namely the original feature space X and an augmented feature space $\Phi(X)$, defined by a map $\Phi : X \to X'$, where X' is an Hilbert space s.t. $dim(X') > dim(X)$ (X' can be even infinite-dimensional). Typically, one uses the *kernel trick* [215] to reduce the computation of an inner product on X' to the computation of a function on X: thus, while the complexity of computing such an inner product is independent of the dimension of X', it is still typically linear in the dimension of X.

Algorithm 3 The DELIN algorithm.

procedure DELIN(d: number of dimensions, k: number of neighbors, n: number of iterations, S: training set)

$S_{temp} \leftarrow \emptyset$

for all $(x, L) \in S$ **do**

Add $(x, \{y : \frac{1}{|S|}\}_{y \in S})$ to S_{temp}

end for

for all iterations $i = 1$ to n **do**

Train LDA on S_{temp} with d dimensions

$S_{red} = LDA(S_{temp})$

Train a k-GNN model on S_{red}

$S_{temp} \leftarrow \emptyset$

for all $(x, L) \in S$ **do**

Add $(x, \{y : GNN(LDA(x))_y\}\}_{y \in Y})$ to S_{temp}

end for

end for

return LDA, GNN

end procedure

sionality reduction methods for superset learning, as well as to improve the generalization of classification algorithms, especially so for instance-based methods [17, 276]. However, despite these advantages, DELIN is affected by some limitations that hinder its applicability in real-world problems. First, being based on linear discriminant analysis, DELIN relies on the parametric assumptions required by this latter model, namely that the data features are distributed as normal, i.i.d. variables. Second, the number of reduced dimensions that DELIN should extract from the original data is a hyper-parameter that has to be fixed a-priori (or discovered via hyper-parameter optimization). Third, DELIN is a dimensionality reduction rather than a feature selection algorithm: while this could be helpful for visualization purposes when the number of selected dimensions is small, this property may also hinder interpretability

in cases where the semantics of the original features is relevant. Finally, the DELIN algorithm has been defined and applied only to the specific case of superset learning, while its generalization to the more general setting of learning from fuzzy data has not been considered in the literature.

As a consequence of the above mentioned limitations of DELIN, the following sections will be focused on the development of a novel feature selection approach based on Rough Set theory [185], a general framework for the representation and management of uncertainty in data which has been widely and successfully applied in feature selection [23, 231], in order to address research problem **P1.2**. In particular, the main aim of this chapter will be to develop a non-parametric feature selection method based on the generalization of Rough Set theory to the setting of learning from imprecise labels. Intuitively, the Rough Set-theoretic approach to feature selection is based on the notion of a *reduct* [118, 134, 135, 222, 228, 231]. Informally, this latter can be described as a subset of features which allows to preserve all information about the target variable, while being minimal with this property, i.e. all proper subsets of a reduct should introduce some new classification error.

More formally, the Rough Set-theoretic framework assumes a factorization of the feature space $X = A_1 \times \dots A_d$, where each A_d corresponds to a specific feature (e.g., in the continuous case where $X = \mathbb{R}^d$, it holds that $A_i = \mathbb{R}$ for each i). In its most general formulation, it also assumes that for each $B \subseteq A = \{A_1, \dots, A_d\}$ it exists a *granulation structure* on X, specifying for each instance x its set of B-neighbors $\mathcal{N}_B(x)$. Depending on the nature of X, various concrete implementations of the notion of a granulation structure have been proposed [49, 222, 269]. For example, when each A_i is discrete, one usually assumes that $\mathcal{N}_B(x) = \{x' \in X : \forall A_i \in B, A_i(x) = A_i(x')\}$, i.e. the granulation structure defines an equivalence relation on X[2]. On the other hand, when the A_i are continuous, typical constructions refer either to some discretization of the features [3] or to some topology on X [255, 270, 284]. Aside, from the granulation structure on the feature space, one also considers the standard granulation on the target space Y given by $\mathcal{N}_Y(x) = \{x' \in X : y_x = y_{x'}\}$.

[2]This corresponds to Pawlak's rough set model[185].

A reduct for a given training set S, is a set of features $B \subseteq A$ such that:

- The granulation structure $\mathcal{N}_B$, is *consistent with* the target granulation $\mathcal{N}_Y$;
- Given any $C \subset B$, the granulation structure $\mathcal{N}_C$ is not consistent with $\mathcal{N}_Y$.

Then, several instantiations of the notion of reduct can be derived based on the precise definition of consistency between two granulations [222]. As an example, one of the most popular definitions of reduct relies on the notion of a generalized decision $D_B(x) = \{y' \in Y : \exists x' \in \mathcal{N}_B(x) \text{ s.t. } y_{x'} = y\}$, and defines a set of feature $B \subseteq A$ to be consistent with Y on the training set S if $\forall x \in X, D_B(x) \subseteq D_A(x)$.

Even though the framework of Rough Set theory has originally been proposed in the settings of either fully unsupervised or fully supervised data [185], it has since been extended to various settings of uncertain data, including missing data [253, 229, 260], interval, incomplete and set-valued data [205, 208, 206, 207], and fuzzy or possibilistic data [68, 175]. Nonetheless, as highlighted in the recent review [49], no previous work has focused on the task of feature selection in the setting of learning from imprecise data, which, as described in the previous sections, features a complex interaction between learning (i.e. generalization to new data) and disambiguation (i.e. discrimination of the most likely data instantiations in the training data). To address this gap, starting from the more specific setting of superset learning, in Section 3.1, the classical Rough Set-theoretic approach to feature selection, and in particular the notion of a reduct, will be generalized to the learning from fuzzy label setting, in Section 3.2, by adapting the optimistic risk minimization and minimum description length [117] principle to the Rough Set-theoretic framework. In both the superset learning and the learning from fuzzy label settings a complete characterization of the computational complexity of the proposed feature selection methods will be provided, showing that, generally, the problem of feature selection is computationally hard. Then, as a way to resolve this computational complexity issue, in Section 3.3, several approximation algorithms, based either on heuristics or meta-heuristics, will be discussed and analyzed from a computational perspective as well as in terms of their empirical performance, in comparison with DELIN.

International Journal of Approximate Reasoning 136 (2021) 150–167

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Rough set-based feature selection for weakly labeled data

Andrea Campagner [a,*], Davide Ciucci [a], Eyke Hüllermeier [b]

[a] Department of Informatics, Systems and Communication University of Milano–Bicocca, viale Sarca 336, 20126 Milano, Italy
[b] Institute of Informatics, University of Munich (LMU), Germany

ARTICLE INFO

Article history:
Received 28 December 2020
Received in revised form 8 June 2021
Accepted 9 June 2021
Available online 18 June 2021

Keywords:
Superset Learning
Rough Sets
Feature Selection
Evidence Theory
Entropy

ABSTRACT

Supervised learning is an important branch of machine learning (ML), which requires a complete annotation (labeling) of the involved training data. This assumption is relaxed in the settings of *weakly* supervised learning, where labels are allowed to be imprecise or partial. In this article, we study the setting of superset learning, in which instances are assumed to be labeled with a set of *possible* annotations containing the correct one. We tackle the problem of learning from such data in the context of *rough set theory* (RST). More specifically, we consider the problem of RST-based feature reduction as a suitable means for *data disambiguation*, i.e., for the purpose of figuring out the most plausible precise instantiation of the imprecise training data. To this end, we define appropriate generalizations of decision tables and reducts, using tools from generalized information theory and belief function theory. Moreover, we analyze the computational complexity and theoretical properties of the associated computational problems. Finally, we present results of a series of experiments, in which we analyze the proposed concepts empirically and compare our methods with a state-of-the-art dimensionality reduction algorithm, reporting a statistically significant improvement in predictive accuracy.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

Weakly supervised learning [69] refers to machine learning tasks in which training instances are not required to be associated with a precise target label. Instead, the annotations can be imprecise or partial. Such tasks could be the consequence of certain data pre-processing operations such as anonymization [15,49] or censoring [17], could be due to imprecise measurements or expert opinions, or meant to limit data annotation costs [45]. Some examples of weakly supervised learning tasks include semi-supervised learning, but also more general tasks like learning from soft labels [8,12,13,48], (in which partial labels are represented through belief functions) which, in turn, encompasses both learning from fuzzy labels [14,28] (in which partial labels are represented through possibility distributions) and superset learning [29,40,44]. In this latter setting, which will be the focus of this article, each instance x is annotated with a set S of candidate labels that are deemed (equally) *possible*. In other words, we know that the label of x is an element of S, but nothing more. For example, an image could be tagged with $\{\text{horse}, \text{pony}, \text{zebra}\}$, suggesting that the animal shown on the picture is one of these three, though it is not exactly known which of them.

In the recent years, the superset learning task has been widely investigated both under the classification perspective [19,30,64,66] and from a theoretical standpoint [39]. The latter result is particularly relevant, as it shows that, as in the standard PAC learning model, superset learnability is characterized by combinatorial dimensions (e.g., Vapnik-Chervonenkis

* Corresponding author.

https://doi.org/10.1016/j.ijar.2021.06.005
0888-613X/© 2021 Elsevier Inc. All rights reserved.

or Natarajan dimension) which, in general, depend on the dimensionality (i.e., the number of features) of the learning problem. Thus, the availability of effective *feature selection* [24] or dimensionality reduction algorithms would be of critical importance in order to control model capacity and, hence, ensure proper model generalization. Nevertheless, this task has not received much attention so far [61].

In this article, which is an extension of our previous article [6], we study the application of *rough set theory* in the setting of superset learning. In particular, adhering to the generalized risk minimization principle [28], we consider the problem of feature reduction as a mean for *data disambiguation*, i.e., for the purpose of figuring out the most plausible precise instantiation of the imprecise training data. Compared to our previous work, we provide a finer characterization of the theoretical properties and relations among the proposed definitions of reduct through Theorems 3.4, 3.5, 3.7 that were previously left as open problems. In Section 4, which has been newly added, we also discuss two computational experiments by which we study the empirical performance of the proposed reduct definitions, also in comparison with the state-of-the-art method for dimensionality reduction in superset learning.

2. Background

In this section, we recall basic notions of rough set theory (RST) and belief function theory, which will be used in the main part of the article.

2.1. Rough set theory

Rough set theory has been proposed by Pawlak [46] as a framework for representing and managing uncertain data, and has since been widely applied for various problems in the ML domain (see [4] for a recent overview and survey). We briefly recall the main notions of RST, especially regarding its applications to feature reduction.

A decision table (DT) is a triple $DT = \langle U, Att, t\rangle$ such that U is a universe of objects and Att is a set of *attributes* employed to represent objects in U. Formally, each attribute $a \in Att$ is a function $a : U \to V_a$, where V_a is the domain of values of a. Moreover, $t \notin Att$ is a distinguished *decision* attribute, which represents the target decision (also labeling or annotation) associated with each object in the universe. We say that DT is *inconsistent* if the following holds: $\exists x_1, x_2 \in U, \forall a \in Att, a(x_1) = a(x_2)$ and $t(x_1) \neq t(x_2)$.

Given $B \subseteq Att$, we can define the *indiscernibility relation* with respect to B as xI_Bx' iff $\forall a \in B$, $a(x') = a(x)$. Clearly, it is an equivalence relation that partitions the universe U in equivalence classes, also called *granules of information*, $[x]_B$. Then, the *indiscernibility partition* is denoted as $\pi_B = \{[x]_B \mid x \in U\}$.

We say that $B \subseteq Att$ is a *decision reduct* for DT if $\pi_B \leq \pi_t$ (where the order $\leq$ is the refinement order for partitions, that is, π_t is a coarsening of π_B) and there is no $C \subsetneq B$ such that $\pi_C \leq \pi_t$. Then, evidently, a reduct of a decision table DT represents a set of non-redundant and necessary features to represent the information in DT. We say that a reduct R is *minimal* if it is among the smallest (with respect to cardinality) reducts.

Given $B \subseteq Att$ and a set $S \subseteq U$, a *rough approximation* of S (with respect to B) is defined as the pair $B(S) = \langle l_B(S), u_B(S)\rangle$, where $l_B(S) = \bigcup\{[x]_B \mid [x]_B \subseteq S\}$ is the *lower approximation* of S, and $u_B(S) = \bigcup\{[x]_B \mid [x]_B \cap S \neq \emptyset\}$ is the corresponding *upper approximation*.

Finally, given $B \subseteq Att$, the *generalized decision* with respect to B for an object $x \in U$ is defined as $\delta_B(x) = \{t(x') \mid x' \in [x]_B\}$. Notably, if DT is consistent and B is a reduct, then $\delta_B(x) = \{t(x)\}$ for all $x \in U$.

We notice that in the RST literature, there exist several definitions of reduct that, while equivalent on consistent DTs, are generally non-equivalent for inconsistent ones. We refer the reader to [55] for an overview of such a list and a study of their dependencies, while here we report two specific definitions that are useful for the following:

Definition 2.1. $B \subset Att$ is a δ-reduct if $\forall x \in U,\ \ \delta_B(x) = \delta_{Att}(x)$.

Definition 2.2. $B \subset Att$ is μ-reduct if $\forall x \in U, \forall v \in V_t,\ \ Pr(v|[x]_B) = Pr(v|[x]_{Att})$, where

$$Pr(v|[x]_B) = \frac{|\{x' \in [x]_B : t(x') = v\}|}{|[x]_B|}.$$

Further, we recall the following result:

Theorem 2.1. *Let DT be a decision table. Then, every μ-reduct of DT is also a δ-reduct of DT, but not vice versa.*

We further notice that, given a decision table, the problem of finding the minimal reduct is in general NP-hard (by reduction to the *Shortest Implicant* problem [53,59]).

2.2. Belief function theory

Belief Function Theory (BFT), also known as Dempster-Shafer theory (DST) or Evidence theory (ET), has originally been introduced by Dempster in [10] and subsequently formalized by Shafer in [50] as a generalization of probability theory (although this interpretation has been disputed [47]). The starting point is a *frame of discernment* X, which represents all possible states of a system under study, together with a *basic belief assignment* (bba) $m : 2^X \to [0, 1]$, such that $m(\emptyset) = 0$ and $\sum_{A \in 2^X} m(A) = 1$. From this bba, a pair of functions, called respectively *belief* and *plausibility*, can be defined as follows:

$$Bel_m(A) = \sum_{B:B \subseteq A} m(B) \tag{1}$$

$$Pl_m(A) = \sum_{B:B \cap A \neq \emptyset} m(B) \tag{2}$$

As can be seen from these definitions, there is a clear correspondence between belief functions (resp., plausibility functions) and lower approximations (resp., upper approximations) in RST: this connection has been first established in [63], in which the authors showed that every belief function can be derived from a corresponding (generalized) decision table. More recently, the connection between BFT and RST have been investigated from both the theoretical point of view, for example in [68], where the authors provide a characterization of belief functions in terms of lower and upper approximation operators, and in [65], where the author discusses a novel approach to decision-theoretic rough sets based on BFT; and also from the application point of view: in [67] the authors propose an algorithm for feature reduction based on BFT in the setting of Pythagorean fuzzy rough approximation spaces; while in [7] the authors propose an algorithm to induce weighted decision rules based on RST and BFT.

Starting from a bba, a probability distribution, called *pignistic probability*, can be obtained [57]:

$$P^m_{Bet}(x) = \sum_{A:x \in A} \frac{m(A)}{|A|} \tag{3}$$

Finally, we recall that appropriate generalizations of information-theoretic concepts [51], specifically the concept of *entropy* (which was also proposed to generalize the definition of reducts in RST [54]), have been defined for evidence theory. These include measures of non-specificity [1,16], measures of conflict or dissonance [26,35,56,62], and measures of total uncertainty [2,25,34]: see [34] for a comprehensive review on generalizations of entropy for evidence theory. Most relevantly for the purposes of this article, we recall the definition of *aggregate uncertainty* [25]:

$$AU(m) = \max_{p \in \mathcal{P}(m)} H_p(X)\,, \tag{4}$$

where $\mathcal{P}(m)$ is the set of probability distributions p such that $Bel_m \leq p \leq Pl_m$, and $H_p(X) = -\sum_{x \in X} p(x) log_2 p(x)$ the Shannon entropy of p. While this measure is not compatible with Dempster combination rule (see [34]; note, however, that we do not rely on Dempster combination rule in this paper), it complies with the generalized risk minimization approach [28] to superset learning and, more in particular, with the pessimistic loss approach to generalized risk minimization [22, 23,31]. Another relevant approach is the *normalized pignistic entropy* (see [36] for the non-normalized definition)

$$H_{Bet}(m) = \frac{H(P^m_{Bet})}{H(\hat{p}_m)}\,, \tag{5}$$

where $\hat{p}_m$ is the probability distribution that is uniform on the support of $P^m_{Bet}(x)$, i.e., on the set of elements $\{x \mid P^m_{Bet}(x) > 0\}$. Similarly to the AU, also the pignistic entropy is not compatible with Dempster combination rule, but has the advantage of being efficiently computable.

2.3. Superset learning

As already mentioned in the introduction, *superset learning* is a specific type of *weakly supervised learning* and, more precisely, a specific type of the *learning from soft labels* [8,11,13,48] task. While in learning from soft labels the partial labels are represented through general belief functions [11], in the case of superset learning each instance (or object) $x \in U$, where U is a data set (e.g., the training data in a machine learning setting), is annotated with a collection of labels $S \subseteq \mathcal{Y}$ (i.e., in BFT terminology, the partial labels are represented by belief functions with a single focal set). The common interpretation of S is in terms of a set of candidates of an underlying ground-truth: There is a true label y, which is not precisely known, but which is known to be an element of S. In other words, S is a superset of y, hence the name "superset learning".

As an illustration, consider the famous Iris data, where the objects are iris plants characterized by four attributes $a_1, \ldots, a_4$ (sepal length, sepal width, petal length, petal width). Moreover, each plant belongs to either of the three categories Setosa, Versicolor, Virginica. Thus, a labeled instance in a data set might be given by $(6.1, 2.9, 4.7, 1.4, \text{Versicolor})$. Now, imagine that a botanist who is responsible for the categorization is not entirely certain about the type of a plant

with features $x = (6.1, 2.9, 4.7, 1.4)$, but can at least exclude Setosa as an option. She could then label the instance with $S = \{\text{Versicolor}, \text{Virginica}\}$.

In spite of the ambiguous, set-valued training data, the goal that is commonly considered in superset learning is to induce a unique model, i.e., a map $h : \mathcal{X} \to \mathcal{Y}$ that generalizes beyond the training data and can be used to make predictions $h(x) \in \mathcal{Y}$ for any new query instance $x \in \mathcal{X}$. In one way or the other, this requires the "disambiguation" of the training data. To this end, various methods and algorithmic approaches have been proposed in the literature, for example based on maximum likelihood estimation [8,13,33,40,48], generalizations of empirical risk minimization [28,30,31], convex optimization [9,18], and instance-based approaches [11,29,66]. In [39], superset learning has been studied from a theoretical perspective in the framework of PAC learning.

Superset learning has mostly been studied for classification problems so far, while other (related) machine learning tasks have been considered much less. This also includes feature selection, despite its important influence on model complexity, generalization performance, and transparency of learning algorithms. Indeed, while many works have studied feature selection and dimensionality reduction in the setting of semi-supervised learning [3,52], which is actually a special case of superset learning, to the knowledge of the authors, the only work focusing on the more general setting of superset learning is the DELIN algorithm proposed in [61]. Compared to the method put forward in this article, we note two main differences. First, being based on Linear Discriminant Analysis (LDA), DELIN relies on specific assumptions regarding the statistical distribution of the data, whereas our method (based on Rough Set Theory) is completely non-parametric. Second, DELIN is a dimensionality reduction algorithm, which means that it constructs a new set of attributes that is not (in general) a subset of the original one. By contrast, our approach is a feature selection algorithm, which selects a subset of the original set of attributes. In Section 4, we will provide an experimental comparison of the two methods.

As for the notation and connection to RST, it should be clear that the attribute y and its domain $\mathcal{Y}$ in superset learning play the role, respectively, of the decision attribute t and its domain V_t in RST. As an aside, let us note that the information provided in superset learning may also be interpreted in a different way, which provides an alternative motivation for the superset extension of decision tables in general and the search for reducts of such tables in particular. As explained above, the superset extension is mostly motivated by the assumption of imprecise labeling: The value of the decision attribute is not known precisely but only characterized in terms of a set of possible candidates. As will be seen further below, finding a reduct is then supposed to help disambiguate the data, i.e., figuring out the most plausible among the candidates. Instead of this "don't know" interpretation, a superset S can also be given a "don't care" interpretation: In a certain context characterized by x, all decisions in S are sufficiently good, or "satisficing" in the sense of March and Simon [42]. A reduct can then be considered as a maximally simple (least cognitively demanding) yet satisficing decision rule.

3. Superset decision tables and reducts

In this section, we extend some key concepts of rough set theory to the setting of superset learning.

3.1. Superset decision tables

In superset learning, an object $x \in U$ is not necessarily assigned a single annotation $t(x) \in V_t$, but instead a set S of candidate annotations, one of which is assumed to be the true annotation associated with x. To model this idea in terms of RST, we generalize the definition of a decision table as follows.

Definition 3.1. A *superset decision table* (SDT) is a tuple $SDT = \langle U, Att, t, d \rangle$, where $\langle U, Att, t \rangle$ is a decision table, i.e.:

- U is a universe of objects of interest;
- Att is a set of attributes (or features);
- t is the (real) decision attribute (whose value, in general, is not known);
- $d \notin Att$ is a candidate decision attribute, that is, a set-valued map $d : U \to \mathcal{P}(V_t)$ such that the *superset property* holds: $t(x) \in d(x)$ for all $x \in U$.

The intuitive meaning of the set-valued information d is that, if $|d(x)| > 1$ for some $x \in U$, then the real decision associated with x (i.e., $t(x)$) is not known precisely, but is known to be in $d(x)$. Notice that Definition 3.1 is a proper generalization of decision tables: if $|d(x)| = 1$ for all $x \in U$, then we have a standard decision table.

Remark 3.1. In Definition 3.1, a set-valued decision attribute is modeled as a function $d : U \to \mathcal{P}(V_t)$. While this mapping is formally well-defined for a concrete decision table, let us mention that, strictly speaking, there is no functional dependency between x and $d(x)$. In fact, $d(x)$ is not considered as a property of x, but rather represents *information* about a property of x, namely the underlying decision attribute $t(x)$. As such, it reflects the epistemic state of the decision maker.

A SDT can be associated with a collection of compatible (standard) decision tables, which we call instantiations of the SDT.

Table 1
An example of superset decision table.

	a_1	a_2	a_3	a_4	d
x_1	0	0	0	0	0
x_2	0	0	0	1	$\{0,1\}$
x_3	0	1	1	0	0
x_4	0	1	1	1	$\{0,1\}$
x_5	0	1	0	1	1
x_6	0	1	0	0	$\{0,1\}$

Definition 3.2. An *instantiation* of a SDT $\langle U, Att, t, d\rangle$ is a standard DT $I = \langle U, Att, t_I\rangle$ such that $t_I(x) \in d(x)$ for all $x \in U$. The set of instantiations of SDT is denoted $\mathcal{I}(SDT)$.

The notion of inconsistency of a SDT has to reflect this richness. The following definition reflects the idea that no instantiations are consistent.

Definition 3.3. For $B \subset Att$, the SDT is B-inconsistent if

$$\exists x_1, x_2 \in U, \forall a \in B, a(x_1) = a(x_2) \text{ and } d(x_1) \cap d(x_2) = \emptyset. \tag{6}$$

We call such a pair x_1, x_2 inconsistent. If condition (6) is not satisfied, the SDT is B-consistent.

Thus, inconsistency implies the existence of (at least) two indiscernible objects with non-overlapping superset decisions. We say that an instantiation I is *consistent with a SDT S* (short, is consistent) if the following holds for all x_1, x_2: if x_1, x_2 are consistent in S, then they are also consistent in I.

3.2. Superset reducts

Learning from superset data is closely connected to the idea of *data disambiguation* in the sense of figuring out the most plausible instantiation of the set-valued training data [27,31]. But what makes one instantiation more plausible than another one? One approach originally proposed in [29] refers to the principle of simplicity in the spirit of *Occam's razor* (which can be given a theoretical justification in terms of *Kolmogorov complexity* [38]): An instantiation that can be explained by a simple model is more plausible than an instantiation that requires a complex model. In the context of RST-based data analysis, a natural measure of model complexity is the size of the reduct. This leads us to the following definition.

Definition 3.4. A set of attributes $R \subseteq Att$ is a (consistent) *superset reduct* if there exists a (consistent) instantiation $I = \langle U, Att, t_I\rangle$ such that R is a reduct for I and there is no other (consistent) instantiation $I' = \langle U, Att, t_{I'}\rangle$ with reduct $R' \subset R$. We denote with R_{super} (resp., R^c_{super}) the set of superset reducts (resp., consistent superset reducts). A *minimum description length (MDL) instantiation* is one of the (consistent) instantiations of SDT that admits a reduct of minimum size compared to all the reducts of all possible (consistent) instantiations. We will call the corresponding reducts *MDL reducts*.

First of all, in order to clarify these concepts, we show a brief example.

Example 3.1. Consider the superset decision table

$$SDT = \left\langle U = \{x_1, ..., x_6\}, A = \{a_1, a_2, a_3, a_4\}, d \right\rangle$$

given in Table 1.

It is easy to observe that the SDT admits 8 possible instantiations:

- I_1 s.t. $t_{I_1}(x_1) = t_{I_1}(x_2) = t_{I_1}(x_3) = t_{I_1}(x_4) = t_{I_1}(x_6) = 0$ and $t_{I_1}(x_5) = 1$;
- I_2 s.t. $t_{I_2}(x_1) = t_{I_2}(x_2) = t_{I_2}(x_3) = t_{I_2}(x_4) = 0$ and $t_{I_2}(x_5) = t_{I_2}(x_6) = 1$;
- I_3 s.t. $t_{I_3}(x_1) = t_{I_3}(x_2) = t_{I_3}(x_3) = t_{I_3}(x_6) = 0$ and $t_{I_3}(x_5) = t_{I_3}(x_4) = 1$;
- I_4 s.t. $t_{I_4}(x_1) = t_{I_4}(x_3) = t_{I_4}(x_4) = t_{I_4}(x_6) = 0$ and $t_{I_4}(x_5) = t_{I_4}(x_2) = 1$;
- I_5 s.t. $t_{I_5}(x_1) = t_{I_5}(x_2) = t_{I_5}(x_3) = 0$ and $t_{I_5}(x_4) = t_{I_5}(x_5) = t_{I_5}(x_6) = 1$;
- I_6 s.t. $t_{I_6}(x_1) = t_{I_6}(x_3) = t_{I_6}(x_4) = 0$ and $t_{I_6}(x_2) = t_{I_6}(x_5) = t_{I_6}(x_6) = 1$;
- I_7 s.t. $t_{I_7}(x_1) = t_{I_7}(x_3) = t_{I_7}(x_6) = 0$ and $t_{I_7}(x_2) = t_{I_7}(x_4) = t_{I_7}(x_5) = 1$;
- I_8 s.t. $t_{I_8}(x_1) = t_{I_8}(x_3) = 0$ and $t_{I_8}(x_2) = t_{I_8}(x_4) = t_{I_8}(x_5) = t_{I_8}(x_6) = 1$.

All of the instantiations are Att-consistent, since no two $x, x' \in U$ are associated with the same representation. It is easy to observe that the single shortest reduct among all instantiations is $R = \{a_4\}$, with corresponding instantiation I_7: thus I_7

is a MDL instantiation and $\{a_4\}$ is the unique MDL reduct (and thus also a superset reduct). The SDT also admits another superset reduct, namely $\{a_2, a_3\}$ (with corresponding instantiation I_2).

Then, we briefly comment on the fact that the definition of MDL reduct generalizes the standard definition of (minimal) reduct. Indeed, in a classical decision table, there is only one possible instantiation, hence the MDL reduct is exactly (one of) the minimal reducts of the decision table. Further, if we denote by R_{MDL} the set of MDL reducts, and by R^c_{MDL} the set of consistent MDL reducts (i.e., the MDL reducts corresponding only to consistent instantiations), then we can prove the following result:

Theorem 3.1. *$R_{MDL} \subseteq R_{super}$ and $R^c_{MDL} \subseteq R^c_{super}$. Furthermore, if $R \in R^c_{MDL}$ (resp., R^c_{super}), then $\exists R' \in R_{MDL}$ (resp., R_{super}) s.t. $R' \subseteq R$.*

Proof. If R is a consistent MDL reduct, then by definition it is also a consistent superset reduct, thus $R^c_{MDL} \subsetneq R^c_{super}$. The same holds for R_{MDL}, R_{super}.

As regard the second pair of statement, it is obviously the case that if we consider also inconsistent instantiations then the set of superset super-reducts (denoted with SR_{super}) contains the set of superset super-reducts that we would obtain were we to consider only consistent instantiations (denoted SR^c_{super}): this implies that if $R \in SR^c_{super}$ then $R \in SR_{super}$ and the result easily follows. □

Algorithm 1 The brute-force algorithm for finding MDL reducts of a superset decision table S.

```
procedure BRUTE-FORCE-MDL-REDUCT(S: superset decision table)
    reds ← ∅
    l ← ∞
    ists ← enumerate-instantiations(S)
    for all i ∈ ists do
        tmp-reds ← find-shortest-reducts(i)
        len ← |red| where red ∈ tmp-reds
        if len < l then
            reds ← tmp-reds
            l ← len
        else if len = l then
            reds ← reds ∪ tmp-reds
        end if
    end for
    return reds                                   ▷ The MDL reducts for S
end procedure
```

An algorithmic solution to the problem of finding the MDL reduct for an SDT can be given as a brute-force algorithm, which computes the reducts of all the possible instantiations, see Algorithm 1. It is easy to see that the worst case runtime complexity of this algorithm is exponential in the size of the input. Unfortunately, it is unlikely that an asymptotically more efficient algorithm exists. Indeed, if we consider the problem of finding *any* MDL reduct, then the number of instantiations of S is, in the general case, exponential in the number of objects, and for each such instantiation one should find the shortest reduct for the corresponding decision table, which is known to be NP-hard. Interestingly, we can prove that the following decision problem (i.e., does there exist a superset reduct of size $\leq k$?) related to finding MDL-Reducts is in NP^{NP} (i.e., the class of problems that can be checked in polynomial time with access to an oracle for SAT).

Theorem 3.2. *Let $MDL\text{-}Reduct$ be the problem of deciding if, given an SDT S and $k \in \mathbb{N}^+$, the MDL reducts of S are of size $\leq k$. Then, $MDL\text{-}Reduct$ is in NP^{NP}.*

Proof. We need to show that there is an algorithm for verifying instances of $MDL\text{-}Reduct$ whose runtime is polynomial given access to an oracle for an NP-complete problem. Indeed, a certificate can be given by an instantiation I (whose size is clearly polynomial in the size of the input SDT) together with a minimal reduct r for I s.t. $|r| \leq k$. Verifying whether r is a minimal reduct for I can then be done in polynomial time with an oracle for NP, from which the result follows. □

From the above proof we can observe that the pair (I, r), used as a certificate, only requires that r is a reduct of I, which means that in general it is a superset super-reduct of S and not necessarily also a superset reduct.

While heuristics could be applied to speed up the computation of reducts [58] (specifically, to reduce the complexity of the *find-shortest-reducts* step in Algorithm 1) the approach described in Algorithm 1 still requires enumerating all the possible instantiations. Therefore, in the following section, we propose two alternative definitions of reduct in order to reduce the computational costs.

3.3. Entropy reducts

We begin with a definition based on the notion of entropy [54], which simplifies the complexity of finding a reduct for an SDT. Indeed, while finding Superset and MDL reducts requires to enumerate all possible instantiations of a given SDT (which, in general, are exponentially many in the size of the SDT), the two alternative notions of entropy-based reducts that we propose in this Section do not require such an enumeration.

Given a decision d, we can associate with it a pair of belief and plausibility functions. Let $v \in V_t$ and $[x]_B$ for $B \subseteq Att$ an equivalence class, i.e. $[x]_B = \{x' \in U : \forall a \in B,\ \ a(x') = a(x)\}$. Then:

$$Bel_S(v|[x]_B) = \frac{|\{x' \in [x]_B : d(x') = \{v\}\}|}{|[x]_B|}$$
$$Pl_S(v|[x]_B) = \frac{|\{x' \in [x]_B : v \in d(x')\}|}{|[x]_B|}$$

For each $W \subseteq V_t$, the corresponding basic belief assignment is defined as

$$m(W|[x]_B) = \frac{|\{x' \in [x]_B : d(x') = W\}|}{|[x]_B|}\,. \tag{7}$$

Given this setting, we now consider two different entropies. The first one is the pignistic entropy $H_{Bet}(m)$ as defined in (5). As regards the second definition, we will not directly employ the AU measure (see equation (4)). This measure, in fact, corresponds to a quantification of the degree of conflict in the bba m, which is not appropriate in our context, as it would imply finding an instantiation which is maximally inconsistent. We thus consider a modification of the AU measure called *Optimistic Aggregate Uncertainty* (OAU), which is consistent with the optimistic approach to generalized risk minimization [28,30,31]. This measure, which has already been studied in the context of evidence theory [1], superset decision tree learning [29] and soft clustering [5], is defined as follows:

$$OAU(m) = \min_{p \in \mathcal{P}(m)} H_p(X)\,, \tag{8}$$

where m is a bba, and H is the Shannon entropy (see Section 2).

We now show how these two entropies can be defined for a given SDT. Let $SDT = \langle U, Att, t, d\rangle$ be an SDT, $B \subseteq Att$ be a set of attributes and denote by $IND_B = \{[x]_B : x \in U\}$ the collection of equivalence classes (granules) determined by B. Let $d_{[x]_B}$ be the restriction of d on the equivalence class $[x]_B$, that is $d_{[x]_B} = \{d(x') : x' \in [x]_B\}$. The H_{Bet} and OAU entropy of d, conditional on B, are defined as

$$\begin{aligned} H_{Bet}(d|B) &= \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} H_{Bet}(d_{[x]_B}) \\ &= \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} \frac{H(P^m_{Bet}(d_{[x]_B}))}{H(\hat{p}_m(d_{[x]_B}))} \\ &= \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} \frac{\sum_{v \in d_{[x]_B}} P^{m(\cdot|[x]_B)}_{Bet}(v) * log(P^{m(\cdot|[x]_B)}_{Bet}(v))}{\sum_{v \in d_{[x]_B}} \frac{1}{|d_{[x]_B}|} * log(\frac{1}{|d_{[x]_B}|})} \end{aligned} \tag{9}$$

$$\begin{aligned} OAU(d|B) &= \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} OAU(d_{[x]_B}) \\ &= \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} min_{I \in \mathcal{I}(SDT)} \sum_{v \in \delta^I_B(x)} Pr(v|[x]^I_B) * log(\frac{1}{Pr(v|[x]^I_B)}) \end{aligned} \tag{10}$$

where $m(\cdot|[x]_B)$ is the bba determined by the granule $[x]_B$ (see Eq. (7)), $P^{m(\cdot|[x]_B)}_{Bet}$ is the pignistic probability distribution (see Section 2), $[x]^I_B$ is the granule of x determined by $B \subset Att$ in the instantiation $I \in \mathcal{I}(SDT)$, δ^I_B is the generalized decision w.r.t. B for the instantiation $I \in \mathcal{I}(SDT)$ (see Section 2), and $Pr(v|[x]^I_B)$ is the probability of class label v in the granule of x determined by $B \subset Att$ in instantiation I (see the definition of μ-reduct in Section 2).

Definition 3.5. We say that $B \subseteq Att$ is

- an OAU super-reduct (resp., H_{Bet} super-reduct) if $OAU(d \mid B) \leq OAU(d \mid Att)$ (resp., $H_{Bet}(d \mid B) \leq H_{Bet}(d \mid Att)$);
- an OAU reduct (resp., H_{Bet} reduct) if no proper subset of B is also a super-reduct.

As a further heuristic, we introduce approximate reducts as follows.

Definition 3.6. We say that $B \subseteq Att$ is

- an OAU ϵ-approximate super-reduct (resp., H_{Bet} ϵ-approximate super-reduct), with $\epsilon \in [0,1)$, if $OAU(d \mid B) \leq OAU(d \mid Att) - log_2(1-\epsilon)$ (resp., $H_{Bet}(d \mid B) \leq H_{Bet}(d \mid Att) - log_2(1-\epsilon)$);
- an OAU ϵ-approximate reduct (resp., H_{Bet} ϵ-approximate reduct) if no proper subset of B is also an ϵ-approximate super-reduct.

It is easy to observe that both OAU and H_{Bet} naturally define two families of instantiations of the underlying SDT. Indeed, let B be an OAU reduct and let $[x]_B$ be one of the granules with respect to an OAU reduct. Then, a *OAU instantiation* is any instantiation $I_{OAU} \in \mathcal{I}(SDT)$ s.t.:

$$dec_{OAU}([x]_B) = \arg\max_{v \in V_t} \Big\{ Pr(v|[x]^I_B) : I \in \{ \arg\min_{J \in \mathcal{I}(SDT)} \sum_{v \in \delta^J_B(x)} Pr(v|[x]^J_B) * log(\frac{1}{Pr(v|[x]^J_B)}\} \Big\}. \tag{11}$$

That is, an OAU reduct determines an instantiation in which each object is assigned to the most probable among the classes, under the probability distribution which corresponds to the minimum value of entropy.

Similarly, a H_{Bet} *instantiation* with respect to $[x]_B$ is given by

$$dec_{H_{Bet}}([x]_B) = \arg\max_{v \in V_t} P^{m(\cdot|[x]_B)}_{Bet}(v) \tag{12}$$

We note that, in general, neither $dec_{OAU}([x]_B)$ nor $dec_{H_{Bet}}([x]_B)$ are unique: for the case of $dec_{OAU(B)}([x]_B)$ there may exist two instantiations $I, I' \in \mathcal{I}(SDT)$ with corresponding probability distributions p, p' (over the labels $v \in V_t$) s.t. both $p, p' \in \arg\min_{p \in P_m} H_p(X)$; while for the case of $dec_{H_{Bet}}([x]_B)$ there may be two classes $v', v'' \in V_t$ s.t.

$$P^{m(\cdot|[x]_B)}_{Bet}(v') = P^{m(\cdot|[x]_B)}_{Bet}(v'') = max_{v \in V_t} P^{m(\cdot|[x]_B)}_{Bet}(v).$$

The following example shows, for a simple SDT, the OAU reducts, MDL reducts, and H_{Bet} reducts and their relationships.

Example 3.2. Consider the superset decision table

$$SDT = \langle U = \{x_1, ..., x_6\}, A = \{a_1, a_2, a_3, a_4\}, d \rangle$$

given in Table 1 and described in Example 3.1. We have that, for $B = \{a_2, a_3\}$:

$$OAU(d \mid A) = OAU(d \mid B) = 0.$$

Thus, B is an OAU reduct of SDT, as $OAU(d \mid a_2) = OAU(d \mid a_3) > 0$. It can easily be seen that B admits only a single OAU instantiation, which is given by $\{a_2, a_3\}$ is $dec_{a_2,a_3}(\{x_1, x_2\}) = dec_{a_2,a_3}(\{x_3, x_4\}) = 0$, $dec_{a_2,a_3}(\{x_5, x_6\}) = 1$. Indeed, every other possible assignment of class labels to the equivalence classes determined by B would result in a greater entropy.

Note that $\{a_4\}$ is also an OAU reduct and also in this case there exists a single corresponding OAU instantiation: this is given by $\{a_4\}$ is $dec_{a_4}(\{x_1, x_3, x_6\}) = 0$, $dec_{a_4}(\{x_2, x_4, x_5\}) = 1$.

On the other hand, $H_{Bet}(d \mid A) = \frac{1}{2}$, while $H_{Bet}(d \mid \{a_2, a_3\}) = 0.81$. Therefore, $\{a_2, a_3\}$ is not an H_{Bet} reduct. Notice that, in this case, there are no H_{Bet} reducts (excluding A). However, it can easily be seen that $\{a_2, a_3\}$ is an H_{Bet} approximate reduct when $\epsilon \geq 0.20$. We note that there exists 8 different H_{Bet} instantiations corresponding to the H_{Bet} reduct A: in all these instantiations we have that $dec_A(x_1) = dec_A(x_3) = 0$ and $dec_A(x_5) = 1$, while we have a different instantiation for each of the possible class assignments for the remaining objects.

As shown in Example 3.1, the unique MDL reduct is $\{a_4\}$, with corresponding MDL instantiation $dec_{MDL}(\{x_1, x_3, x_6\}) = 0$, $dec_{MDL}(\{x_2, x_4, x_5\}) = 1$. Thus, in this case, the MDL reduct is equivalent to one of the OAU reducts.

We note that also the other possible superset reduct (i.e. $\{a_2, a_3\}$, as shown in Example 3.1) is an OAU reduct: as we'll show in the next Section, this is a general property of OAU reducts.

Before studying the formal properties of the proposed entropy reducts, we observe that the computation of H_{Bet} and OAU entropies do not require one to enumerate all instantiations of a SDT, and can be performed in polynomial time. This is clearly immediate for the computation of H_{Bet}:

Proposition 3.1. *H_{Bet} can be computed in polynomial time, without enumerating the instantiations $I \in \mathcal{I}(SDT)$.*

Proof. In Equation (9) we only perform $|IND_B||d_{[x]_B}| \in O(|U||V_t|)$ operations, and there is clearly no dependency on $|\mathcal{I}(SDT)|$. □

The analogous result for the computation of the OAU entropy is less immediate (indeed, in Eq. (10) we need to solve a minimization problem over $\mathcal{I}(SDT)$), and rests on a previous characterization of this uncertainty measure [5,29]:

Proposition 3.2. *OAU can be computed in polynomial time, without enumerating the instantiations $I \in \mathcal{I}(SDT)$.*

Proof. The OAU entropy can be computed efficiently through the P-LLE (Polynomial Lower Logical Entropy) algorithm proposed in [5]: the time complexity of this procedure is $\Omega(|V_t|)$ and $O(|U||V_t| * log|V_t|)$, and has no dependency on $|\mathcal{I}(SDT)|$. □

These properties imply that the computation of H_{Bet} and OAU reduct does not require one to enumerate the instantiations $I \in \mathcal{I}(SDT)$, and instead the required computations can be performed by directly relying on the statistics in the original SDT: this property will be useful for designing efficient (heuristic) procedures for searching reducts, as we show in Section 3.4, and is similarly useful for computing OAU and H_{Bet} instantiations. Indeed, it can easily be seen that, as a consequence of Propositions 3.1 and 3.2, the OAU (resp. H_{Bet}) instantiations, corresponding to a given OAU (resp. H_{Bet}) reduct, can be computed in polynomial time.

3.4. Properties of reducts

In this section, we study the properties of, and relationships among, the different definitions of reducts on superset decision tables. In Example 3.2, it is shown that the MDL reduct is one of the OAU reducts. Indeed, we can prove that this holds in general.

Theorem 3.3. *Let R be an MDL reduct whose MDL instantiation is consistent (that is, $R \in R^c_{MDL}$). Then R is also an OAU reduct.*

Proof. As the instantiation corresponding to R is consistent, $OAU(d \mid R) = 0$. Thus, R is an OAU reduct. □

Corollary 3.1. *Finding the minimal OAU reduct for a consistent SDT is NP-hard.*

Proof. As any MDL reduct of a consistent SDT is also an OAU reduct and MDL reducts are by definition minimal, the complexity of finding a minimal OAU reduct is equivalent to that of finding MDL reducts. □

More in general, if we consider a consistent SDT, we can prove that the collection of OAU reducts and (consistent) superset reducts are equivalent, that is, the following result holds.

Theorem 3.4. *Let S be a consistent SDT, then $R^c_{super} = R_{OAU}$, that is each OAU reduct is a consistent superset reduct (and vice-versa). Furthermore, for each $r \in R_{OAU}$ there exists $r' \in R_{super}$ (i.e. a superset reduct) s.t. $r' \subseteq r$, that is each OAU reduct is a superset super-reduct.*

Proof. Let $r \in R^c_{super}$, then its instantiation is consistent and hence $OAU(d|r) = 0$, thus $r \in R_{OAU}$. Conversely, let $r \in R_{OAU}$ and notice that every OAU instantiation (i.e., an instantiation s.t. $\forall [x]_r, d([x]_r) = dec_{OAU(r)}([x]_r)$) is necessarily consistent (as $OAU(d|r) = 0$). Hence, r is a reduct of a consistent instantiation, thus $r \in R^c_{super}$.

For the last part of the theorem, it suffices to notice that no inconsistent instantiation can be an OAU instantiation, and that each consistent superset reduct is also a (not necessarily consistent) superset super-reduct (by Theorem 3.1). The result follows. □

In inconsistent SDTs, only the last part of the previous theorem holds, as shown by the following theorem.

Theorem 3.5. *Let S be an inconsistent SDT. Then, for each $r \in R_{OAU}$, there exists $r' \in R_{super}$ s.t. $r' \subseteq r$.*

Proof. We can notice that each $r \in R_{OAU}$ corresponds to a OAU instantiation, whose Shannon entropy (by definition of the OAU measure) is minimal with respect to all possible instantiations. Thus, R_{OAU} is the collection of superset super-reducts whose corresponding instantiations have minimal entropy. Further, note that there may be $r' \in R_{super}$ s.t. $r' \subseteq r$. □

On the other hand, as shown in Example 3.2, the relationship between MDL reducts (or OAU reducts) and H_{Bet} reducts is more complex as, in general, an OAU reduct is not necessarily a H_{Bet} reduct. In particular, one could be interested in knowing whether an H_{Bet} (smaller than the whole set of attributes Att) exists and whether there exists a H_{Bet} reduct

which is able to disambiguate objects that are not disambiguated when taking in consideration the full set of attributes Att. The following two results provide a characterization in the binary (i.e., $V_t = \{0, 1\}$), consistent case.

Theorem 3.6. *Let $B \subseteq Att$ be a set of attributes, $[x_1]_{Att}, [x_2]_{Att}$ be two distinct equivalence classes (i.e., $[x_1]_{Att} \cap [x_2]_{Att} = \emptyset$) that are merged by B (i.e., $[x_1]_B = [x_1]_{Att} \cup [x_2]_{Att}$), that are consistent and such that $|[x_1]_{Att}| = n_1 + m_1$, $|[x_2]_{Att}| = n_2 + m_2$, where the n_1 (resp., n_2) objects are such that $|d(x)| = 1$ and the m_1 (resp., m_2) objects are such that $|d(x)| = 2$. Then, $H_{Bet}(d \mid B) \geq H_{Bet}(d \mid Att)$, with equality holding iff one of the following two holds:*

1. *$m_1 = m_2 = 0$ and $n_1, n_2 > 0$;*
2. *$m_1, m_2 > 0$ and $n_1 \geq 0$, $n_2 = \frac{m_2 n_1}{m_1}$ (and, symmetrically when changing n_1, n_2).*

Proof. A sufficient and necessary condition for $H_{Bet}(d \mid B) \geq H_{Bet}(d \mid Att)$ is:

$$\frac{n_1 + \frac{m_1+m_2}{2} + n_2}{n_1 + m_1 + n_2 + m_2} \geq \max\left\{\frac{n_1 + \frac{m_1}{2}}{n_1 + m_1}, \frac{\frac{m_2}{2} + n_2}{n_2 + m_2}\right\} \tag{13}$$

under the constraints $n_1, n_2, m_1, m_2 \geq 0$, as the satisfaction of this inequality implies that the probability is more peaked on a single alternative. The integer solutions for this inequality provide the statement of the theorem. Further, one can see that the strict inequality is not achievable. □

Corollary 3.2. *On a binary consistent SDT, a subset $B \subseteq Att$ is a H_{Bet} reduct iff, whenever it merges a pair of equivalence classes, the conditions expressed in Theorem 3.6 are satisfied.*

Notably, these two results also provide an answer to the second question, that is, whether an H_{Bet} reduct can disambiguate instances that are not disambiguated when considering the whole attribute set Att. Indeed, Theorem 3.6 provides sufficient conditions for this property and shows that, in the binary case, disambiguation is possible only when at least one of the equivalence classes (w.r.t. Att), that are merged by the reduct, is already disambiguated.

As we described in the statement of Theorem 3.6, our result applies only to the binary case: indeed, the general n-ary case is much more complex and, in such cases, disambiguation could happen also in more general situations. This is shown by the following example.

Example 3.3. Let $SDT = \langle U = \{x_1, ..., x_{10}\}, Att = \{a_1, a_2\}, d\rangle$ such that $\forall i \leq 5$, $d(x_i) = \{0, 1\}$ and $\forall i > 5, d(x_i) = \{1, 2\}$. Then, assuming the equivalence classes determined by Att are $\{x_1, ..., x_5\}, \{x_6, ..., x_{10}\}$, it holds that $H_{Bet}(d \mid Att) = 1$. If we further assume that a_1 determines a single equivalence class U, then it is easy to observe that $H_{Bet}(d \mid a_1) < 0.95 < H_{Bet}(d \mid Att)$ and hence a_1 is a H_{Bet} reduct.

Note that the conditions expressed in Theorem 3.6 are satisfied for the set of all attributes Att, but Att is not a H_{Bet} reduct: indeed, if we consider the equivalence classes determined by Att, then $n_1 = n_2 = 0$ while $m_1 = m_2 = 5$ and therefore condition 2 in Theorem 3.6 holds. However, as previously shown, Att is not a H_{Bet} reduct.

Furthermore, note that Att is not able to disambiguate, since

$$dec_{H_{Bet}(Att)}([x_1]_{Att}) = \{0, 1\},$$
$$dec_{H_{Bet}(Att)}([x_6]_{Att}) = \{1, 2\}.$$

On the other hand, $dec_{H_{Bet(a_1)}}(x_i) = 1$ for all $x_i \in U$. Notice that, in this case, $\{a_1\}$ would also be an OAU reduct (and hence a MDL reduct, as it is minimal).

On the other hand, regarding the relationships between H_{Bet} reducts and the other families of reducts, it is easy to show that, even on consistent SDTs, the conditions for existence of H_{Bet} reducts (smaller than the whole set of attributes Att) are quite restrictive. Indeed, the following result holds.

Theorem 3.7. *Let S be an SDT and r be an H_{Bet} reduct. Then, there exists $r' \subseteq r$ s.t. r' is an OAU reduct. That is, the collection of H_{Bet} reducts is a sub-collection of the OAU super-reducts.*

Proof. First, let us assume that S is consistent, and let $r \in R_{H_{Bet}}$. Then, since S is consistent, each $[x]_r$ is also consistent and therefore, by definition, $OAU(d|r) = 0$ and r is an OAU super-reduct (but not necessarily also a OAU reduct). Consequently, the result holds for consistent SDTs.

For the inconsistent case, let r be an H_{Bet} reduct, and $\{[x]^i_r\}_i$ be the collection of the equivalence classes w.r.t. r. By definition of H_{Bet} reducts, we have $\sum_i Pr([x]^i_r) \cdot H_{Bet}(d|[x]^i_r) \leq H_{Bet}(d|Att)$. Therefore, for the (weighted) majority of equivalence classes the probability distributions $P_{Bet}(d|[x]^i_r)$ are more peaked (equivalently, less uniform) and, hence, there exists an instantiation I s.t. the probability distributions $P_I(d^I \mid x^i_r)$ are also more peaked. Hence, $OAU(d \mid r) \leq OAU(d \mid Att)$ holds. Notice, however, that this only guarantees that r is a OAU super-reduct, thus the result. □

As we did not find an appropriate generalization of Theorem 3.6 for the general multi-class case, we leave this as an open problem: such a result would be useful to provide general existence conditions for H_{Bet} reducts. Moreover, we also leave as open problem that of finding conditions required for an H_{Bet} to also be an OAU (or MDL) reduct.

Concerning the computational complexity of finding OAU or H_{Bet} reducts, since as we shown in the previous Section, both OAU and H_{Bet} can be computed in polynomial time, the following result holds as a simple consequence of the general hardness result for finding reducts in standard decision tables.

Theorem 3.8. *Finding all OAU (resp. H_{Bet}) reduct is NP-hard.*

Finally, we notice that, while the complexity of finding OAU (resp. H_{Bet}) reducts is still NP-hard, even in the approximate case, these definitions are more amenable to optimization through heuristics, as they employ a quantitative measure of quality for each attribute. Indeed, a simple greedy procedure can be implemented, as shown in Algorithm 2, which obviously has polynomial time complexity, and is guaranteed to find an OAU (resp., H_{Bet}) reduct (albeit not necessarily a minimal one).

Proposition 3.3. *Algorithm 2 returns an OAU (resp. H_{Bet}) reduct in polynomial time. In particular:*

- *The complexity of finding a OAU reduct is $O\left(|Att|^2|U||V_t| * log|V_t|\right)$;*
- *The complexity of finding a H_{Bet} reduct is $O\left(|Att|^2|U||V_t|\right)$*

Proof. That the algorithm returns an OAU (resp. H_{Bet}) reduct is obvious, thus we only need to prove that its complexity is polynomial in the size of the SDT.

Indeed, Algorithm 2 requires a polynomial number of evaluations of the OAU (resp. H_{Bet}) entropy: in particular, the number of such evaluations is $O\left(|Att|^2\right)$. As shown in Propositions 3.1 and 3.2, both OAU and H_{Bet} can be computed in polynomial time, thus the result follows. □

Thus, Algorithm 2 has a linear dependence in the number of objects, a linear (or log-linear, depending on whether H_{Bet} or OAU reducts are searched for) dependence in the number of possible class labels, and a quadratic dependence in the number of conditional attributes: we note that since usually $|V_t| \ll min\{|U|, |Att|\}$, one can assume w.l.o.g. that the complexity of searching reducts is dominated by the leading term among $\{|U|, |Att|^2\}$, and it is thus more or less independent of the number of possible class labels.

Algorithm 2 A heuristic greedy algorithm for finding approximate entropy reducts of a superset decision table S.

procedure HEURISTIC-ENTROPY-REDUCT(S: superset decision table, ϵ: approximation level, $E \in \{OAU, H_{Bet}\}$)
  $red \leftarrow Att$
  $Ent \leftarrow E(d \mid red)$
  $check \leftarrow True$
  while check **do**
    Find $a \in red$ s.t. $\begin{cases} E(d \mid red \setminus \{a\}) \leq E(d \mid Att) - log_2(1-\epsilon) \\ E(d \mid red \setminus \{a\}) \text{ is minimal} \end{cases}$
    if a exists **then**
      $red \leftarrow red \setminus \{a\}$
    else
      $check \leftarrow False$
    end if
  end while
  return red
end procedure

4. Experiments

In this section, we present a series of experimental studies meant to evaluate the different definitions of reduct in superset learning as put forward in this paper, as well as the performance of the proposed algorithms in light of the state-of-the-art in superset dimensionality reduction (DELIN algorithm, see Section 2). More specifically, our experiments are aimed at studying the following aspects:

- Reduct approximation: The ability of the different types of reducts to recover the true reducts (i.e., the reducts w.r.t. the true, but generally unknown, decision attribute t) when varying both the number of objects associated with a set-valued decision and the size of the set-valued decision.
- Predictive Performance: The quality of the selected feature subsets from a machine learning point of view. We measured the latter in terms of the predictive accuracy of a model trained on that subset of features, using a suitable algorithm for superset learning.

We conduct experiments with the following datasets from the UCI repository [20]:

- Iris: 150 objects, 3 classes, 4 attributes
- Boston house prices (Boston): 506 objects, 3 classes, 13 attributes
- Wine: 178 objects, 3 classes, 13 attributes
- Breast Cancer: 569 objects, 2 classes, 30 attributes
- Diabetes: 442 objects, 3 classes, 10 attributes
- Adult Census Income: 48842 objects, 2 classes, 14 attributes
- Abalone: 4177 objects, 15 classes, 8 attributes
- Forest Fires: 517 objects, 5 classes, 13 attributes

For the second experiment, we used the PL-KNN [29] classifier, a simple generalization of the k-nearest neighbor algorithm for superset learning. Of course, more sophisticated methods for superset learning might be used as well, and the choice of the learning methods may clearly influence the results. However, as one advantage of a simple nearest neighbor approach, let us mention that its performance critically depends on the underlying feature representation, which is exactly what we seek to capture. Many other algorithms have in-built feature selection or transformation capabilities, which may bias the results.

For each UCI dataset, we created 5 different SDTs, each one generated through random coarsening: For each value $y \in V_t \setminus \{t(x)\}$, a biased coin with success probability γ was flipped to decide whether or not it is added to the true decision $t(x)$ as an additional candidate. Obviously, the parameter γ allows for varying and controlling the degree of *ambiguity* [9,39]. We considered the following values: 0% (i.e., the case in which $d(x) = t(x)$, which allows us to compute the true reducts for the SDT as a reference comparison), 5%, 10%, and 25%.

To estimate predictive performance, we adopted a 5-fold cross-validation approach: during each iteration, 4 folds were used for training while the remaining fold was used for testing. The training folds were used for feature selection, using the proposed methods and the DELIN algorithm, and for training the PL-KNN algorithm. The test fold was then used to measure the accuracy of the trained PL-KNN models. Specifically, we measured both the average accuracy across the 5 folds and the corresponding 95% confidence intervals.

4.1. Comparison of reducts for superset decision tables

In the first experiment, each dataset was discretized in a pre-processing step, i.e., numerical attributes were replaced by categorical attributes. In particular, since *Boston*, *Abalone* and *Forest Fires* are originally regression datasets (i.e., the target attribute t is continuous), we also discretized the target attribute. The discretization was performed by applying the k-means algorithm [37] with $k = 5$ (on the values of the respective numerical attribute, i.e., running k-means on a one-dimensional dataset) and $k = 2$ (on the values of the target attribute).

We evaluated five different algorithms: the brute-force enumeration algorithm for computing MDL reducts (see Algorithm 1), the brute-force enumeration algorithms for computing H_{Bet} and OAU reducts, and the greedy algorithms to compute H_{Bet} and OAU reducts (see Algorithm 2). The algorithms were compared with respect to both their running time and their ability to recover the true reducts (that is, the reducts on the SDT with 0% ambiguity degree). A time budget of 10,000 seconds was assigned to each algorithm. The results of the experiments are reported in Tables 2–10. Based on these results, the following observations can be made:

- Computing MDL reducts, at least through the application of the brute-force algorithm (see Algorithm 1), is in general infeasible in terms of computation time. Indeed, among all 8 examined datasets, only on two 5% SDT and only on one 10% SDT, the algorithm finished the computation within the time budget. The two datasets were the smallest in terms of number of objects and attributes. This is hardly surprising, as the time complexity of Algorithm 1 is exponential in both the number of attributes and the number of objects. In the average case, we expect the algorithm to have a time complexity of $O(2^{|Att|} \cdot 2^{\epsilon|U|})$ on an ϵ% SDT. Let us also note that for all three datasets, the MDL reducts coincided with the minimal OAU reducts. This finding is interesting as, in light of Theorems 3.3 and 3.5, we know that the two definitions of reducts are equivalent only for consistent SDT, while all the considered SDTs were actually inconsistent.
- Regarding OAU reducts, it is interesting to observe that on all datasets, in the 5% and 10% SDT, the true reducts (that is, the reducts on the 0% SDT) were among the OAU reducts, and in all cases but three (Wine, Boston, and Forest Fires), the OAU reducts coincided with the true reducts. For the 25% SDT, on all datasets but three (Boston House Prices, Breast Cancer, Adult Census Income), the true reducts were among the OAU reducts, while on the three remaining datasets, the OAU reducts were subsets of the true reducts. Thus, from an empirical point of view, the notion of OAU reduct seems to be effective as a method to discover the true reducts.
- On the other hand, regarding H_{Bet} reducts, in only three 5% SDT (Forest Fires, Abalone, Iris) and in only one 10% SDT (Forest Fires), the H_{Bet} (minimal) reducts were among the true (minimal) reducts. In only one case (the 5% SDT for dataset Iris), the H_{Bet} reducts coincided with the true reducts, while in all other cases the H_{Bet} reducts were either a sub-family of the true reducts or super-reducts (indeed, in most cases the only H_{Bet} reduct was the set Att of all attributes). Thus, compared with OAU reducts, the requirement imposed by H_{Bet} entropy seems to be too conservative.

Table 2
Results for dataset Iris.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	1 reduct (2,3)				
5%	1 reduct (2,3) 60 s	1 reduct (2,3) 0.24 s	1 reduct (2,3) 0.17 s	(2,3) 0.15 s	(2,3) 0.13 s
10%	1 reduct (2,3) 5570 s	1 reduct (2,3) 0.24 s	1 reduct *A* 0.17 s	(2,3) 0.15 s	*A* 0.13 s
25%	-	2 reducts (1,2) (2,3) 0.24 s	1 reduct *A* 0.17 s	(1,2) 0.15 s	*A* 0.13 s

Table 3
Results for dataset Boston house prices.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	2 reducts, 1 minimal (0, 3, 5, 6, 7, 10, 11, 12)				
5%	-	2 reducts, 1 minimal (0, 3, 5, 6, 7, 10, 11, 12) 86 s	1 reduct *A* 70 s	(0, 1, 5, 6, 7, 8, 10, 11, 12) 1.73 s	*A* 1.26 s
10%	-	2 minimal reducts (0, 3, 5, 6, 7, 10, 11, 12) (0, 5, 6, 7, 8, 10, 11, 12) 86 s	1 reduct *A* 70 s	(0, 3, 5, 6, 7, 10, 11, 12) 1.73 s	*A* 1.26 s
25%	-	1 minimal reducts (0, 5, 6, 7, 10, 11, 12) 86 s	1 reduct *A* 70 s	(0,5,6,7,10,11,12) 1.73 s	*A* 1.26 s

Table 4
Results for dataset Breast Cancer.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	6 reducts (0, 3, 4, 5, 6, 10, 11, 12, 13, 14) (0, 3, 4, 5, 7, 10, 11, 12, 13, 14) (0, 3, 4, 5, 8, 10, 11, 12, 13, 14) (1, 3, 4, 5, 6, 10, 11, 12, 13, 14) (1, 3, 4, 5, 7, 10, 11, 12, 13, 14) (1, 3, 4, 5, 8, 10, 11, 12, 13, 14)				
5%	-	6 reducts As in the 0% SDT 3316 s	1 reduct *A* 3186 s	(0, 3, 4, 5, 6, 10, 11, 12, 13, 14) 15.3 s	*A* 13.8 s
10%	-	6 reducts As in the 0% SDT 3316 s	1 reduct *A* 3186 s	(0, 3, 4, 5, 6, 10, 11, 12, 13, 14) 15.3 s	*A* 13.8 s
25%	-	1 minimal reducts (0, 5, 6, 7, 10, 11, 12) 86 s	1 reduct *A* 70 s	(0, 5, 6, 7, 10, 11, 12) 1.73 s	*A* 1.26 s

This provides a stronger empirical counterpart of Theorems 3.2 and 3.7 and suggests that, in most practical cases, the requirements for the existence of H_{Bet} reducts are strictly stronger than those for OAU reducts.

- As for the approximate entropy (both OAU and H_{Bet}) computed according to Algorithm 2, the computed reduct was in all cases except two (the Wine dataset for H_{Bet} reducts, Boston House Prices for OAU reducts) a minimal reduct (according to the respective definition of entropy reducts). In particular, the approximate Algorithm for computing OAU reducts was able to recover one of the true minimal reducts in most datasets, at a computational cost which was, on average, at least ten times smaller. Thus, the heuristic greedy algorithm for finding OAU reducts seems to be effective in finding minimal reducts with significantly reduced time complexity.

4.2. Comparison between rough set feature selection and DELIN

Based on the results of the first experiment, we decided to use the algorithm for computing OAU reducts for the second study, since this algorithm has shown strong performance in discovering the real reducts, as discussed in Section 4.1.

Table 5
Results for dataset Diabetes.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	2 reducts (0, 1, 2, 3, 4, 6, 7, 8, 9) (0, 1, 2, 3, 5, 6, 7, 8, 9)				
5%		2 reducts	1 reduct	(0, 1, 2, 3, 4, 6, 7, 8, 9)	A
	-	As in the 0% SDT	A		
		147 s	147 s	16.2 s	16.7 s
10%		2 reducts	1 reduct	(0, 1, 2, 3, 4, 6, 7, 8, 9)	A
	-	As in the 0% SDT	A		
		147 s	147 s	16.2 s	16.7 s
25%		2 reducts	1 reduct	(0, 1, 2, 3, 4, 6, 7, 8, 9)	A
	-	As in the 0% SDT	A		
		147 s	147 s	16.2 s	16.7 s

Table 6
Results for dataset Adult Census Income.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	8 reducts, 2 minimal (1, 2, 4, 5, 7, 8, 11, 12, 13, 14) (1, 2, 4, 5, 7, 10, 11, 12, 13, 14)				
5%		8 reducts, 2 minimal	4 reducts, 2 minimal	$A \setminus \{0, 3, 9, 10\}$	$A \setminus \{0, 9, 10\}$
	-	As in the 0% SDT	(1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14) (1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14)		
		2645 s	2637 s	15 s	11 s
10%		8 reducts, 2 minimal	3 reducts, 2 minimal	$A \setminus \{0, 3, 9, 10\}$	$A \setminus \{0, 9, 10\}$
	-	As in the 0% SDT	As in the 5% SDT		
		2645 s	2637 s	15 s	11 s
25%		4 reducts, 1 minimal	2 reducts, 1 minimal	(1, 2, 4, 5, 7, 11, 12, 13, 14)	$A \setminus \{8, 9\}$
	-	(1, 2, 4, 5, 7, 11, 12, 13, 14)	$A \setminus \{8, 9\}$		
		2645 s	2637 s	15 s	11 s

Table 7
Results for dataset Abalone.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	1 reduct (2, 3, 5, 6, 7)				
5%		1 reduct	1 reduct	(2, 3, 5, 6, 7)	(2, 3, 5, 6, 7)
	-	As in the 0% SDT	As in the 0% SDT		
		475 s	421 s	16 s	16 s
10%		1 reduct	1 reduct	(2, 3, 5, 6, 7)	(0, 2, 3, 5, 6, 7)
	-	As in the 0% SDT	(0, 2, 3, 5, 6, 7)		
		475 s	421 s	16 s	16 s
25%		2 reducts	1 reduct	(2, 3, 5, 6)	A
	-	(2, 3, 5, 6) (2, 3, 5, 7)	A		
		475 s	421 s	16 s	11 s

Specifically, we evaluated the greedy algorithm for computing OAU reducts (see Algorithm 2), in order to limit the execution time, as the evaluation was implemented using 5-fold cross-validation. For comparison, as already said, we used the DELIN algorithm. For the DELIN algorithm, at each iteration of 5-fold cross-validation procedure, the number of dimensions to be selected was set equal to the size of the minimal reduct found by the greedy OAU algorithm.[1]

For each iteration of the 5-fold cross-validation procedure, the training fold was used to both compute the minimal reducts (respectively, applying dimensionality reduction using the DELIN algorithm) and the reduced training set was then used to train the PL-KNN algorithm and the performance of the two feature selection approaches was compared by assessing the accuracy of the trained models on the reduced test fold. The results were then averaged across the 5 folds. The results are reported in Table 9. In most datasets (6 out of 8), the rough set-based feature selection algorithm performed better (in terms of average predictive accuracy) than the DELIN algorithm. In order to evaluate if the reported differences are statistically significant, we performed a Wilcoxon signed rank test [60] with a confidence level of 95% ($\alpha = 0.05$). The ob-

[1] Moreover, since DELIN requires numerical features, categorical features were first one-hot encoded.

Table 8
Results for dataset Forest Fires.

	MDL reducts	OAU reducts	H_{Bet} reducts	Greedy OAU reduct	Greedy H_{Bet} reduct
0%	3 reducts, 2 minimal (0, 1, 3, 4, 5, 6, 7, 8, 10) (0, 1, 3, 5, 6, 7, 8, 9, 10)				
5%		3 reducts, 2 minimal	2 reducts, 1 minimal	(0, 1, 3, 4, 5, 6, 7, 8, 10)	(0, 1, 3, 5, 6, 7, 8, 9, 10)
	-	As in the 0% SDT	(0, 1, 3, 5, 6, 7, 8, 9, 10)		
		1258 s	1212 s	14 s	12 s
10%		4 reducts, 3 minimal	1 reduct	(0, 1, 2, 3, 5, 6, 7, 8, 10)	(0, 1, 3, 5, 6, 7, 8, 9, 10)
	-	As in the 0% SDT plus (0, 1, 2, 3, 5, 6, 7, 8, 10)	(0, 1, 3, 5, 6, 7, 8, 9, 10)		
		1258 s	1212 s	14 s	12 s
25%		4 reducts, 3 minimal	1 reduct	(0, 1, 2, 3, 5, 6, 7, 8, 10)	*A*
	-	As in the 10% SDT	*A*		
		1258 s	1207 s	14 s	9 s

Table 9
Accuracy of the PL-KNN algorithm on reduced datasets, using both the OAU algorithm and the DELIN algorithm. For each dataset and level of ambiguity, the numbers in bold denote the feature selection algorithm resulting in the best performance.

Dataset	5%		10%		25%	
	OAU	DELIN	OAU	DELIN	OAU	DELIN
Iris	**0.91 ± 0.09**	**0.91 ± 0.07**	**0.90 ± 0.10**	0.83 ± 0.13	**0.90 ± 0.10**	0.82 ± 0.15
Cancer	0.91 ± 0.03	**0.92 ± 0.03**	0.89 ± 0.05	**0.90 ± 0.03**	0.89 ± 0.05	**0.90 ± 0.05**
Wine	**0.82 ± 0.12**	0.78 ± 0.08	**0.81 ± 0.12**	0.72 ± 0.17	**0.79 ± 0.13**	0.71 ± 0.17
Boston	**0.81 ± 0.10**	0.73 ± 0.12	**0.81 ± 0.10**	0.71 ± 0.12	**0.79 ± 0.11**	0.70 ± 0.12
Diabetes	**0.72 ± 0.03**	**0.71 ± 0.03**	**0.71 ± 0.03**	**0.71 ± 0.05**	**0.70 ± 0.04**	0.69 ± 0.05
Adult	**0.73 ± 0.04**	0.72 ± 0.04	**0.73 ± 0.04**	0.72 ± 0.04	**0.73 ± 0.04**	0.72 ± 0.04
Forest Fires	**0.86 ± 0.07**	0.82 ± 0.07	**0.86 ± 0.07**	0.82 ± 0.07	**0.83 ± 0.09**	0.79 ± 0.10
Abalone	**0.76 ± 0.07**	**0.76 ± 0.07**	**0.75 ± 0.07**	**0.75 ± 0.07**	**0.75 ± 0.09**	**0.75 ± 0.09**

tained statistic was $z = -3.2797$ (p-value $= 0.001$), which means the difference between the two algorithms is statistically significant at the selected confidence level. Thus, our results provide evidence in favor of the conjecture that the features selected by the rough set-based approach are more informative than the features constructed using the DELIN algorithm.

That said, these results should of course be taken with some caution. Indeed, one may argue that a direct comparison between the two algorithms is difficult, for example because OAU requires discrete data while DELIN is working on numerical attributes. Moreover, DELIN relies on certain assumptions regarding the distribution of the data, so that its performance will depend on whether or not these assumptions are met. Rough set-based feature selection methods, on the other side, are entirely non-parametric and thus allow more flexibility in modeling the relationship between the target and the features. While this is clearly an advantage, some information might be lost through the discretization of numerical features: future work should be devoted toward generalizing the proposed approach to encompass rough set-techniques that can directly manage continuous features.

In terms of computational complexity and running time, the DELIN algorithm is vastly more efficient than the standard brute-force algorithm to compute OAU reducts. Indeed, the algorithm for finding OAU reducts is combinatorial and, in general, has exponential running time (in the number of features). Compared to this, DELIN is based on LDA and has a running time which is essentially quadratic (more precisely, $O(|U||Att|^2)$) and can easily be implemented using standard linear algebra and optimization software. We note, though, that Algorithm 2, which was the method we adopted in our comparison, has the same computational complexity as DELIN, and in our experiments was shown to still be effective at finding the true reducts. Thus, the heuristic greedy approach to finding OAU reducts could be seen as a useful trade-off on large-scale datasets.

5. Conclusion

Addressing the problem of superset learning in the context of rough set theory, as we did in this paper, appears to be interesting and mutually beneficial for both sides:

- RST provides natural tools for *data disambiguation*, which is at the core of methods for superset learning, most notably the notion of a reduct. Here, the basic idea is that the plausibility of an instantiation of the data is in direct correspondence with the (information-theoretic) complexity it implies for the dependency between input features and target (decision) variable (and a reduct in turn captures just this complexity).
- For RST itself, the setting of superset learning is a quite natural extension of the standard setting of supervised learning, and comes with a number of interesting challenges and non-trivial generalizations of existing concepts.

Table 10
Results for dataset Wine.

	MDL	OAU	H_{Bet}	Greedy OAU	Greedy H_{Bet}
0%	163 reducts, 9 minimal (0, 1, 4, 5, 8, 9), (0, 2, 3, 5, 7, 10), (0, 2, 3, 5, 10, 11) (0, 2, 3, 7, 10, 11), (0, 2, 5, 8, 9, 11), (0, 3, 4, 5, 7, 10) (0, 3, 4, 5, 8, 9), (0, 4, 5, 6, 8, 9), (4, 5, 6, 9, 10, 12)				
5%	174 reducts, 16 minimal (0, 1, 2, 5, 6, 9) (0, 1, 4, 5, 8, 9) (0, 2, 3, 5, 6, 9) (0, 2, 3, 5, 6, 10) (0, 2, 3, 5, 7, 10) (0, 2, 3, 5, 9, 10) (0, 2, 3, 5, 10, 11) (0, 2, 3, 7, 10, 11) (0, 2, 5, 6, 8, 9) (0, 2, 5, 8, 9, 11) (0, 3, 4, 5, 7, 10) (0, 3, 4, 5, 8, 9) (0, 4, 5, 6, 8, 9) (1, 2, 6, 7, 9, 12) (4, 5, 6, 9, 10, 12) (4, 5, 8, 9, 10, 12)	174 reducts, 16 minimal Same as MDL reducts	9 reducts, 6 minimal (0, 2, 3, 6, 7, 8, 9, 11) (0, 2, 3, 7, 8, 9, 10, 11) (0, 2, 4, 7, 8, 9, 10, 11) (0, 3, 4, 6, 7, 8, 9, 11) (0, 3, 5, 6, 7, 8, 9, 11) (0, 3, 5, 7, 8, 9, 10, 11)	(0, 2, 5, 6, 8, 9)	(0, 2, 3, 6, 7, 8, 9, 10, 11)
	9281 s	98 s	70 s	1.73 s	1.26 s
10%		188 reducts, 16 minimal The minimal reducts are as in the 5% SDT	9 reducts, 3 minimal (0, 1, 2, 3, 7, 8, 9, 10, 11) (0, 2, 3, 4, 7, 8, 9, 10, 11) (0, 2, 4, 5, 7, 8, 9, 10, 11)	(0, 2, 5, 6, 8, 9)	(0, 2, 3, 6, 7, 8, 9, 10, 11)
		98 s	70 s	1.73 s	1.26 s
25%	-	191 reducts, 34 minimal The minimal reducts are as in the 5% SDT plus (0, 1, 2, 7, 9, 10) (0, 1, 2, 7, 10, 12) (0, 1, 4, 5, 9, 10) (0, 2, 3, 4, 5, 10) (0, 2, 3, 4, 6, 7) (0, 2, 3, 4, 7, 10) (0, 2, 3, 6, 7, 10) (0, 2, 3, 7, 8, 10) (0, 2, 3, 7, 9, 10) (0, 2, 3, 7, 10, 12) (0, 2, 4, 7, 9, 10) (0, 2, 5, 7, 8, 9) (0, 2, 7, 8, 9, 10) (0, 3, 4, 5, 6, 10) (0, 3, 4, 5, 9, 10) (0, 3, 4, 5, 10, 11) (0, 4, 5, 7, 9, 10) (0, 4, 5, 8, 9, 11)	1 reduct A	(0, 1, 2, 7, 9, 10)	A
		98 s	70 s	1.73 s	1.26 s

One such challenge has been tackled in this paper, namely the question how to generalize the notion of a reduct as well as devising algorithms for feature selection on the basis of this notion.

To this end, we first proposed a generalization of decision tables and then studied a purely combinatorial definition of reducts inspired by the Minimum Description Length principle, which we called MDL reducts. Since, as we showed, the computational complexity of finding this type of reducts is NP-hard, we proposed two alternative definitions based on the notion of entropy and harnessing the natural relationship between superset learning and belief function theory. We then provided a characterization for both these notions in terms of their relationship with MDL reducts, their existence conditions and their disambiguation power. Moreover, we developed simple heuristic algorithms for computing approximate entropy reducts.

Finally, we conducted experiments on real datasets in order to empirically compare the different definitions of reducts for superset learning and the algorithms for computing them. As a result of these experiments, we conclude that the definition based on OAU entropy seems to be more effective in terms of its ability to recover the true but unknown reducts, compared with the definition based on H_{Bet} entropy. We have also shown that our heuristic algorithm for computing approximate entropy provides an effective approach to finding minimal reducts with limited computational resources. Finally, we compared the proposed feature selection methods with a state-of-the-art dimensionality reduction algorithm for superset learning and showed that the proposed method leads to a significantly higher classification accuracy on a collection of benchmark datasets, thus highlighting its usefulness in applications.

While this paper provides a promising direction for the application of RST-based feature reduction in superset learning, it naturally leaves many questions open. Specifically, we plan to address the following problems in future works:

- In Theorem 3.5, we proved that, in general, OAU reducts are a sub-family of the superset super-reducts. However, our experiments also showed that in most cases (in which the MDL reducts were actually computed within the assigned time budget) the MDL reducts were exactly equivalent to the OAU reducts. Thus, the conditions for such an equivalence between the two definitions should be investigated in more depth;
- In Theorems 3.6 and 3.7, we described two characterizations of H_{Bet} reducts: first, showing sufficient and necessary conditions for their existence on *binary* decision tables; second, showing that, in general, H_{Bet} reducts are OAU super-reducts. Therefore, the generalization of Theorem 3.6 to the multi-class case, together with a characterization of the conditions for the equivalence between H_{Bet} reducts and OAU reducts, should be investigated;

- The proposed RST feature reduction methods require the available data to be discrete: otherwise, data discretization techniques need to be applied which, in turn, could have an impact on the results and performance of the feature selection procedure. While, at least in principle, scaling techniques [21] (such as those applied in Formal Concept Analysis) could be applied to manage continuous features, these would likely have a huge impact on the computational complexity of the proposed methods. Thus, the generalization of the proposed approach to also encompass RST techniques that can directly manage continuous features, such as neighborhood- [43] or fuzzy-rough [32] based approaches, should be investigated.
- We studied the application of RST feature reduction to the superset learning task, however, it would also be interesting to study an extension of the proposed framework to other, even more general settings, such as learning from fuzzy [14,28] or evidential [8,11,13,48,41] data.
- In this paper, the superset assumption was motivated by the problem of imprecise labeling. As explained in Section 2.3, this "don't know" interpretation can be distinguished from a "don't care" interpretation. Proceeding from the latter, a reduct can be considered as a maximally simple (least cognitively demanding) yet satisfying decision rule. Interestingly, in spite of very different interpretations, the theoretical problems that arise are essentially the same as those studied in this paper. Nevertheless, elaborating on the idea of reduction as a means for finding satisfying decision rules from a more practical point of view is another interesting direction for future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Joaquin Abellan, Combining nonspecificity measures in Dempster–Shafer theory of evidence, Int. J. Gen. Syst. 40 (6) (2011) 611–622.
[2] Joaquin Abellan, Serafin Moral, Completing a total uncertainty measure in the Dempster-Shafer theory, Int. J. Gen. Syst. 28 (4–5) (1999) 299–314.
[3] Pierre C. Bellec, Arnak S. Dalalyan, Edwin Grappin, Quentin Paris, et al., On the prediction loss of the lasso in the partially labeled setting, Electron. J. Stat. 12 (2) (2018) 3443–3472.
[4] Rafael Bello, Rafael Falcon, Rough sets in machine learning: a review, in: Thriving Rough Sets, Springer, 2017, pp. 87–118.
[5] Andrea Campagner, Davide Ciucci, Orthopartitions and soft clustering: soft mutual information measures for clustering validation, Knowl.-Based Syst. 180 (2019) 51–61.
[6] Andrea Campagner, Davide Ciucci, Eyke Hüllermeier, Feature reduction in superset learning using rough sets and evidence theory, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2020, pp. 471–484.
[7] Leilei Chang, Chao Fu, Wei Zhu, Weiyong Liu, Belief rule mining using the evidential reasoning rule for medical diagnosis, Int. J. Approx. Reason. 130 (2021) 273–291.
[8] Etienne Côme, Latifa Oukhellou, Thierry Denoeux, Patrice Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognit. 42 (3) (2009) 334–348.
[9] Timothee Cour, Ben Sapp, Ben Taskar, Learning from partial labels, J. Mach. Learn. Res. 12 (2011) 1501–1536.
[10] Arthur P. Dempster, Upper and lower probabilities induced by a multivalued mapping, in: Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, 2008, pp. 57–72.
[11] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Trans. Syst. Man Cybern. 25 (5) (1995) 804–813.
[12] Thierry Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, in: Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, 2008, pp. 737–760.
[13] Thierry Denoeux, Maximum likelihood estimation from uncertain data in the belief function framework, IEEE Trans. Knowl. Data Eng. 25 (1) (2011) 119–130.
[14] Thierry Denœux, Lalla Meriem Zouhal, Handling possibilistic labels in pattern classification using evidential reasoning, Fuzzy Sets Syst. 122 (3) (2001) 409–424.
[15] Adrian Dobra, Stephen E. Fienberg, Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, Proc. Natl. Acad. Sci. USA 97 (22) (2000) 11885–11892.
[16] Didier Dubois, Henri Prade, Properties of measures of information in evidence and possibility theories, Fuzzy Sets Syst. 24 (2) (1987) 161–182.
[17] Bradley Efron, Censored data and the bootstrap, J. Am. Stat. Assoc. 76 (374) (1981) 312–319.
[18] Lei Feng, Bo An, Leveraging latent label distributions for partial label learning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 2107–2113.
[19] Lei Feng, Bo An, Partial label learning with self-guided retraining, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3542–3549.
[20] A. Frank, A. Asuncion, UCI machine learning repository, 2010.
[21] Bernhard Ganter, Rudolf Wille, Conceptual scaling, in: Applications of Combinatorics and Graph Theory to the Biological and Social Sciences, Springer, 1989, pp. 139–167.
[22] Romain Guillaume, Didier Dubois, Robust parameter estimation of density functions under fuzzy interval observations, in: 9th International Symposium on Imprecise Probability: Theories and Applications, ISIPTA'15, 2015, pp. 147–156.
[23] Romain Guillaume, Didier Dubois, A maximum likelihood approach to inference under coarse data based on minimax regret, in: International Conference Series on Soft Methods in Probability and Statistics, Springer, 2018, pp. 99–106.
[24] Isabelle Guyon, André Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar 2003) 1157–1182.
[25] David Harmanec, George J. Klir, Measuring total uncertainty in Dempster-Shafer theory: a novel approach, Int. J. Gen. Syst. 22 (4) (1994) 405–419.
[26] Ulrich Hohle, Entropy with respect to plausibility measures, in: Proceedings of 12th IEEE International Symposium on Multiple Valued Logic, Paris, 1982, 1982.
[27] E. Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (7) (2014) 1519–1534.
[28] Eyke Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (7) (2014) 1519–1534.

[29] Eyke Hüllermeier, Jürgen Beringer, Learning from ambiguously labeled examples, Intell. Data Anal. 10 (5) (2006) 419–439.
[30] Eyke Hüllermeier, Weiwei Cheng, Superset learning based on generalized loss minimization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 260–275.
[31] Eyke Hüllermeier, Sébastien Destercke, Ines Couso, Learning from imprecise data: adjustments of optimistic and pessimistic variants, in: Nahla Ben Amor, Benjamin Quost, Martin Theobald (Eds.), Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compiègne, France, December 16-18, 2019, Proceedings, in: Lecture Notes in Computer Science, vol. 11940, Springer, 2019, pp. 266–279.
[32] Richard Jensen, Qiang Shen, Fuzzy-rough sets assisted attribute selection, IEEE Trans. Fuzzy Syst. 15 (1) (2007) 73–89.
[33] Rong Jin, Zoubin Ghahramani, Learning with multiple labels, in: Advances in Neural Information Processing Systems, 2003, pp. 921–928.
[34] Radim Jiroušek, Prakash P. Shenoy, A new definition of entropy of belief functions in the Dempster–Shafer theory, Int. J. Approx. Reason. 92 (2018) 49–65.
[35] Radim Jiroušek, Prakash P. Shenoy, On properties of a new decomposable entropy of Dempster-Shafer belief functions, Int. J. Approx. Reason. 119 (2020) 260–279.
[36] A-L. Jousselme, Chunsheng Liu, Dominic Grenier, Éloi Bossé, Measuring ambiguity in the evidence theory, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 36 (5) (2006) 890–903.
[37] Sotiris Kotsiantis, Dimitris Kanellopoulos, Discretization techniques: a recent survey, GESTS Int. Trans. Comput. Sci. Eng. 32 (1) (2006) 47–58.
[38] Ming Li, Paul Vitányi, et al., An Introduction to Kolmogorov Complexity and its Applications, 3rd edition, Springer, 2008.
[39] Liping Liu, Thomas Dietterich, Learnability of the superset label learning problem, in: International Conference on Machine Learning, 2014, pp. 1629–1637.
[40] Liping Liu, Thomas G. Dietterich, A conditional multinomial mixture model for superset label learning, in: Advances in Neural Information Processing Systems, 2012, pp. 548–556.
[41] Liyao Ma, Sébastien Destercke, Yong Wang, Online active learning of decision trees with evidential data, Pattern Recognition 52 (2016) 33–45.
[42] J.G. March, H.A. Simon, Organizations, Wiley, New York, 1958.
[43] Michinori Nakata, Hiroshi Sakai, Keitarou Hara, Rule induction based on rough sets from information tables having continuous domains, CAAI Trans. Intell. Technol. 4 (4) (2019) 237–244.
[44] Nam Nguyen, Rich Caruana, Classification with partial labels, in: Proceedings of the 14th ACM SIGKDD, 2008, pp. 551–559.
[45] Qiang Ning, Hangfeng He, Chuchu Fan, Dan Roth, Partial or complete, that's the question, preprint, arXiv:1906.04937, 2019.
[46] Zdzisław Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[47] Judea Pearl, Reasoning with belief functions: an analysis of compatibility, Int. J. Approx. Reason. 4 (5–6) (1990) 363–389.
[48] Benjamin Quost, Thierry Denoeux, Shoumei Li, Parametric classification with soft labels using the evidential em algorithm: linear discriminant analysis versus logistic regression, Adv. Data Anal. Classif. 11 (4) (2017) 659–690.
[49] Hiroshi Sakai, Chenxi Liu, Michinori Nakata, Shusaku Tsumoto, A proposal of a privacy-preserving questionnaire by non-deterministic information and its analysis, in: 2016 IEEE International Conference on Big Data, Big Data, IEEE, 2016, pp. 1956–1965.
[50] Glenn Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[51] Claude E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[52] Razieh Sheikhpour, Mehdi Agha Sarram, Sajjad Gharaghani, Mohammad Ali Zare Chahooki, A survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.
[53] Andrzej Skowron, Cecylia Rauszer, The discernibility matrices and functions in information systems, in: Intelligent Decision Support, Springer, 1992, pp. 331–362.
[54] Dominik Slezak, Approximate entropy reducts, Fundam. Inform. 53 (3–4) (2002) 365–390.
[55] Dominik Slezak, Soma Dutta, Dynamic and discernibility characteristics of different attribute reduction criteria, Lecture Notes in Computer Science 11103 (2018) 628–643.
[56] Philippe Smets, Information content of an evidence, Int. J. Man-Mach. Stud. 19 (1) (1983) 33–43.
[57] Philippe Smets, Robert Kennes, The transferable belief model, Artif. Intell. 66 (2) (1994) 191–234.
[58] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, Appl. Soft Comput. 9 (1) (2009) 1–12.
[59] Christopher Umans, On the complexity and inapproximability of shortest implicant problems, in: Automata, Languages and Programming, Springer, Berlin, Heidelberg, 1999, pp. 687–696.
[60] Frank Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.
[61] Jing-Han Wu, Min-Ling Zhang, Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 416–424.
[62] Ronald R. Yager, Entropy and specificity in a mathematical theory of evidence, in: Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, 2008, pp. 291–310.
[63] Y.Y. Yao, P.J. Lingras, Interpretations of belief functions in the theory of rough sets, Inf. Sci. 104 (1–2) (1998) 81–106.
[64] Fei Yu, Min-Ling Zhang, Maximum margin partial label learning, in: Asian Conference on Machine Learning, 2016, pp. 96–111.
[65] Chunying Zhang, Xiaoze Feng, Ruiyan Gao, Three-way decision models and its optimization based on Dempster–Shafer evidence theory and rough sets, Granul. Comput. 6 (2021) 411–420.
[66] Min-Ling Zhang, Fei Yu, Solving the partial label learning problem: an instance-based approach, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.
[67] Shao-Pu Zhang, Pin Sun, Ju-Sheng Mi, Tao Feng, Belief function of Pythagorean fuzzy rough approximation space and its applications, Int. J. Approx. Reason. 119 (2020) 58–80.
[68] Yan-Lan Zhang, Chang-Qing Li, Relationships between relation-based rough sets and belief structures, Int. J. Approx. Reason. 127 (2020) 83–98.
[69] Zhi-Hua Zhou, A brief introduction to weakly supervised learning, Nat. Sci. Rev. 5 (1) (2018) 44–53.

Feature Selection and Disambiguation in Learning from Fuzzy Labels Using Rough Sets

Andrea Campagner( ) and Davide Ciucci

Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, viale Sarca 336, 20126 Milano, Italy
a.campagner@campus.unimib.it

Abstract. In this article, we study the setting of *learning from fuzzy labels*, a generalization of supervised learning in which instances are assumed to be labeled with a fuzzy set, interpreted as an epistemic possibility distribution. We tackle the problem of feature selection in such task, in the context of *rough set theory* (RST). More specifically, we consider the problem of RST-based feature selection as a means for *data disambiguation*: that is, retrieving the most plausible precise instantiation of the imprecise training data. We define generalizations of decision tables and reducts, using tools from generalized information theory and belief function theory. We study the computational complexity and theoretical properties of the associated computational problems.

Keywords: Fuzzy labels · Rough sets · Feature selection · Belief functions · Entropy

1 Introduction

Weakly supervised learning [34] refers to Machine Learning tasks in which training instances are not required to be associated with a precise target label: the annotations can be either imprecise or partial. Such tasks could be a consequence of certain data pre-processing operations such as anonymization [24]; could be due to imprecise measurements or expert opinions; or to limit data annotation costs [20]. Some particularly relevant instances of weakly supervised learning are *superset learning* [15] (i.e. instances are associated with sets of candidate labels), *learning from evidential labels* [6,9] (i.e., instances are associated with belief functions over the labels) and *learning from fuzzy labels* [10,13]. In this latter setting, which is the focus of this article, each instance x is annotated with a fuzzy set μ of candidate labels. These fuzzy sets have an epistemic semantics and represent possibility distributions π_μ: only one of the labels is the correct one and the fuzzy membership degrees, then, describe the possibility degree of the labels. For example, an image could be tagged with $\{\text{horse} : 1, \text{pony} : 0.8, \text{zebra} : 0.5, \text{dog} : 0.0\}$, suggesting that the animal shown

© Springer Nature Switzerland AG 2021
S. Ramanna et al. (Eds.): IJCRS 2021, LNAI 12872, pp. 164–179, 2021.
https://doi.org/10.1007/978-3-030-87334-9_14

on the picture is one among {horse, pony, zebra}: though it is not exactly known which of them, it is known that *horse* is deemed more plausible than *pony*, which in turn is deemed more plausible than *zebra*.[1]

While in recent years the superset learning task has been widely investigated [4,16,17], also using approaches based on Rough Set theory [4,25], the *learning from fuzzy labels* task has, comparatively, been given less attention mainly due to the high computational complexity of the problem [13] and to the difficulty of acquiring such data [14]: most works have focused on the problem of classification [6,9,10,13,23] (in particular in those tasks where the acquisition of such fuzzy labels is easier, e.g. multi-rater learning and self-regularized learning [11]), while other tasks such as *feature selection*, despite their importance, have mostly been ignored.

In this article, drawing from our previous work on superset feature selection [4], we attempt to close this gap by proposing methods, based on Rough Set Theory (RST), Belief Function Theory (BFT) and possibility theory, to address the problem of feature selection. Remarkably, in line with the generalized risk minimization paradigm [13], we consider this task as a means for *data disambiguation*, i.e., for the purpose of figuring out the most plausible precise instantiation of the imprecise training data. For this purpose we propose a generalization of standard Decision Tables and we describe different definitions of reducts. In particular, in Sect. 2 we provide the necessary background knowledge on possibility theory, Rough Set theory and Belief Function theory; in Sect. 3.1 we define a generalization of decision tables to the learning from fuzzy label settings; in Sect. 3.2 we introduce several notions of reducts and study their relationships and computational complexity properties; in Sect. 3.3 we propose a generalization of entropy reducts, in order to provide an approach for performing feature selection which is more apt at the design of heuristics or approximation algorithms; finally, in Sect. 4, we summarize our results and describe some open problems.

2 Background

In this section, we recall basic notions of rough set theory (RST) and evidence theory, which will be used in the main part of the article.

2.1 Possibility Theory

Possibility theory is a theory of uncertainty, alternative to probability theory, which allows for the quantification of degrees of possibility on the basis of a fuzzy set [33]. We recall that a fuzzy set (equivalently, a possibility distribution)

[1] We note that in the *learning from fuzzy labels* setting, the set of candidate labels (that is, the labels with a membership degree greater than 0) is given a disjunctive interpretation: only one of those labels is correct, but we don't precisely know which one, and the membership degrees represent *degrees of belief*. Thus, in this article, we do not consider the conjunctive interpretation, in which the membership degrees are *degrees of truth* (and, thus, could be seen as a generalization of multi-label learning).

F can be seen as a function $F : X \mapsto [0,1]$, that is, a generalization of the characteristic function representation of classical sets. A possibility measure is a function $pos_F : 2^X \mapsto [0,1]$ such that

1. $pos_F(\emptyset) = 0$ and $pos_F(X) = 1$,;
2. if $A \cap B = \emptyset$ then $pos_F(A \cup B) = max(pos_F(A), pos_F(B))$.

It can be easily seen that every possibility measure is induced by a fuzzy set F as $pos_F(A) = max_{x \in A} F(x)$: in this case we say that F is the possibility distribution corresponding to the possibility measure pos_F.

A possibility distribution F is *normal* if $\exists x \in X. F(x) = 1$: in this article we will focus on normal possibility distributions. Given $\alpha \in [0,1]$, the *alpha*-cut of F is defined as $F^{\alpha} = \{x \in X : F(x) \geq \alpha\}$, while the *strong* α-cut is defined as $F^{\alpha+} = \{x \in X : F(x) > \alpha\}$: we recall that the collection of α-cuts of F is sufficient to determine F [19].

The epistemic view [8] of possibility distributions refers to the common interpretation under which a possibility distribution represents the degrees of belief (of an agent) towards a set of possible alternatives. We refer the reader to [7,13] for a discussion of epistemic possibility distributions in Machine Learning.

2.2 Rough Set Theory

Rough set theory has been proposed by Pawlak [22] as a framework for representing and managing uncertain data, and has since been widely applied for various problems in the ML domain (see [2] for a recent overview and survey). We briefly recall the main notions of RST, especially regarding its applications to feature selection.

A decision table (DT) is a triple $DT = \langle U, Att, t \rangle$ such that U is a universe of objects and Att is a set of *attributes* employed to represent objects in U. Formally, each attribute $a \in Att$ is a function $a : U \to V_a$, where V_a is the domain of values of a. Moreover, $t \notin Att$ is a distinguished *decision* attribute, which represents the target decision (also labeling or annotation) associated with each object in the universe. We say that DT is *inconsistent* if the following holds: $\exists x_1, x_2 \in U, \forall a \in Att, a(x_1) = a(x_2)$ and $t(x_1) \neq t(x_2)$.

Given $B \subseteq Att$, we can define the *indiscernibility relation* with respect to B as xI_Bx' iff $\forall a \in B$, $a(x') = a(x)$. Clearly, it is an equivalence relation that partitions the universe U in equivalence classes, also called *granules of information*, $[x]_B$. Then, the *indiscernibility partition* is denoted as $\pi_B = \{[x]_B \,|\, x \in U\}$.

We say that $B \subseteq Att$ is a *decision reduct* for DT if $\pi_B \leq \pi_t$ (where the order $\leq$ is the refinement order for partitions, that is, π_t is a coarsening of π_B) and there is no $C \subsetneq B$ such that $\pi_C \leq \pi_t$. Then, evidently, a reduct of a decision table DT represents a set of non-redundant and necessary features to represent the information in DT. We say that a reduct R is *minimal* if it is among the smallest (with respect to cardinality) reducts.

Given $B \subseteq Att$ and a set $S \subseteq U$, a *rough approximation* of S (with respect to B) is defined as the pair $B(S) = \langle l_B(S), u_B(S) \rangle$, where $l_B(S) = \bigcup \{[x]_B \,|\, [x]_B \subseteq$

$S\}$ is the *lower approximation* of S, and $u_B(s) = \bigcup\{[x]_B \mid [x]_B \cap S \neq \emptyset\}$ is the corresponding *upper approximation*.

Finally, given $B \subseteq Att$, the *generalized decision* with respect to B for an object $x \in U$ is defined as $\delta_B(x) = \{t(x') \mid x' \in [x]_B\}$. Notably, if DT is not inconsistent and B is a reduct, then $\delta_B(x) = \{t(x)\}$ for all $x \in U$.

We note that in the RST literature, there exist several definitions of reduct that, while equivalent on consistent Decision Tables, are generally non-equivalent for inconsistent ones. We refer the reader to [28] for an overview of such a list and a study of their dependencies. Moreover, we remark that, given a decision table, the problem of finding the minimal reduct is in general Σ_2^P-complete (by reduction to the *Shortest Implicant* problem [31]). We recall that Σ_2^P is the complexity class defined by problems that can be verified in polynomial time given access to an oracle for an NP-complete problem [1].

Finally, we recall that some previous works have investigated the generalization of Rough Set Theory to the case of imprecise data, both in the case of set-valued data [4,21,25] and in the case of possibility distributions [5], or more general uncertainty representations [30]. Nakata et al. [18] discuss a generalization of Rough Set Theory to the case where every attribute value is expressed as a possibility distribution and study generalized notions of rough approximations: though this approach uses a cut-based approach similar to the one we adopt in this paper, the authors do not study generalizations of reducts to this setting. Ciucci et al. [5] focus on a specific type of possibility distribution (certainty distributions) and study different notions for both rough approximations and reducts: in our work we consider the case of general possibility distributions, but only for the decision attribute. Also, we note that both articles [5,18] do not consider applications to the learning from fuzzy labels setting. Finally, Trabelsi et al. [30] considered the generalization of RST to account for evidential data in the decision attribute and proposed a definition of reducts in that setting: while the approach adopted by the authors shares some similarities with the approach we propose, the former does not agree with the generalized risk minimization principle [13] and hence cannot be applied to the task of data disambiguation.

2.3 Belief Function Theory

Belief Function theory (BFT), also known as Dempster-Shafer theory or evidence theory, has been introduced by Dempster and subsequently formalized by Shafer in [26]. Given a *frame of discernment* X, which represents all possible states of a system under study, a *basic belief assignment* (bba) is a function $m : 2^X \to [0, 1]$, such that $m(\emptyset) = 0$ and $\sum_{A \in 2^X} m(A) = 1$. The *support* of m is defined as $supp(m) = \{A \subseteq X : m(A) > 0\}$.

From a bba, a pair of functions, called respectively *belief* and *plausibility*, can be defined as follows:

$$Bel_m(A) = \sum_{B:B \subseteq A} m(B) \qquad Pl_m(A) = \sum_{B:B \cap A \neq \emptyset} m(B) \tag{1}$$

As can be seen from these definitions, there is a clear correspondence between BFT and, respectively, RST and possibility theory. In the first case, it is easy to note that belief functions (resp., plausibility functions) correspond to lower approximations (resp., upper approximations) in RST whenever the support m is a partition of X; we refer the reader to [32] for further connections between the two theories. In the case of possibility theory, any possibility measure (resp. necessity measure) is a plausibility (resp. belief) function: indeed, it can be shown that possibility theory can be seen as a special case of BFT where we require that m is *consonant* [26], that is $\forall A_1, A_2 \in supp(m) \,.\, A_1 \subseteq A_2 \vee A_2 \subseteq A_1$ (i.e., $supp(m)$ with the order given by $\subseteq$ is a linear order).

Finally, we recall that several generalizations of information-theoretic concepts, specifically the concept of *entropy* (which was also proposed to generalize the definition of reducts in RST [27]), have been defined for BFT. Most relevantly, we recall the definition of *optimistic aggregate uncertainty* [3,4]:

$$OAU(m) = \min_{p \in \mathcal{P}(m)} H(p), \tag{2}$$

where $\mathcal{P}(m)$ is the set of probability distributions p such that $Bel_m \leq p \leq Pl_m$ and $H(p) = -\sum_{x \in X} p(x) log_2 p(x)$ is the Shannon entropy of p.

3 Possibilistic Decision Tables and Reducts

In this section, we extend some key concepts of rough set theory to the setting of learning from fuzzy labels.

3.1 Possibilistic Decision Tables

In the *learning from fuzzy labels* setting, an object $x \in U$ is not necessarily assigned a single annotation $t(x) \in V_t$, but may instead be associated with an epistemic statement (elicited by an agent, human or computational) encoding the relative plausibility of a set S of candidate annotations, one of which is assumed to be the true annotation associated with x. The relative plausibility of the candidate annotations is expressed as a possibility distribution (or, equivalently, as a fuzzy set) over the label set. To model this idea in terms of RST, we generalize the definition of a decision table as follows.

Definition 1. *A* possibilistic decision table *(PDT) is a tuple* $P = \langle U, Att, t, d \rangle$, *where* $\langle U, Att, t \rangle$ *is a decision table, i.e.:*

- U *is a universe of objects of interest;*
- Att *is a set of attributes (or features);*
- t *is the (real) decision attribute (whose value, in general, is not known);*
- $d \notin Att$ *is a map from objects to possibility distributions over* V_t, $d : U \to \mathcal{F}(V_t)$ *such that the* weak superset property *holds:* $d(x)_{t(x)} > 0$ *for all* $x \in U$.

Remark 1. By $d(x)_y$ we denote the possibility degree assigned to class label y for object x. We adopt this convention (over the alternative $d(x)(y)$) in order to simplify the notation.

The intuitive meaning of the possibility distribution d is that, if $|d(x)^{0+}| > 1$ for some $x \in U$, then the real decision associated with x (i.e. $t(x)$) is not known precisely, but is known to be in $d(x)^{0+}$. Furthermore, if $d(x)_a > d(x)_b$ then the decision a is considered more plausible than decision b for object x. Nonetheless, an alternative *preferential* interpretation can also be considered (similarly to the superset learning setting [4,16]): in this context, the inequality $d(x)_y \leq d(x)_{y'}$ would mean that, for object x, the label y' is preferred to y. Interestingly, while in the superset learning setting the two interpretations coincide (in the sense that they define the same notion of reducts), this is not the case in the learning from fuzzy labels setting. In the following, we will mainly focus on the epistemic interpretation, though we will occasionally make reference also to the preferential one when the two differ. First, we note that Definition 1 is a proper generalization of both standard and superset decision tables (SDT) [4]: indeed, if $d(x)^{0+} = d(x)^1$ for all $x \in U$, then we have a superset decision table; and, in the particular case where it also holds that $|d(x)^{0+}| = 1$ for all $x \in U$, then we have a standard decision table. We remark that the *weak superset property* forbids the real decision $t(x)$, for any object x, to be considered impossible (that is, we assume that there are no labeling errors) but nothing more is assumed: in particular, the stronger requirement that $d(x)_{t(x)} = 1$ (which means that $t(x)$ is considered fully plausible) is not guaranteed to hold. We call this latter requirement the *strong superset property*.

While both conditions can be seen as proper generalizations of the *superset property* in superset learning [16,17], we argue that under the epistemic interpretation of a PDT, the strong superset property is, in a specific sense, trivial: indeed, were this property be satisfied, then the PDT P would be equivalent to a SDT (specifically, the SDT $S = \langle U, Att, t, d_S \rangle$ s.t. $\forall x \in U. d_S(x) = d(x)^1$) as under the strong superset condition (i.e. $d(x)_{t(x)} = 1$) the real annotation $t(x)$ is guaranteed to lie among those with an associated possibility degree equal to 1.

By contrast, in the preferential interpretation, the strong superset property only implies that $t(x)$ is the most preferred annotation for x: this, in general, does not imply that other possible annotations should not be considered.

A PDT can be associated with a collection of compatible (standard) decision tables, which we call instantiations of the PDT:

Definition 2. *An* instantiation *of a PDT $P = \langle U, Att, t, d \rangle$ is a standard decision table $T = \langle U, Att, t' \rangle$ such that $d(x)_{t'(x)} > 0$ for all $x \in U$. The collection of instantiations of P is denoted $\mathcal{I}(P)$.*

We note that the collection $\mathcal{I}(P)$ inherits a ranking of the instantiations from the definition of the possibilistic decision attribute d:

Definition 3. *Let $I_1, I_2 \in \mathcal{I}(P)$ be two instantiations of a PDT P. Then we say that I_1 is* (conservatively) less possible *than I_2, denoted $I_1 \leq_C I_2$, if:*

$$min_{x \in U} d(x)_{t'}^{I_1} \leq min_{x \in U} d(x)_{t'}^{I_2} \tag{3}$$

We say that I_1 *is* dominated in possibility *by* I_2*, denoted* $I_1 \leq_D I_2$*, if:*

$$\forall x \in U.\ d(x)^{I_1}_{t'} \leq d(x)^{I_2}_{t'} \tag{4}$$

where, in both definitions $d(x)^{I_i}_{t'}$ *refers to the value of the decision attribute* d *(in* P*) on the label* $t'(x)$ *in the instantiation* I_i*.*

It is easy to observe that the following result holds:

Proposition 1. *The order* $\leq_C$ *determines a possibility distribution (equivalently, a fuzzy set)* $\mu_{\mathcal{I}(P)}$ *on the collection* $\mathcal{I}(P)$ *where, for each* $I \in \mathcal{I}(P)$*:*

$$\mu_{\mathcal{I}(P)}(I) = min_{x \in U} d(x)^{I}_{t'} \tag{5}$$

Proof. The result easily follows from the observation that $\leq_C$ is a weak ordering on $\mathcal{I}(P)$. Using the product fuzzy set construction [19], it is then easy to see that we can associate with $\leq_C$ a possibility distribution which is equivalent to $\mu_{\mathcal{I}(P)}$. □

The order $\leq_D$, on the other hand, cannot be directly associated with a (standard) possibility distribution on $\mathcal{I}(P)$, as it only defines a partial order: thus, it defines an L-fuzzy set over the set of instantiations where, in general, $L \neq ([0,1], min, max)$. Interestingly, the $\leq_D$ order is equivalent to the notion of dominance [12] in multi-criteria decision making: this could suggest that this ordering over instantiations (and the corresponding definitions of reducts) could be of particular interest in the *preferential* interpretation of the *learning from fuzzy-label* setting.

The following definition generalizes the notion of inconsistency for a PDT:

Definition 4. *For* $B \subset Att$ *and* $\alpha \in [0,1)$ *the PDT* P *is* (α, B)-inconsistent *if*

$$\exists x_1, x_2 \in U, \forall a \in B, a(x_1) = a(x_2) \text{ and } d(x_1)^{\alpha+} \cap d(x_2)^{\alpha+} = \emptyset\,. \tag{6}$$

We call such a pair x_1, x_2 *(*α*,* B*)-inconsistent. If condition (6) is not satisfied, then* P *is* (α, B)-consistent*. In particular, we say that* P *is* weakly B-consistent *if it is* (0, B)-consistent*; while we say that* P *is* B-consistent *when it is* (α, B)-consistent *for every* α*.*

From the definition, we see that the notion of consistency (dually, inconsistency) for a PDT is richer than its classical counterpart and, in general, implies the non-existence of indiscernible objects with non-overlapping decisions, at any given α-cut of the possibility distribution defined by d. We say that an instantiation I is *α-consistent with a PDT* P if the following holds for all x_1, x_2: if x_1, x_2 are (α, Att)-consistent in S, then they are consistent in I.

3.2 Possibilistic Reducts

Learning from fuzzy labels, as a proper generalization of superset learning, encompasses the idea of *data disambiguation*: the goal of such a task is to jointly

learn a function, mapping novel objects to the corresponding correct decision, and figuring out the most plausible instantiation of the available data.

In the case of superset learning the notion of *plausibility* of an instantation can be entirely captured through the principle of simplicity [13] as any two instantiations are, a priori, equally plausible as they are both associated a possibility degree equal to 1: Thus, an instantiation that can be explained by a simple model is more plausible than an instantiation that requires a more complex one (this approach is, in turn, inspired by the *Occam's razor* principle).

In Rough Set Theory, the most natural measure of model complexity is the size of a reduct: indeed, this approach has been applied, in superset learning, to define so-called Minimum Description Length (MDL) reducts [4] which refer to the minimal reducts among all reducts of all possible instantiations. The most natural generalization of this notion to the setting of learning from fuzzy labels leads to the following definition:

Definition 5. *A set of attributes $R \subseteq Att$ is a* possibilistic reduct *of a PDT P if there exists an instantiation $I \in \mathcal{I}(P)$ s.t. R is a reduct for I. A minimum description length (MDL) instantiation is one of the instantiations of P admitting a reduct of minimum size (compared to all the reducts of all possible instantiations). We call the corresponding reducts* possibilistic MDL reducts.

While meaningful from a conceptual perspective, it is easy to observe that this definition of reducts disregards the most important difference between the superset learning and learning from fuzzy label settings: that is, the instantiations can be associated with an inherent measure of plausibility, given by the orders $\leq_C, \leq_D$. Indeed, the following result trivially holds:

Proposition 2. *Let P be a PDT, and let $\mathcal{S}(P) = \langle U, Att, t, d_S \rangle$ be the SDT defined from P s.t. $\forall x.\ d_S(x) = d(x)^{0+}$. Then, R is a possibilistic reduct (resp. possibilistic MDL reduct) of P iff it is a superset reduct (resp. MDL reduct) of $\mathcal{S}(P)$.*

Proposition 2 shows that the notion of a possibilistic reduct discards the epistemic information expressed by the decision attribute, and is thus equivalent to the notion of a superset reduct. In order to capture the richer semantics of PDTs, we argue that any proper definition of reduct should take into account not only the simplicity of the induced model (that is, the size of the reducts) but also the epistemic information encoded by the (possibilistic) decision attribute d. For this reason, we consider the following definitions of reducts:

Definition 6. *For each $\alpha \in (0, 1]$, let $\mathcal{S}(P)^\alpha$ be the SDT defined from P s.t. $\forall x.\ d_S^\alpha(x) = d(x)^\alpha$. For each possibilistic reduct R, denote by $\mathcal{I}(R) \subseteq \mathcal{I}(P)$ the collection of instantiations of P for which R is a reduct. Then, R:*

- *Is an α-possibilistic reduct if it is a superset reduct of $\mathcal{S}(P)^\alpha$, and an α-MDL reduct if it is also a MDL reduct of $\mathcal{S}(P)^\alpha$;*
- *Is a C-reduct if it is a possibilistic reduct and $\nexists R' \subseteq Att$ s.t. both $|R'| \leq |R|$ and $\exists I_1 \in sup_{\leq_C} \mathcal{I}(R), I_2 \in sup_{\leq_C} \mathcal{I}(R').\ I_1 <_C I_2$[2];*

[2] Here $sup_{\leq_C} \mathcal{I}(R) = \{I \in \mathcal{I}(R) : \nexists I' \in \mathcal{I}(R) \text{ s.t. } I <_C I'\}$.

- *Is a λ-reduct, with $\lambda \in [0,1]$, if it is a possibilistic reduct and $sup_{I \in \mathcal{I}(R)}(1-\lambda)\mu_{\mathcal{I}(P)}(I) - \lambda\frac{|R|}{|Att|}$ is maximal among all possibilistic reducts;*
- *Is a D-reduct if it is a possibilistic reduct and there is no $R' \subseteq Att$ s.t. both $|R'| \leq |R|$ and $\exists I_1 \in sup_{\leq_D}\mathcal{I}(R), I_2 \in sup_{\leq_D}\mathcal{I}(R').\ I_1 <_D I_2$;*

Given a possibilistic reduct R of a given PDT P, we denote by α^R the maximum α s.t. R is an α-possibilistic reduct of P. We note the following basic properties:

Theorem 1. *The problem of finding all possibilistic reducts (resp. all C-reducts, all λ-reducts for any given value of λ) can be polynomially reduced to the problem of finding all α-possibilistic reducts and α-MDL reducts. In particular:*

- *R is a 0-possibilistic reduct iff it is a possibilistic reduct iff it is a λ-reduct ($\lambda = 1$);*
- *R is a C-reduct iff $\nexists R'$ s.t. both $|R'| \leq |R|$ and $\alpha^{R'} \geq \alpha^R$.*

Proof. As regards possibilistic reducts, it is trivial to show that the collection of possibilistic reducts is the same as the collection of 0-possibilistic reducts. For all other types of reducts, the proof is constructive: we describe an algorithm that finds all α-possibilistic and α-MDL reducts (see Algorithm 1), and show that this procedure can be effectively used (see Algorithms 2, 3) for finding all other types of reducts with no more than polynomial (in the number of reducts) overhead. For a PDT P let $\alpha(P) = \{\alpha' \in (0,1] : \exists x \in U, y \in V_t \text{ s.t. } d(x)_t = \alpha'\}$. The overhead for Algorithm 2 is $O(n^2)$ and for Algorithm 3 is $\Theta(n)$ (where n is the number of reducts). Thus, the main statement of the theorem holds. The other statements can be easily proved. □

Algorithm 1. The brute-force algorithm for finding the α-possibilistic and α-MDL reducts of a possibilistic decision table P.

```
procedure α-POSSIBILISTIC-REDUCTS(P: possibilistic decision table)
    for all α ∈ α(P) in decreasing order do
        poss-reds_α ← Superset-Reducts(S(P)^α)
        MDL-reds_α ← Find-Shortest(poss-reds_α)
    end for
    return poss-reds_α, MDL-reds_α        ▷ The collections of α-possibilistic and α-MDL reducts
end procedure
```

We do not know if a similar technique could also be applied to compute the D-reducts: we leave this as open problem.

As a direct consequence of Theorem 1, we can see that the problem of finding all α-possibilistic (resp. α-MDL) reducts is not harder than finding all superset (resp. MDL) reducts of a given SDT.

Theorem 2. *The problem of finding all α-possibilistic (resp. α-MDL) reducts is no computationally harder than the problem of finding all superset (resp. MDL) reducts. Thus, in particular the problem of deciding whether, given a PDT P and $k \in \mathbb{N}^+$, the α-MDL reducts of P are of size $\leq k$ is Σ_2^P complete.*

Algorithm 2. The algorithm for finding the C-reducts of a possibilistic decision table P.

procedure C-REDUCTS(P: possibilistic decision table)
  $MDL\text{-}reds_\alpha \leftarrow \alpha\text{-}Possibilistic\text{-}Reducts(P)$
  $C\text{-}reds \leftarrow MDL\text{-}reds_1$
  for all $\alpha \in \alpha(P) \setminus \{1\}$ **do**
    for all $r \in MDL\text{-}reds_\alpha$ **do**
      if $\nexists r' \in C\text{-}reds$ s.t. $|r'| < |r|$ **then**
        $C\text{-}reds.append(r)$
      end if
    end for
  end for
  return $C\text{-}reds$ ▷ The set of C-reducts
end procedure

Algorithm 3. The algorithm for finding the λ-reducts of a possibilistic decision table P.

procedure λ-REDUCTS(P: possibilistic decision table, $\lambda \in [0, 1]$)
  $poss\text{-}reds_\alpha \leftarrow \alpha\text{-}Possibilistic\text{-}Reducts(P)$
  $\lambda\text{-}reds \leftarrow \emptyset$
  $\theta \leftarrow 0$
  $map \leftarrow \emptyset$
  for all $\alpha \in \alpha(P)$ in decreasing order **do**
    for all $r \in poss\text{-}reds_\alpha$ **do**
      $\theta\text{-}temp \leftarrow (1 - \lambda)\alpha - \lambda\frac{|r|}{|Att|}$
      $map.append(\langle r, \theta\text{-}temp\rangle)$
      if $\theta\text{-}temp \geq \theta$ **then**
        $\theta \leftarrow \theta\text{-}temp$
      end if
    end for
  end for
  $lambda\text{-}reds \leftarrow$ all $r \in map$
  return $\lambda\text{-}reds$ ▷ The set of λ-reducts
end procedure

Proof. For each α the reduction is trivial, as the problem of finding the α-MDL reducts of P is equivalent to finding the MDL reducts of $\mathcal{S}(P)^\alpha$. Note also that $|\alpha(P)| \leq |U||V_t|$: this implies that the problem $\alpha\text{-}MDL\ Reduct$ requires, in the worst case, a polynomial (in the size of the PDT P) number of calls to a procedure for checking MDL Reducts. This can also be easily seen from Algorithm 1. □

Despite this result, showing that finding minimal reducts (that is, α-MDL, C-reducts or λ-reducts) for a PDT is not harder than finding MDL reducts for a SDT (which, in turn, is no harder than finding minimal reducts for a classical DT), all the reduct search problems considered require worst-case exponential time (in the size of the PDT). Indeed, while heuristics could be applied to speed

up the computation of reducts [29] (specifically, to reduce the complexity of the *find-shortest-reducts* step in Algorithm 1) the proposed algorithms still require enumerating all the possible instantiations. Therefore, in the following section, we propose an alternative definition of reducts in order to reduce the computational costs.

3.3 Entropy Reducts

Following [4] we discuss an alternative definition of reduct, based on the notion of entropy [27], which simplifies the complexity of finding a reduct for a SDT. Given a SDT S with decision d, and $W \subseteq V_t$, we can define a basic belief assignment as

$$m(W|[x]_B) = \frac{|\{x' \in [x]_B : d(x') = W\}|}{|[x]_B|}. \tag{7}$$

Let P be a PDT, $\alpha \in [0,1]$, $B \subseteq Att$ be a set of attributes and denote by $IND_B = \{[x]_B\}$ the equivalence classes (granules) with respect to B. Let $d^{\alpha}_{[x]_B}$ be the restriction of d on the equivalence class $[x]_B$ for the derived SDT $\mathcal{S}(P)^{\alpha}$, and let $m(\cdot|[x]^{\alpha}_B)$ the corresponding bba. Then, we define the OAU entropy of d, conditional on B and possibility degree α, as:

$$OAU(d|B,\alpha) = \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} OAU(m(\cdot|[x]^{\alpha}_B)) \tag{8}$$

That is, the OAU entropy of a PDT (conditional on a set of attributes B and a possibility degree α) is obtained by first computing the derived SDT $\mathcal{S}(P)^{\alpha}$, and then computing the (weighted) average of the OAU entropies of the bbas (see Eq. 2) determined by the granules $\{[x]_B : x \in U\}$. Based on the OAU entropy of a PDT, we can define entropy reducts for PDTs:

Definition 7. *Let $B \subseteq Att$ be a set of attributes, $\alpha \in [0,1]$. Then, we say that B is:*

- *An α-OAU super-reduct if $OAU(d \,|\, B, \alpha) \leq OAU(d \,|\, Att, \alpha)$;*
- *An α-OAU reduct if no proper subset of B is also a α-OAU super-reduct;*
- *An α-OAU ϵ-approximate super-reduct, with $\epsilon \in [0,1)$, if $OAU(d \,|\, B, \alpha) \leq OAU(d \,|\, Att, \alpha) - log_2(1-\epsilon)$;*
- *An α-OAU ϵ-approximate reduct if no proper subset of B is also an α-OAU ϵ-approximate super-reduct.*

Let $[x]_B$ be one of the granules with respect to an α-OAU reduct. Then, the *α-OAU instantiation* with respect to $[x]_B$ is given by

$$dec_{OAU(B,\alpha)}([x]_B) = \arg\max_{v \in V_t} \left\{ p(v) \,|\, p \in \arg\min_{p \in P_{Bel}} H(p) \right\}, \tag{9}$$

that is, (one of) the most probable among the classes under the probability distributions which corresponds to the minimum value of entropy.

Table 1. An example of possibilistic decision table

	w	x	v	z	d
x_1	0	0	0	0	0
x_2	0	0	0	1	$\{0:0.5, 1:1.0\}$
x_3	0	1	1	0	0
x_4	0	1	1	1	$\{0:1.0, 1:0.5\}$
x_5	0	1	0	1	1
x_6	0	1	0	0	$\{0:0.5, 1:1.0\}$

Example 1. Let $P = \langle U = \{x_1, ..., x_6\}, A = \{w, x, v, z\}, d \rangle$ be the PDT in Table 1. We have that $\alpha(P) = \{0.5, 1\}$, thus in particular $\mathcal{S}(P)^{0.5}$ assigns $\{0, 1\}$ to objects x_2, x_4, x_6; while $\mathcal{S}(P)^1$ assigns 1 to objects x_2, x_6 and 0 to object x_4.

We have $OAU(d \,|\, A, 0.5) = OAU(d \,|\, B, 0.5) = 0$ for $B = \{x, v\}$. Also, it holds that $OAU(d \,|\, \{x\}, 0.5) = OAU(d \,|\, \{v\}, 0.5) > 0$. Thus, B is a 0.5-OAU reduct of SDT. We note that $\{z\}$ is also a 0.5-OAU reduct since, similarly, $OAU(d \,|\, z, 0.5) = 0$.

The 0.5-OAU instantiation given by $\{x, v\}$ is $dec_{x,v}(\{x_1, x_2\}) = 0$, and, similarly, $dec_{x,v}(\{x_3, x_4\}) = 0$ (since for objects $x_1, ..., x_4$ the instantiation with minimal OAU value is the one where all objects are assigned the label 0), while $dec_{x,v}(\{x_5, x_6\}) = 1$. By contrast, 0.5-OAU instantiation given by $\{z\}$ is $dec_z(\{x_1, x_3, x_6\}) = 0$, $dec_z(\{x_2, x_4, x_5\}) = 1$.

There is a single 0.5-MDL instantiation, that is $dec_{MDL}(\{x_1, x_3, x_6\}) = 0$, and $dec_{MDL}(\{x_2, x_4, x_5\}) = 1$, which corresponds to the 0.5-MDL reduct $\{z\}$. Thus, in this case, the 0.5-MDL reduct is equivalent to a 0.5-OAU reduct.

As regards $\mathcal{S}(P)^1$, we note that the decision attribute d is single-valued (hence, there is a single instantiation) and the corresponding DT is consistent. In this case there is a single reduct, namely $C = \{x, v, z\}$: therefore C is the only 1-MDL reduct and the only 1-OAU reduct.

Therefore we have that the set of MDL reducts is equivalent to the set of 0.5-MDL reducts (i.e. $\{\{z\}\}$); while the set of C-reducts is $\{\{z\}, \{x, z, v\}\}$; on the other hand we notice that the set of λ-reducts (for varying λ) is structured as follows:

$$\begin{cases} \{z\} & \lambda \geq 0.5 \\ \{\{z\}, \{x, z, v\}\} & \lambda = 0.5 \\ \{\{x, z, v\}\} & 0 \leq \lambda < 0.5 \end{cases}$$

Note that the set of λ-reducts and C-reducts (and possibilistic reducts, by extension) can include two sets $R, R' \subseteq Att$ s.t. $R \subset R'$ as long as they correspond to two different instantiations of the PDT from which they are derived.

In Example 1, the set of α-MDL reducts was exactly the set of minimal (w.r.t. size) α-OAU reducts: this is not a coincidence, we can show that this is a general property of OAU reducts on consistent PDTs.

Theorem 3. *Let P be a PDT, $\alpha \in (0,1]$ and assume that $\mathcal{S}(P)^\alpha$ is consistent. Then, the set of consistent α-possibilistic reducts (i.e., the α-possibilistic reducts whose corresponding instantiations are consistent) coincides with the set of α-OAU reducts. Thus, in particular:*

1. *The set of consistent α-MDL reducts coincides with the set of minimal α-OAU reducts;*
2. *Finding the sets of consistent C-reducts and λ-reducts (for all values of λ) can be reduced to finding the set of α-OAU reducts for all values of α;*
3. *Finding the minimal α-OAU reducts is Σ_2^P-complete.*

Proof. We show the proof only for the main statement: the other statements directly follow from the definition of α-MDL reducts and from Theorems 1, 2. Indeed, suppose that R is a consistent α-possibilistic reduct: this means that there exists $I \in \mathcal{I}(R)$, instantiation of $\mathcal{S}(P)^\alpha$ that is consistent. As a consequence $OAU(d|R,\alpha) = 0$ and thus R is a α-OAU super-reduct. Suppose, further, that R were not a α-OAU reduct: then $\exists R' \subset R$ s.t. $OAU(d|R',\alpha) = 0$, but this means that R' is a consistent reduct of $\mathcal{S}(P)^\alpha$ which is a contradiction. Therefore R is an α-OAU reduct and the claim follows. □

While, as a consequence of Theorem 3, the complexity of finding minimal α-OAU reducts is the same as finding α-MDL ones, even in the approximate case, the former approach to finding reducts is more amenable to optimization as it does not require an explicit enumeration of the instantiations of the PDT. Furthermore, as this approach relies on a quantitative quality measure (i.e., entropy), simple greedy procedures can be implemented with polynomial time complexity (specifically, $O(m^2 \cdot n)$, where m is the number of attributes and n the number of objects), and the guarantee to find an α-OAU reduct (albeit not necessarily minimal w.r.t. size).

4 Conclusion

In this article we studied the problem of feature selection in the learning from fuzzy label setting, and introduced generalized notions of reducts as well as algorithms for feature selection on the basis of this notion. While this paper provides a promising direction for the application of RST-based feature selection in weakly supervised learning, it naturally leaves many questions open. Specifically, we plan to address the following problems in future works:

- In Theorem 1, we showed that most definitions of reducts in a PDT can be derived from the definition of α-possibilistic reducts. Similar characterization also for D-reducts should be investigated in order to better understand the relationship between the latter and other types of reducts;
- In Theorem 3, we showed the equivalence of α-OAU and α-possibilistic reducts in the consistent case. The relation between these two definitions of reduct in the general, non-necessarily consistent case, should also be investigated;

– The definitions of reducts considered in this article, being based on the Pawlak definition of rough approximations, can only be applied to discrete data: thus, the generalization of the proposed approaches to encompass RST techniques that can be applied to continuous data (neighborhood-based or fuzzy-rough approaches) should be investigated.
– We plan to evaluate the performance of the proposed reduct definitions on real PDTs: These, in turn, can be obtained from multi-rater annotations, or through self-labeling techniques [11].

References

1. Arora, S., Barak, B.: Computational Complexity: A modern Approach. Cambridge University Press, Cambridge (2009)
2. Bello, R., Falcon, R.: Rough sets in machine learning: a review. In: Wang, G., Skowron, A., Yao, Y., Ślęzak, D., Polkowski, L. (eds.) Thriving Rough Sets. SCI, vol. 708, pp. 87–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54966-8_5
3. Campagner, A., Ciucci, D.: Orthopartitions and soft clustering: soft mutual information measures for clustering validation. Knowl.-Based Syst. **180**, 51–61 (2019)
4. Campagner, A., Ciucci, D., Hüllermeier, E.: Feature reduction in superset learning using rough sets and evidence theory. In: Lesot, M.J., et al. (eds.) IPMU 2020. CCIS, vol. 1237, pp. 471–484. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50146-4_35
5. Ciucci, D., Forcati, I.: Certainty-based rough sets. In: Polkowski, L., et al. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10314, pp. 43–55. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60840-2_3
6. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recogn. **42**(3), 334–348 (2009)
7. Couso, I., Borgelt, C., Hullermeier, E., Kruse, R.: Fuzzy sets in data analysis: from statistical foundations to machine learning. IEEE Comput. Intell. Mag. **14**(1), 31–44 (2019)
8. Couso, I., Dubois, D., Sánchez, L.: Random sets and random fuzzy sets as ill-perceived random variables. SpringerBriefs in Computational Intelligence (2014)
9. Denoeux, T.: A k-nearest neighbor classification rule based on dempster-shafer theory. In: Yager, R.R., Liu, L. (eds.) Classic Works of the Dempster-Shafer Theory of Belief Functions. Studies in Fuzziness and Soft Computing, vol. 219, pp. 737–760. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-44792-4_29
10. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst. **122**(3), 409–424 (2001)
11. El Gayar, N., Schwenker, F., Palm, G.: A study of the robustness of KNN classifiers trained using soft labels. In: Schwenker, F., Marinai, S. (eds.) ANNPR 2006. LNCS (LNAI), vol. 4087, pp. 67–80. Springer, Heidelberg (2006). https://doi.org/10.1007/11829898_7
12. Greco, S., Matarazzo, B., Slowinski, R.: Rough sets theory for multicriteria decision analysis. Eur. J. Oper. Res. **129**(1), 1–47 (2001)

13. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approx. Reason. **55**(7), 1519–1534 (2014)
14. Hüllermeier, E.: Does machine learning need fuzzy logic? Fuzzy Sets Syst. **281**, 292–299 (2015)
15. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intell. Data Anal. **10**(5), 419–439 (2006)
16. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: Appice, A., Rodrigues, P.P., Santos Costa, V., Gama, J., Jorge, A., Soares, C. (eds.) ECML PKDD 2015. LNCS (LNAI), vol. 9285, pp. 260–275. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23525-7_16
17. Liu, L., Dietterich, T.: Learnability of the superset label learning problem. In: ICML, pp. 1629–1637 (2014)
18. Nakata, M., Sakai, H.: An approach based on rough sets to possibilistic information. In: Laurent, A., Strauss, O., Bouchon-Meunier, B., Yager, R.R. (eds.) IPMU 2014. CCIS, vol. 444, pp. 61–70. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08852-5_7
19. Nguyen, H.T., Walker, C., Walker, E.A.: A First Course in Fuzzy Logic. CRC Press, Boca Raton (2018)
20. Ning, Q., He, H., Fan, C., Roth, D.: Partial or complete, that's the question. arXiv preprint arXiv:1906.04937 (2019)
21. Orlowska, E. (ed.): IncompleteIinformation: Rough Set Analysis. Physica (2013)
22. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**(5), 341–356 (1982)
23. Quost, B., Denoeux, T.: Clustering and classification of fuzzy data using the fuzzy em algorithm. Fuzzy Sets Syst. **286**, 134–156 (2016)
24. Sakai, H., Liu, C., Nakata, M., Tsumoto, S.: A proposal of a privacy-preserving questionnaire by non-deterministic information and its analysis. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1956–1965. IEEE (2016)
25. Sakai, H., Nakata, M., Yao, Y.: Pawlak's many valued information system, non-deterministic information system, and a proposal of new topics on information incompleteness toward the actual application. In: Wang, G., Skowron, A., Yao, Y., Ślęzak, D., Polkowski, L. (eds.) Thriving Rough Sets. SCI, vol. 708, pp. 187–204. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54966-8_9
26. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
27. Ślęzak, D.: Approximate entropy reducts. Fundam. Inform. **53**(3–4), 365–390 (2002)
28. Ślęzak, D., Dutta, S.: Dynamic and discernibility characteristics of different attribute reduction criteria. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 628–643. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_49
29. Thangavel, K., Pethalakshmi, A.: Dimensionality reduction based on rough set theory: a review. Appl. Soft Comput. **9**(1), 1–12 (2009)
30. Trabelsi, S., Elouedi, Z., Lingras, P.: Dynamic reduct from partially uncertain data using rough sets. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS (LNAI), vol. 5908, pp. 160–167. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10646-0_19
31. Umans, C.: On the complexity and inapproximability of shortest implicant problems. In: Wiedermann, J., van Emde Boas, P., Nielsen, M. (eds.) ICALP 1999. LNCS, vol. 1644, pp. 687–696. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48523-6_65

32. Yao, Y.Y., Lingras, P.J.: Interpretations of belief functions in the theory of rough sets. Inf. Sci. **104**(1–2), 81–106 (1998)
33. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **1**(1), 3–28 (1978)
34. Zhou, Z.-H.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**(1), 44–53 (2018)

Rough-set Based Genetic Algorithms for Weakly Supervised Feature Selection

Andrea Campagner( ) and Davide Ciucci

Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano – Bicocca, Viale Sarca 336 – 20126, Milan, Italy
a.campagner@campus.unimib.it

Abstract. In this article, we study the problem of feature selection under weak supervision, focusing in particular on the *fuzzy labels* setting, where the weak supervision is provided in terms of possibility distributions over candidate labels. While traditional Rough Set-based approaches have been applied for tackling this problem, they have high computational complexity and only provide local search heuristic methods. In order to address these issues, we propose a global optimization algorithm, based on genetic algorithms and Rough Set theory, for feature selection under fuzzy labels. Based on a set of experiments, we illustrate the effectiveness of the proposed approach in comparison to state-of-the-art methods.

Keywords: Weak supervision · Feature selection · Fuzzy labels · Genetic algorithms · Rough sets

1 Introduction

Learning from fuzzy labels [9,12] is a weakly supervised learning problem, in which each instance x is associated with a possibility distribution μ over candidate labels, having an epistemic semantics: only one of the labels is the correct one and μ, then, describes the possibility degree of the labels. For example, an image could be tagged with $\{\text{car} : 1, \text{bus} : 0.8, \text{bicycle} : 0.0\}$: the picture then depicts either a *car* or a *bus*, and *car* is deemed more plausible than *bus*.

In the recent years, increasing interest has been devoted to the development of algorithms for the learning from fuzzy labels task [7,9,12,17]. Even though these techniques can be effective on small-scale benchmarks, they can fail to scale to more complex and higher-dimensional problems, as their generalization ability depends (without further assumptions) on the dimensionality of the feature space [3]. While *feature selection* or *data dimensionality* methods could be helpful in mitigating this issue, their development has mostly been ignored.

While Rough Set-theoretic approaches have been applied effectively to address the above mentioned issues for other weakly supervised learning problem [4], their extension to the learning from fuzzy labels case [3] is more difficult,

© Springer Nature Switzerland AG 2022
D. Ciucci et al. (Eds.): IPMU 2022, CCIS 1602, pp. 761–773, 2022.
https://doi.org/10.1007/978-3-031-08974-9_60

due to increased computational complexity costs and the local heuristic nature of the greedy algorithms currently existing in the literature.

To address these limitations, in this article which represents a continuation of our previous work in this line of research [3], we propose a global optimization approach that combines Rough Set-based feature selection with genetic algorithms to solve the feature selection from fuzzy labels problem. In Sect. 2, we provide the necessary background knowledge on possibility theory and Rough Set theory. In Sect. 3, we first introduce the generalization of Rough Set theory to the learning from fuzzy labels setting, as well as the existing methods for performing feature selection in this setting, and then we introduce the proposed genetic algorithm-based approach and discuss its properties; in Sect. 4 we illustrate the effectiveness of the proposed method on a comprehensive set of benchmarks; finally, in Sect. 5, we summarize our results and describe some open problems.

2 Background

In this section, we recall basic notions of rough set theory (RST) and possibility theory, which will be used in the main part of the article.

2.1 Possibility Theory

Possibility theory is a theory of uncertainty which allows for the quantification of degrees of possibility on the basis of a fuzzy set [20]. We recall that a fuzzy set (equivalently, a possibility distribution) F can be seen as a function $F : X \mapsto [0, 1]$, that is, a generalization of the characteristic function representation of classical sets. A possibility measure is a function $pos_F : 2^X \mapsto [0, 1]$ such that

1. $pos_F(\emptyset) = 0$ and $pos_F(X) = 1$,;
2. if $A \cap B = \emptyset$ then $pos_F(A \cup B) = max(pos_F(A), pos_F(B))$.

Thus, every possibility measure is associated with a fuzzy set F, s.t. $pos_F(A) = max_{x \in A} F(x)$: F is the possibility distribution associated with pos_F. We will focus on *normal* possibility distributions, that is on possibility distributions F such that $\exists x \in X$, $F(x) = 1$. Given $\alpha \in [0, 1]$, the *alpha*-cut of F is defined as $F^{\alpha} = \{x \in X : F(x) \geq \alpha\}$, while the *strong* α-cut is defined as $F^{\alpha+} = \{x \in X : F(x) > \alpha\}$.

In this article, we will adopt the epistemic interpretation [8] of possibility theory, in which possibility distributions represent the degrees of belief (of an agent) w.r.t. a set of possible alternatives. We refer the reader to [12] for a discussion of epistemic possibility distributions in Machine Learning.

2.2 Rough Set Theory

Rough set theory has been proposed by Pawlak [16] as a framework for representing and managing uncertain data, and has since been widely applied for various problems in the ML domain (see [1] for an overview and survey).

A decision table (DT) is a triple $DT = \langle U, Att, t\rangle$ such that U is a universe of objects and Att is a set of *attributes* employed to represent objects in U. Each attribute $a \in Att$ is a function $a : U \to V_a$, where V_a is the domain of values of a. Moreover, $t \notin Att$ is a distinguished *decision* attribute, which represents the target label (or, decision) associated with each object in the universe.

Given $B \subseteq Att$, we can define the *indiscernibility relation* with respect to B as xI_Bx' iff $\forall a \in B$, $a(x') = a(x)$. Clearly, it is an equivalence relation that partitions the universe U in equivalence classes, also called *granules of information*, $[x]_B$. Then, the *indiscernibility partition* is denoted as $\pi_B = \{[x]_B \,|\, x \in U\}$.

We say that $B \subseteq Att$ is a *decision reduct* for DT if $\pi_B \leq \pi_t$ (where the order $\leq$ is the refinement order for partitions, that is, π_t is a coarsening of π_B) and there is no $C \subsetneq B$ such that $\pi_C \leq \pi_t$. Then, evidently, a reduct of a decision table represents a set of non-redundant and necessary features: therefore, reduct search can be understood as a process of feature selection. We say that a reduct R is *minimal* if it is among the smallest (with respect to cardinality) reducts. We remark that, given a decision table, the problem of finding minimal reducts is in general NP-HARD.

3 Rough Set-Based Weakly Supervised Feature Selection

In this section we recall the basic definitions regarding the generalization of Rough Set theory to the fuzzy labels setting, and discuss the existing feature selection methods for this setting. Then, we introduce the proposed weakly supervised genetic rough set feature selection method and we discuss its properties.

3.1 Possibilistic Decision Tables and Reducts

In this work, we will refer to the approach for Rough Set-based weakly supervised feature selection proposed in [4]. For other approaches to generalize Rough Set Theory to the case of imprecise data, we refer the reader to [6,15,18].

In the *learning from fuzzy labels* setting, each object $x \in U$ is generally not associated a single annotation $t(x) \in V_t$. Instead, each such object x is associated with a possibility distribution $\pi(x)$, which describes the state of knowledge of the annotating agent (either human or computational): in particular, $\pi(x)_y$ represents the relative plausibility of label y being the true annotation associated with x (as compared to other labels y'). These notions can be modeled within RST by generalizing the definition of a decision table:

Definition 1. *A* possibilistic decision table *(PDT) is a tuple $P = \langle U, Att, d, t\rangle$, where $d : U \mapsto [0,1]^{|V_t|}$ is a collection of normalized possibility distributions. $t : U \mapsto V_t$ is the true decision attribute, i.e. it is a function s.t. $\langle U, Att, t\rangle$ is a DT and s.t. the* weak superset property *w.r.t. d holds: $d(x)_{t(x)} > 0$ for all $x \in U$.*

As mentioned in the definition, t is the true decision attribute: for each object x, its true label is $t(x)$. However, t is assumed to be unknown and only the

possibility distribution $d(x)$ is available. In regard to this latter, if $|d(x)^{0+}| > 1$ for some $x \in U$, then the correct decision $t(x)$ is not known precisely. Note that, by the weak superset property, the true label $t(x)$ is never considered impossible. Furthermore, if $d(x)_a > d(x)_b$ then a is considered more plausible than b for object x.

A PDT can be associated with a collection of compatible (standard) decision tables, called *instantiations* of the PDT:

Definition 2. *An* instantiation *of a PDT $P = \langle U, Att, t, d\rangle$ is a standard decision table $T = \langle U, Att, t'\rangle$ such that $d(x)_{t'(x)} > 0$ for all $x \in U$. The collection of instantiations of P is denoted $\mathcal{I}(P)$. In particular, $\langle U, Att, t\rangle \in \mathcal{I}(P)$.*

Thus, the collection $\mathcal{I}(P)$ contains all standard decision tables that are compatible (i.e., should not be considered impossible) with the imprecise knowledge descibed by the possibility distribution d. Furthermore, the collection $\mathcal{I}(P)$ inherits a ranking from the definition of the possibilistic decision attribute d:

Definition 3. *Let $I_1, I_2 \in \mathcal{I}(P)$ be two instantiations of a PDT P. Then we say that I_1 is* (conservatively) less possible *than I_2, denoted $I_1 \leq_C I_2$, if:*

$$min_{x\in U} d(x)_{t'}^{I_1} \leq min_{x\in U} d(x)_{t'}^{I_2} \tag{1}$$

We say that I_1 is dominated in possibility *by I_2, denoted $I_1 \leq_D I_2$, if:*

$$\forall x \in U.\ d(x)_{t'}^{I_1} \leq d(x)_{t'}^{I_2} \tag{2}$$

where, in both definitions $d(x)_{t'}^{I_i}$ refers to the value of the decision attribute d (in P) on the label $t'(x)$ in the instantiation I_i.

So as to capture not only the simplicity of the induced model (that is, the size of the reducts), but also the epistemic information encoded by the possibility distribution d, we [3] considered the following definitions of reducts:

Definition 4 ([3]). *For each $\alpha \in (0, 1]$ and PDT P, let P^α be the α-cut of P, that is $P^\alpha = \langle U, Att, t, d^\alpha\rangle$, where $\forall x \in U, d^\alpha(x) = \{y \in V_t : d(x)_y \geq \alpha\}$. For each set of attributes $R \subseteq Att$, denote by $\mathcal{I}(R) \subseteq \mathcal{I}(P)$ the collection of instantiations of P for which R is a reduct. Then, $R \subseteq Att$:*

- *Is an α-possibilistic reduct if it is a reduct for some instantiation of P^α, and an α-MDL reduct if it is a size-minimal α-possibilistic reduct;*
- *Is a C-reduct if it is a possibilistic reduct and $\nexists R' \subseteq Att$ s.t. both $|R'| \leq |R|$ and $\exists I_1 \in sup_{\leq_C}\mathcal{I}(R), I_2 \in sup_{\leq_C}\mathcal{I}(R').\ I_1 <_C I_2$[1];*
- *Is a λ-reduct, with $\lambda \in [0, 1]$, if it is a possibilistic reduct and $sup_{I\in\mathcal{I}(R)}(1 - \lambda)\mu_{\mathcal{I}(P)}(I) - \lambda\frac{|R|}{|Att|}$ is maximal among all possibilistic reducts;*
- *Is a D-reduct if it is a possibilistic reduct and there is no $R' \subseteq Att$ s.t. both $|R'| \leq |R|$ and $\exists I_1 \in sup_{\leq_D}\mathcal{I}(R), I_2 \in sup_{\leq_D}\mathcal{I}(R').\ I_1 <_D I_2$;*

[1] Here $sup_{\leq_C}\mathcal{I}(R) = \{I \in \mathcal{I}(R) : \nexists I' \in \mathcal{I}(R) \text{ s.t. } I <_C I'\}$.

Remarkably, the problem of finding C-reducts and λ-reducts can be (polynomially) reduced to the problem of finding the α-possibilistic reducts:

Theorem 1 ([3]). *The problem of finding all C-reducts (resp., λ-reducts, for any given value of λ) can be polynomially reduced to the problem of finding all α-MDL reducts (resp., α-possibilistic reducts), for all values of α. In particular, all the problems in the statement are in NP-HARD.*

Even though the reduct search problems in Theorem 1 are unlikely to be computationally feasible [4], a local search greedy algorithm whose runtime is $O(|U|^2|Att|^2)$ has been proposed to find approximated C-reducts or λ-reducts [3]. This latter approach, however, suffers from several limitations. First, it is only a local search algorithm, therefore it does not provide any guarantee about the quality of its results. Second, though polynomial, the complexity of this approach is quadratic in both the number of attributes and the number of objects. Consequently, it doesn't scale-well to big data or high-dimensional tasks.

3.2 Genetic Rough Set Selection

The definition of C-reducts, λ-reducts and D-reducts (see Def. 4) is intimately tied to the notion of an instantiation of a PDT. The complexity of finding a reduct for a PDT, therefore, could be understood as stemming from the large size of the search space of all such instantiations. In this section, we show how genetic algorithms can be used to effectively harness the structure of the above mentioned search space, by providing an efficient global search algorithm. In particular, we aim to show (as described also through the results shown in Sect. 4) that a simple global search strategy is sufficient to out-perform the local search methods previously proposed in the literature: for this reason, the approach we propose grounds on basic genetic operators, and does not employ more advanced strategies such as elitism or diversity control.

In the proposed approach, each candidate solution is represented as a pair $\langle I, F\rangle$, where $I \in V_t^{|U|}$ is a vector of decision labels s.t. $\forall x \in U, d(I_x) > 0$, and $F \in \{0,1\}^{|Att|}$. Intuitively, I represents a candidate instantiation, while F is a corresponding candidate reduct: in particular if $F_a = 1$, then attribute a is included in the candidate reduct. We next define the adopted fitness functions, the mutation and crossover criteria, and the selection algorithm.

In regard to the fitness function, we consider three different functions, in order to take into account the differences among C-reducts, λ-reducts and D-reducts. Namely, the three fitness functions are defined as:

$$Fitness_C(\langle I, F\rangle) = \langle r, p\rangle, \tag{3}$$

$$Fitness_\lambda(\langle I, F\rangle) = (1-\lambda)p - \lambda\frac{r}{|Att|}, \tag{4}$$

$$Fitness_D(\langle I, F\rangle) = \langle r, s\rangle, \tag{5}$$

where $p = \min_{x \in U} d(x)_{I_x}$, $r = \begin{cases} |F| & F \text{ is a super-reduct for } (U, Att, I) \\ \infty & \text{otherwise} \end{cases}$, and $s \in [0,1]^{|U|}$ is a vector s.t. $s_x = d(x)_{I_x}$. Note, in particular, that only $Fitness_\lambda$ is single-valued, while the other two fitness functions are multi-valued. Consequently, for these latter two fitness functions we will consider an approach based on multi-objective optimization. In particular, given two candidate solutions $\langle I_1, F_1 \rangle$, $\langle I_2, F_2 \rangle$ we say that:

$$\langle I_1, F_1 \rangle \geq_C^F \langle I_2, F_2 \rangle \text{ iff } r_1 \leq r_2 \wedge p_1 \geq p_2, \tag{6}$$

$$\langle I_1, F_1 \rangle \geq_\lambda^F \langle I_2, F_2 \rangle \text{ iff } Fitness_\lambda(\langle I_1, F_1 \rangle) \geq Fitness_\lambda(\langle I_2, F_2 \rangle), \tag{7}$$

$$\langle I_1, F_1 \rangle \geq_D^F \langle I_2, F_2 \rangle \text{ iff } r_1 \leq r_2 \wedge \forall x \in U,\ s_x \geq s_x. \tag{8}$$

Given these definitions, selection is performed by non-dominated tournament selection [14], as described in Algorithm 1.

Algorithm 1. The selection algorithm.

procedure NON-DOMINATED TOURNAMENT SELECTION (P: population, t: tournament size, c: reduct type)
  $T \leftarrow t$ randomly selected candidate solutions from P
  $a \leftarrow$ randomly selected candidate solution in T
  for all $b \in T$ **do**
    if $b >_c^F a$ **then**
      $a \leftarrow b$
    end if
  end for
  return a ▷ A non-dominated candidate solution
end procedure

In regard to mutation, this is performed separately on the possibility degrees and on the candidate reducts. Specifically, in regard to candidate reducts, features are removed or added randomly according to a Bernoulli distribution with parameter b_{mut}. By contrast, possibility degrees are mutated according to a two step procedure: first, for each instance x, a binary value is randomly sampled from a Bernoulli distribution with parameter b_{mut}; then, if the above mentioned value was equal to 1, a new possibility degree is sampled from the probability distribution $\hat{Pr}_{d(x)}$, given by the possibility-probability transform [10] $\hat{Pr}_{d(x)}(y) = \int_0^{d(x)_y} \frac{d\alpha}{|\{y' \in V_t : d(x)_{y'} \geq \alpha\}|}$. In particular, we decided to adopt this sampling distribution as it is the maximally uncertain distribution among all possible probability distributions Pr compatible with $d(x)$, i.e., satisfying $\forall y \in V_t, Pr(y) \leq d(x)_y$ and $d(x)_y \geq d(x)_{y'} \implies Pr(y) \geq Pr(y')$, and hence has minimum bias [10]. The mutation algorithm is summarized in Algorithm 2.

Finally, single-point crossover is applied to I and F, separately. The complete pseudo-code for the proposed method is reported in Algorithm 3.

Algorithm 2. The mutation algorithm.

procedure MUTATION($\langle I, F\rangle$: candidate solution, b_{mut} : mutation probability)
  for all $a \in \{0, ..., |Att| - 1\}$ **do**
    if $Uniform(0, 1) \leq b_{mut}$ **then**
      $F_a \leftarrow 1 - F_a$
    end if
  end for
  for all $x \in \{0, ..., |U| - 1\}$ **do**
    if $Uniform(0, 1) \leq b_{mut}$ **then**
      $I_x \leftarrow$ random label sampled from $\hat{Pr}_{d(x)}$
    end if
  end for
  return $\langle I, F\rangle$ ▷ A new candidate solution
end procedure

Algorithm 3. The proposed weakly supervised genetic rough set feature selection algorithm.

procedure WEAKLY SUPERVISED GENETIC ROUGH SET SELECTION($\langle U, Att, d\rangle$: PDT, *Popsize*: population size, b_{mut}, c: reduct type, t: tournament size)
  $Pop \leftarrow Popsize$ randomly initialized candidate solutions
  $Best \leftarrow \emptyset$
  while Not converged and termination criterion not reached **do**
    Compute fitness according to $Fitness_c$
    $Best \leftarrow$ the non-dominated candidate solutions in $Pop \cup Best$
    $NewPop \leftarrow \emptyset$
    for all $i = 1$ to $Popsize$ **do**
      $NewPop.append(Non - DominatedTournamentSelection(Pop, t, c))$
    end for
    $Pop \leftarrow NewPop$
    Apply $Crossover$ on Pop
    Apply $Mutate$ on Pop
  end while
  return $Best$ ▷ The best candidate solutions
end procedure

We next study the convergence and complexity properties of the proposed method. The following result shows that, asymptotically, Algorithm 3 is guaranteed to return all C-reducts (resp., λ-reducts, D-reducts).

Theorem 2. *Let n be the number of iterations for which Algorithm 3 runs before termination. Let P be a PDT. If $n \to \infty$, then almost surely $\exists R \in Best$ s.t. R is a C-reduct. Furthermore, $Best = C(P)$ almost surely, where $C(P)$ is the collection of C-reducts for PDT P. The same result holds for λ-reducts, D-reducts.*

Proof. We prove the result for C-reducts, as the case of λ-reducts and D-reducts is equivalent. A C-reduct corresponds, by definition, to a non-dominated candidate solution according to order $\leq_C^F$. Since at least one C-reduct is guaranteed

to exist, $Pr(\exists R \in Best : R \text{ is a } C\text{-reduct}) > 0$. Since as, $n \to \infty$, Algorithm 3 is guaranteed to visit all the non-dominated candidate solutions in the search space (since, at each step, each of these candidate solution has non-zero probability of being added to the population), the result follows.

Thus, in the long run, the proposed method is guaranteed to find all the reducts of the desired class. This property provides an advantage w.r.t. the previously described local search methods that, by contrast, do not provide any such guarantee. Nonetheless, we note two limitations of the previous result: 1) the previous result only holds asymptotically, with no bounds on the expected number of iterations required to achieve convergence; 2) the previous results holds irrespective of the population size, thus, as a degenerate case, also for purely random search. While it is reasonable to expect that larger, or even adaptive, population size could improve speed of convergence, we leave the development of such results as open problem.

The computational complexity of Algorithm 3 can be characterized as follows:

Theorem 3. *Let n be the number of iterations for which Algorithm 3 runs before termination. Then, the complexity of Algorithm 3 is $O(n|U||Att|)$.*

Proof. The mutation and crossover steps have both complexity $O(|U|+|Att|)$ per iteration. The per-iteration complexity of the selection step is $O(|U|)$. The per-iteration complexity required for computing the fitness of the candidate solutions in the population is $O(|U||Att|)$. Therefore the result follows.

The previous theorem ensures that, as long as the number of iterations is $o(|U||Att|)$, the proposed method has better computational complexity than the local search methods described in [3]. We leave as open problem the definition of algorithm to automatically tune the number of iterations, based on the available data, so as to guarantee quick convergence with high probability.

4 Experiments and Results

In this section, we discuss the experiments that we designed to evaluate the proposed method, in comparison with other feature selection methods for the learning from fuzzy labels problem, and present and discuss the obtained results.

4.1 Experimental Design

In order to evaluate the proposed genetic algorithm-based feature selection methods we considered a benchmark suite encompassing 14 different datasets:

- Two fuzzy-labeled datasets, previously described, respectively, in [2] and [5];
- 12 datasets from the UCI repository. The precise labels for these datasets were fuzzified by means of nearest neighbors label smoothing [13]. In particular,

setting the number of neighbors equal to k, for each instance x, the associated possibility distribution is obtained as:

$$d(x)_y = \frac{|\{i \in \{1,\ldots,k\} : N_i \text{ nearest neighbor of } x \wedge t(N_i) = y\}| + 1_{t(x)=y}}{k+1}$$

The full list of datasets, including he number of instances, features and classes, is reported in Table 1.

Table 1. List of datasets

Dataset	Instances	Features	Classes
Avila	20768	10	10
Car	864	16	4
Crowd	10845	28	6
Frog family	7195	22	4
HCV	582	12	4
Iranian	7032	45	2
Mushroom	5644	99	6

Dataset	Instances	Features	Classes
Myocardial	1700	111	2
Pen	10992	16	10
Sensorless	20000	48	11
Taiwan	6819	94	2
Wifi	2000	7	4
CTC	617	2500	2
Kyphosis	120	14	7

We considered the following 3 feature selection algorithms:

- The proposed Rough Set-based genetic algorithms, considering the three fitness functions λ, C, D, denoted respectively as $GRSSL, GRSSC, GRSSD$. We selected, in particular, a budget of $n = 1000$ iterations;
- The greedy Rough Set-based local search algorithms for λ-reducts, C-reducts and D-reducts, denoted respectively as $RSSL, RSSC, RSSD$;
- The DELIN algorithm [19], as a comparison baseline. This latter is a dimensionality reduction algorithm based on linear discriminant analysis, whose runtime is $O(|U||Att|^2)$

Performance was evaluated by means of the following experimental design:

1. For each dataset, split in training set Tr (70%) and test set Ts (30%);
2. Apply the feature selection algorithms on the training set Tr, obtaining a reduct F and the reduced training set Tr_F;
3. Train a kNN classifier on the reduced training set Tr_F;
4. Evaluate the trained kNN classifier on the reduced test set Ts_F.

In regard to performance measures, we measured the balanced accuracy, so as to take into account the label imbalance in the considered datasets. The algorithms were also compared in terms of running time. Differences among algorithms (if any) were analyzed by means of a statistical testing approach. In particular, we applied the Friedman rank test to evaluate whether some global statistically significant difference existed among the considered algorithms, and then applied the post-hoc Nemenyi test for pair-wise comparisons of performance. In both cases, p-values smaller than 0.05 were considered to be significant evidence of performance difference (at a confidence level of 95%).

4.2 Results and Discussion

The results of the experiments are reported in Fig. 1 and 2; both in terms of average ranks and p-values for the post-hoc test. As shown in Fig. 1, the three proposed genetic algorithms reported the best average ranks w.r.t. balanced accuracy, and were significantly more accurate than the Rough Set-based local search algorithm. In particular, the GRSSL algorithm (that is, the genetic rough set selection for λ-reducts) was the feature selection algorithm with best performance, being significantly more accurate than all other algorithms. By contrast, while significantly better than the Rough Set-based local search methods, GRSSC and GRSSD were better than the baseline DELIN only on average. We conjecture that this could be due to the intrinsic complexity of the underlying multi-objective optimization problems: indeed, for example, GRSSD involves the solution of a $|U|+1$-dimensional problem. Future research should thus be devoted at exploring more advanced techniques to address this aspect.

Also in terms of running time, the proposed genetic algorithms compare favorably with the Rough Set-based local search methods, as shown in Fig. 2. We note, though, that DELIN had better run-time than all other considered algorithms, being significantly more efficient than GRSSD, while the difference w.r.t. GRSSL and GRSSC was not significant. This could be explained by two different reasons. First, DELIN only uses matrix operations in its execution, that can be performed very efficiently through numerical linear algebra libraries; by contrast Rough Set-based algorithms also include table manipulation operations, having higher computational costs. Second, the complexity of DELIN is $O(|U||Att|^2)$, while for the genetic algorithms the complexity is $O(n|U||Att|)$, with the iterations' budget set to $n = 1000$ iterations: for all datasets, except CTC, n was much greater than $|Att|$, thus the effective complexity of GRSSL, GRSSC and GRSSD was $o(|U||Att|^2)$. We remark, however that, despite this difference, the proposed

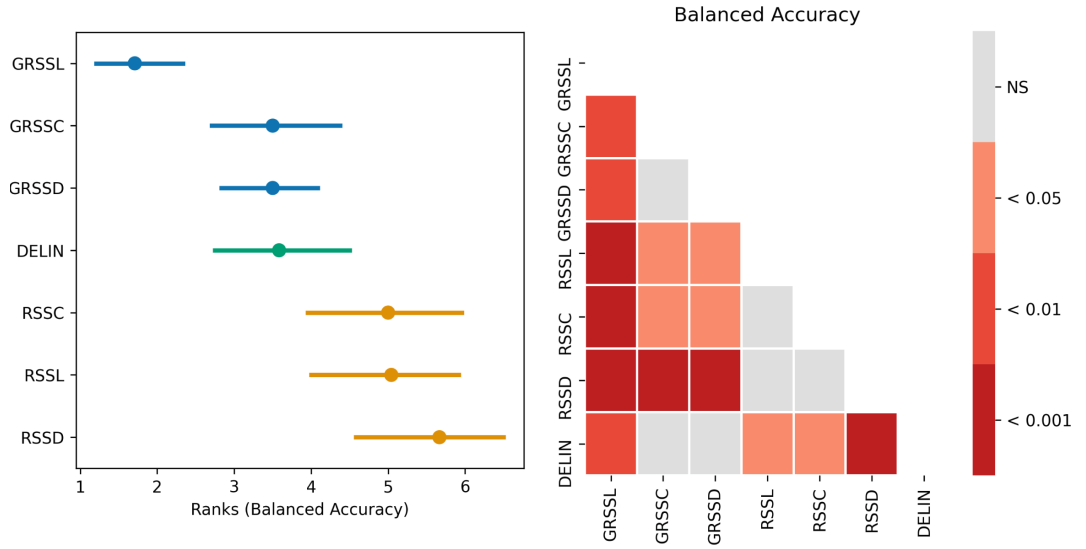


Fig. 1. Average ranks and p-values for balanced accuracy.

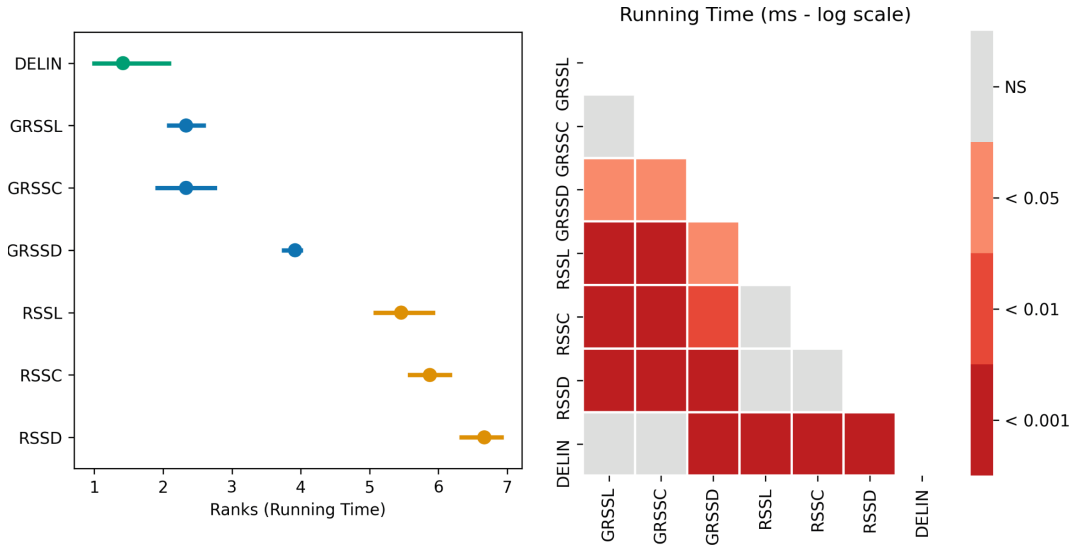


Fig. 2. Average ranks and p-values for running time.

genetic algorithms were significantly more accurate. Nonetheless, future work should be devoted at exploring algorithmic strategies to automatically control the number of iterations n, based on the available data.

5 Conclusion

In this article, we studied the problem of feature selection in the learning from fuzzy label setting, and proposed a method, combining genetic algorithms and Rough Set theory, for efficiently solving this problem. We studied the computational properties of the proposed method and showed its effectiveness in comparison to existing feature selection methods, on a comprehensive set of benchmarks. While this paper provides a promising direction for the application of RST-based feature selection in weakly supervised learning, it naturally leaves many questions open. Specifically, we plan to address the following problems:

- In the proposed method, we did not apply any advanced multi-objective optimization techniques, such as diversity control, elitism or Pareto strength assignment [14]. Future work should evaluate the potential benefits of including such techniques in the proposed method;
- In Theorem 2 we showed that the proposed method is a global search strategy: asymptotically, the best solutions found by the genetic algorithm are exactly the desired reducts. Further research should study non-asymptotic characterizations of the proposed method, in terms of expected time to convergence, or PAC population bounds [11];
- Similarly, as long as the number of iterations n is constant or upper-bounded by the size of the PDT P, the complexity of the proposed algorithm is particularly favourable w.r.t. the standard Rough Set greedy algorithm. Further

research should be devoted at exploring the relationship between n and the quality of the returned solutions.

References

1. Bello, R., Falcon, R.: Rough sets in machine learning: a review. In: Wang, G., Skowron, A., Yao, Y., Ślęzak, D., Polkowski, L. (eds.) Thriving Rough Sets. SCI, vol. 708, pp. 87–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54966-8_5
2. Campagner, A., Cabitza, F., Berjano, P., Ciucci, D.: Three-way decision and conformal prediction: isomorphisms, differences and theoretical properties of cautious learning approaches. Inf. Sci. **579**, 347–367 (2021)
3. Campagner, A., Ciucci, D.: Feature selection and disambiguation in learning from fuzzy labels using rough sets. In: Ramanna, S., Cornelis, C., Ciucci, D. (eds.) IJCRS 2021. LNCS (LNAI), vol. 12872, pp. 164–179. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87334-9_14
4. Campagner, A., Ciucci, D., Hüllermeier, E.: Rough set-based feature selection for weakly labeled data. Int. J. Approx. Reasoning **136**, 150–167 (2021)
5. Campagner, A., Ciucci, D., Svensson, C.M., Figge, M.T., Cabitza, F.: Ground truthing from multi-rater labeling with three-way decision and possibility theory. Inf. Sci. **545**, 771–790 (2020)
6. Ciucci, D., Forcati, I.: Certainty-based rough sets. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10314, pp. 43–55. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60840-2_3
7. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recogn. **42**(3), 334–348 (2009)
8. Couso, I., Dubois, D., Sánchez, L.: Random Sets and Random Fuzzy Sets as ill-perceived Random Variables. SpringerBriefs in Computational Intelligence (2014)
9. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst. **122**(3), 409–424 (2001)
10. Dubois, D., Prade, H., Sandri, S.: On possibility/probability transformations. In: Fuzzy Logic, pp.103–112. Springer (1993)
11. Hernández-Aguirre, A., Buckles, B.P., Martínez-Alcántara, A.: The probably approximately correct (PAC) population size of a genetic algorithm. In: Proceedings of ICTAI 2000, pp. 199–202. IEEE (2000)
12. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approx. Reasoning **55**(7), 1519–1534 (2014)
13. Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S.: Does label smoothing mitigate label noise? In: ICML, pp. 6448–6458. PMLR (2020)
14. Luke, S.: Essentials of Metaheuristics. Lulu, 2nd (edn.) (2013)
15. Nakata, M., Sakai, H.: Rule induction based on rough sets from possibilistic data tables. In: Seki, H., Nguyen, C.H., Huynh, V.-N., Inuiguchi, M. (eds.) IUKM 2019. LNCS (LNAI), vol. 11471, pp. 86–97. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14815-7_8
16. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**(5), 341–356 (1982)

17. Quost, B., Denoeux, T.: Clustering and classification of fuzzy data using the fuzzy em algorithm. Fuzzy Sets Syst. **286**, 134–156 (2016)
18. Sakai, H., Wu, M., Nakata, M.: Apriori-based rule generation in incomplete information databases and non-deterministic information systems. Fundamenta Informaticae **130**(3), 343–376 (2014)
19. Wu, J.-H., Zhang, M.-L.: Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. In: Proceedings of the 25th ACM SIGKDD, pp. 416–424 (2019)
20. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **1**(1), 3–28 (1978)

Chapter 4

Applications of Learning from Imprecise Data

In the previous chapters the main focus has been on the development and study of practical algorithms for dealing with imprecision in the input of ML models, focusing in particular on the case of learning from fuzzy label, as well as on the empirical evaluation of existing techniques on general benchmarks. By contrast, the main aim of this chapter will be to illustrate the application of techniques for learning from imprecise data in practical, real-world settings. Indeed, despite rapid development of algorithmic approaches for these types of problems, the practical evaluation of such approaches on real-world data and problems has rarely been considered and reported about in the specialized literature: relevant examples in this sense include applications in the setting of learning from crowdsourced data [64, 240, 256] as well as in the setting of physical device modeling [209]. To this purpose, this chapter will describe the application of the RRL algorithm proposed in the previous chapters, or variations thereof, to three medical problems: the problem of ground-truthing in multi-rater labeling, the problem of handling vague ordinal terminology in medical datasets, and the problem of handling data affected by individual variation.

In the first case, in Section 4.1, the main focus will be on a relevant problem emerging in medical decision making tasks, in which the ground truth annotations to be used as target supervision by a ML model are not guaranteed to be correct, as the

involved annotators (humans or otherwise) are not necessarily infallible, either due to limited expertise or to the inherent complexity of the annotation task. In this setting, one of the main approaches to counter the above mentioned problem is to collect and aggregate labels produced by many raters who independently provide their labeling of the dataset, similarly to what happens in crowdsourcing [36]. Drawing from the vast literature concerned with learning in crowdsourcing scenarios [275], several state-of-the-art techniques have been proposed for multi-rater ground truthing and learning [136, 198, 227, 226], most of which based on adaptations of the majority aggregation scheme (i.e. the process of selecting for each instance a single, optimal label by majority voting, or a variations thereof, on the multi-rater ground truth and then applying standard or regularized machine learning models), described in Algorithm 4. The main objective of the first section will be to illustrate the application of

Algorithm 4 Generic aggregation-based algorithm for learning from multi-rater labels.

procedure AGGREGATION-BASED MULTI-RATER LEARNING($S = \{(x_1, (y_1^1, \ldots, y_1^k)), \ldots (x_m, (y_m^1, \ldots, y_m^k))$: multi-rater training set, A: aggregation rule, $\mathcal{H}$: model class)

$S_{temp} \leftarrow \emptyset$

$w \leftarrow \emptyset$

for all $r \in \{1, \ldots, k\}$ **do**

end for$w[r] \leftarrow$ Compute weight of rater r based on S

for all $(x, (y^1, \ldots, y^k) \in S$ **do**

$y^* \leftarrow$ aggregate $A(y^1, \ldots, y^k; w)$ based on w

Add (x, y^*) to S_{temp}

end for

$h \leftarrow$ optimal model in $\mathcal{H}$ w.r.t. S_{temp}

return h

end procedure

learning from imprecise data techniques to the problem of learning from multi-rater data, focusing in particular on the application of the RRL algorithm to set-label

or fuzzy-labeled data obtained by imprecisiation of the original multi-rater datasets [152], showing promising results compared to state-of-the-art approaches.

Section 4.2 of this chapter will be focused, in contrast, on the problem of learning from vague ordinal data in medical datasets. In contrast to the case of imprecise data, vague data refers to data representations, usually typical of linguistic data, that may be interpreted differently by different agents, due to the inherent ambiguity or arbitrariness of their semantics [130]. As in the case of imprecise data, standard ML and statistical data analysis techniques cannot be naively applied when dealing with vague data, which generally requires the careful application of ad-hoc techniques [144]. This type of data occurs frequently in the medical domain, especially when referring to medical terminology: a particularly relevant example of this issue is the use of *severity* ordinal scales [212] (i.e. ordinal scales that are used to express the severity of a medical condition in a standardized format) which, while widely used as standard ordinal features in ML-based medical studies, have been shown to be inherently affected by vagueness [10, 41]. The aim of this section will then be to illustrate the application of learning from fuzzy data techniques to the setting of learning from vague ordinal data: to this purpose a general-purpose, questionnaire-based, technique to transform vague data into imprecise data (modeled in the form of possibility distributions) will be illustrated, which will then be used to ground the application of learning from fuzzy data techniques, including an adaptation of the RRL learning algorithm discussed previously, showing that these techniques can provide significant advantages in comparison with standard encoding approaches as well as with approaches based on the use of fuzzy set theory to deal with vague data.

Finally, Section 4.3 will be devoted the application of learning from imprecise data techniques to the problem of managing data affected by within-subject (or, individual) variation [107]. This latter refers to a widely known phenomenon in the biomedical domain, denoting variation and noise in the values of a set of features of interest that is due not to population differences or errors, but rather to the intrinsic and characteristic patterns of variation pertaining to a given individual [189] or the measurement process [258]. Individual variation has been deemed critical for proper

analysis of medical data, and may have relevent implications for data analysis. Indeed, data which may affected by individual variation, even when pertaining with a single individual, may no longer be represented as a point estimate $(x_1, ..., x_d)$, bu rather only as a cloud of possible realizations that are all compatible with the characteristic of the given individual [106]. Thus, in the specialized literature, it is usually assumed that each individual features' distribution can then be represented by a d-dimensional Gaussian $N_x = (x, \Sigma_x)$, where x denotes an averaged, characteristic representation of the individual, called *value at the homeostatic point*, and Σ_x is a covariance matrix denoting the degree of individual variation for each of the features of interests, and their dependencies. Thus, any ML model trained on a single snapshot of data for a given set of patients, may fail to generalize not only when given data pertaining to new patients, but also when given different instantiations of the data pertaining to the same patients considered in its training set, whenever individual variation causes these latter to be sufficiently different for the model to fail to classify them correctly [200]. Despite these characteristics, the impact of individual variation on the development of ML models has seldom been considered in the literature. The aim of this section, then, will be to describe the problem of handling individual variation in ML models with a two-fold objective: first, to show that standard ML models can fail to be robust and generalize properly when confronted with data affected by individual variation; second, to show how techniques for learning from imprecise data, and specifically so adaptations of the generalized nearest neighbors and RRL algorithms to the more general setting of learning from fuzzy data, can be useful in this setting. In particular, the problem of individual variation management will be addressed within the context of COVID-19 diagnosis from laboratory blood exams, so as to rely on the vast literature concerning the estimation of individual variation parameters for this type of data [1, 58, 210].

Information Sciences 545 (2021) 771–790

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Ground truthing from multi-rater labeling with three-way decision and possibility theory

Andrea Campagner [a,*], Davide Ciucci [a], Carl-Magnus Svensson [b], Marc Thilo Figge [b,c], Federico Cabitza [a]

[a] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336 – 20126 Milano, Italy
[b] Applied Systems Biology Research Group, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute (HKI), Jena, Germany
[c] Institute of Microbiology, Faculty of Biological Sciences, Friedrich-Schiller-University Jena, Germany

ARTICLE INFO

Article history:
Received 18 March 2020
Accepted 20 September 2020
Available online 28 September 2020

Keywords:
Machine learning
Multi-rater
Three-way decision
Possibility theory
Uncertainty

ABSTRACT

In recent years, Machine Learning (ML) has attracted wide interest as aid for decision makers in complex domains, such as medicine. Although domain experts are typically aware of the intrinsic uncertainty around it, the issue of *Ground Truth (GT) quality* has scarcely been addressed in the ML literature. GT quality is regularly assumed to be adequate, regardless of the number and skills of raters involved in data annotation. These factors can, however, potentially have a severe negative impact on the reliability of ML models. In this article we study the influence of GT quality, in terms of number of raters, their expertise, and their agreement level, on the performance of ML models. We introduce the concept of *reduction*: computational procedures by which to produce single-target GT from multi-rater settings. We propose three reductions, based on *three-way decision*, *possibility theory*, and *probability theory*. We provide characterizations of these reductions from the perspective of learning theory and propose two ML algorithms. We report the result of experiments, on both real-world medical and synthetic datasets, showing that GT quality strongly impacts on the performance of ML models, and that the proposed algorithms can better handle this form of uncertainty compared with state-of-the-art approaches.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

In recent years, interest in Machine Learning (ML) technologies and systems supporting human decision makers has greatly increased, especially in light of the promising results achieved to date in many decision-intensive domains. A paradigmatic example of such a domain is medicine [41]. One factor that led to the adoption of ML in the medical domain is the large amount of data routinely collected and stored, ranging from simple scalar measurements (e.g., blood pressure), to 3D images (e.g., from magnetic resonance imaging).

Uncovering hidden patterns in data that support human decisions during diagnosis and treatment planning requires the data to be enriched with target annotations that specify their "true nature". Original data and their annotations constitute the "Ground Truth" (GT) that ML models are trained on so that they can subsequently replicate or predict correct associations in new, unseen, instances of data. In the experimental sciences, these target annotations, also called *labels*, are typically

* Corresponding author.
E-mail address: a.campagner@campus.unimib.it (A. Campagner).

https://doi.org/10.1016/j.ins.2020.09.049
0020-0255/© 2020 Elsevier Inc. All rights reserved.

generated under controlled conditions ensuring that a given instance can be associated with a certain class [44]: the resulting GT labels can therefore be considered certain. In the medical domain however, the GT labels are typically generated by human raters who annotate the available data based on their interpretation and skills. Besides it being a time-consuming and error-prone task, human annotation must contend with *inter-observer variability*, which occurs when raters interpret the same instance differently; this phenomenon has been reported and discussed extensively [16,20] and related to the *reliability* of the resulting data. Despite the apparent problem with uncertainties in GT, scholars in the ML and AI communities seldom address the question of how reliable their GT actually is [4,5,42]. We have termed this question the "elephant in the record" [6,3] in an attempt to promote awareness of an enormous issue in ML in medicine, which, although widely recognized, is rarely addressed in the specialist literature.

One approach to counter the potential fallibility of human annotators is to create GT labels from a number of annotations, produced by many raters who independently provide their labeling of the dataset. For many applications, it is both feasible and convenient to involve people from the general public as annotators; these are "crowd sourcing" or "citizen science" projects. In the medical domain, however, raters must be specialists; even so, their degree of competence, which can vary, is very important for the quality of the resulting annotations. Information on annotator competence level, however, is not usually reported in medical ML studies. Sometimes it is stated that medical professionals were involved in producing the GT, but details such as the number of raters and their level of expertise are omitted; or it may be that different raters were associated with different subsets of the dataset [15,17]. In other cases, the number of raters is limited to three [2,18], or the raters were not medical specialists but lay-people specifically trained for the task [42]. This situation is understandable, as the time of highly trained and specialized medical professionals, such as dermatopathologists or licensed ophthalmologists, is very valuable; for these reasons, involving the huge numbers of annotators typical of crowdsourcing settings would be unfeasible in the medical domain.

Having acknowledged the problem of GT uncertainty, it is crucial to investigate the implications of creating GTs involving few annotators [2,18], in terms of both reliability and how effects on the robustness of ML models that are trained using such annotations. The overall goal is to investigate novel methods to mitigate potential shortcomings and make the best possible use of the available annotations. Even in the ideal situation, where one is able to involve many raters with high level of expertise, to take the ground truthing process seriously requires knowing how to deal adequately with inter-observer variability that will necessarily affect the resulting annotations. We focus specifically on the key step by which the original, vector-valued, multi-rater representation is converted into a single label or value. This *reduction* step is typically performed by selecting the label that the majority of voters have chosen. If the number of raters is small, however, the chance of selecting the wrong label is relevant, as is the resulting label noise [39]. Even if the ML model can learn the relevant patterns in datasets with label noise, the performance of the model will nevertheless be affected, because the GT used to test the model will also have the same level of uncertainty.

In this paper, we will investigate how rater expertise, the number of raters, inter-rater variability, and the way the multi-rater representation is reduced into a single value all affect the classification performance of a ML model. Specifically, in this article we offer the following three main contributions:

- We introduce a general theoretical framework grounding on the concept of *reduction* (the transformation of multi-rater annotations to a single label); we will distinguish between the standard reduction, based on majority voting (majority reduction), and alternative reductions that are based on probability theory, possibility theory and three-way decision (TWD) theory;
- We discuss the theoretical properties of the different reductions from the point of view of learning theory and introduce two algorithms that can be applied to general reductions;
- Finally, we report on an empirical investigation in which we analyzed, using a collection of both real and synthetic datasets, the effects of the number of raters, their expertise, their inter-rater agreement, and the chosen reduction method on the performance of ML models. Specifically, we will show that the proposed algorithms are more effective than the traditional majority voting based approach, especially in circumstances where the raters annotating the dataset do not demonstrate high accuracy.

The rest of this article will be structured as follows: in Section 2, we will introduce the necessary mathematical background, and in Section 3 we will introduce the measures of reliability that we use to quantify the level of agreement between the raters. In Section 4.1, we will define the concept of *reduction* and describe a number of alternative reductions. In Section 4.2, we will define the ML models embedding these reductions, and analyze these learning models from the perspective of learning theory. In Section 4.3, we will introduce the concept of *simulated level of expertise* and apply it to a number of datasets to investigate its effect on the performance of ML models; we will show and discuss the obtained results in Section 5. Finally, in Section 6, we will draw our conclusions and outline some open problems, possible ways to address them, and future work in this strand of research.

2. Mathematical background

2.1. Supervised and multi-rater machine learning

To start, let us assume that in a standard ML environment, our knowledge is available in the form of a *decision table* that describes a set of objects and everything we know about it.

Definition 1. A *decision table* is a tuple $\langle U, A, t\rangle$ where

- U is a universe of objects of interest (e.g. images or medical records).
- A is a set of features that we use to represent objects in U. Specifically, we define each feature as a function $a : U \mapsto V_a$ where V_a is the domain of values that the feature a can assume.
- t, with $t \notin A$, is a decision attribute that represents the target attached to the objects in U.

The learning task can then be described as follows. Let $\mathscr{P}$ be a probability distribution over U and $\mathscr{H}$ a class of hypotheses, i.e., a set of functions $h : U \mapsto V_t$. Assuming that we have an instance-level loss function, $l_t : \mathscr{H} \times U \mapsto \mathbb{R}$, we can define the loss function $L_{t,\mathscr{P}} : \mathscr{H} \mapsto \mathbb{R}$ as

$$L_{t,\mathscr{P}}(h) = \int_U l_t(h, x) d\mathscr{P}(x), \tag{1}$$

the goal of the standard ML classification task [36] is to find a hypothesis $h^*(S)$ such that

$$h^*(S) = argmin_{h\in\mathscr{H}} L_{t,\mathscr{P}}(h), \tag{2}$$

where $S \subseteq U$ is a finite sample (also called a training set).

Notice that in general we do not know $\mathscr{P}$, nor the value of $t(x)\ \forall x \in U$. Therefore we can only estimate the loss via the empirical loss, defined as

$$EL_{t,S}(h) = \sum_{x\in S} l_t(h, x). \tag{3}$$

The goal of the learning process then changes to finding a hypothesis $h^*(S)$ that $\forall\mathscr{P}$ satisfies some bounds on the real loss. These bounds depend on the size of the training sample $|S|$, the complexity of $\mathscr{H}$ as measured by the Vapnik–Chervonenkis (VC) dimension [36] or similar constructs, the empirical loss and other approximation parameters.

In the present work, we deal with a more general learning task, which we call learning from *multi-observer labels*.

Definition 2. A *multi-observer decision table* is a tuple $\langle U, A, t, D\rangle$ where

- U is a universe of objects of interest.
- A is a set of features that we use to represent objects in U. Specifically, we define each feature as a function $a : U \mapsto V_a$ where V_a is the domain of values that the feature a can assume.
- $t \notin A$ is a particular decision attribute, i.e., the target, which we assume to be the true (but possibly unknowable) decision associated with the object in U. We will denote the domain of values of t as V_t; specifically we will here assume that $V_t = \{0, 1\}$.
- D, with $D \cap (A \cup \{t\}) = \varnothing$, is a set of decision attributes that represent the decisions that a set of observers assign to objects in U. Notice that when $D = \varnothing$ and t is known $\langle U, A, t\rangle$ is a decision table.

Remark 1. With an abuse of notation we will write $D(x)$ to denote the decision values assumed by object x, and $D(S)$ to indicate the values assumed by all objects x in S.

This kind of setting was proposed and studied in the crowdsourcing domain. While most work in the literature generally acknowledges the limitations of simple majority voting to transform the original multi–observer table into a standard table, relatively few have proposed corrections to this criterion; usually, such corrections have been by means of either computational procedures that estimate the error rates of the single annotators or information provided by raters in addition to the annotation. For instance, Raykar et al. [39] proposed an iterative approach based on Expectation–Maximization to jointly estimate the annotators' error rates and true labels, and train a logistic regression algorithm; while Whitehill et al. [45] describe a parametric Bayesian approach to estimate annotators' error rates and true labels based on an objective indication of the *difficulty* of each instance. Heinecke et al. [19] describe algorithms based on learning theory by which to estimate the annotators' error rates and perform an accuracy-weighted correction to majority voting; Hertwig [21] discusses a confidence-weighted correction to majority voting (in which annotators are supposed to provide also an estimate of their confidence, together with their proposed label); Prelec et al. [38] consider a mechanism for majority voting correction, called *surprisingly popular algorithm*: in this approach, an estimate of the annotators' competence is made by asking annotators to

report which option, in their opinion, will be the most popular. These methods were developed for GT obtained by crowdsourcing or similar approaches and, thus, show several limitations in settings such as medicine (the reference setting for our research). The most prominent of these are the need for an additional source of *objective* or *subjective* information [21,38,45] (such as objective evaluations of case complexity, or raters' confidence) that may be difficult to elicit, the need to involve large number of raters (or annotated examples) to obtain accuracy rates significantly greater than those achieved via simple majority voting [19,39], or, more generally, no significant improvements over simple majority voting.

A similar setting has been studied in the Rough Set Theory (RST) literature under the term *multi-source information systems*. For instance, Sang et al. [40] studied rule learning in this setting using decision-theoretic rough sets; however, that setting differs from that which we consider here in that it does not typically consider annotations' quality.

Most approaches involving multi–observer tables require transforming the original multi–observer table into a standard table, where each instance is associated with a single label (which is plausibly the most appropriate one). To formalize this concept, the notion of *reduction* was introduced in [4] as here:

Definition 3. A *reduction* is a transformation $T : V_D \mapsto \mathscr{C}(V_t)$ that maps each multi-observer labeling to a structure over labels in V_t.

Here, "structure" is intended in a general sense; examples include: a single labeling, thus $\mathscr{C}(V_t) = V_t$, or a probability distribution over labels.

With respect to the ML task, which call *learning from reductions*, the goal is to simultaneously find a reduction T and a hypothesis h^* minimizing the loss with respect to the true, but unknown, labeling t. Note that, in general, each of the raters $d \in D$ may disagree on some subset $S_d \subseteq S$ with the true labeling t, and thus it may also happen that we have $\forall d \in D. t(x) \neq d(x)$ for a given $x \in S$. The task of learning from multi-observer labels can therefore be seen as a generalization of the task of learning from noisy labels [1], as the reduction may produce the wrong label.

Finally, we also recall the pseudo-labels [28] approach, which has been widely applied in semi-supervised learning: this approach uses synthetic labels for the unsupervised instances, and updates them iteratively on the basis of the output of an underlying classifier. Since its proposal this approach has shown remarkable results in solving different ML tasks: semi-supervised learning [32], ensemble learning [46] and, most relevant to this paper, learning from noisy labels [30]. As we show in Section 4.2, the algorithms that we propose to address the learning from reductions task are based on the pseudo-label approach, as they involve training an ensemble of models on labels sampled from a prior distribution (which is specified by the reduction employed). Despite this similarity, the two approaches differ in their implementation details: in the pseudo-labels approach the probability distribution over the labels is iteratively updated on the ground of the output provided by the underlying classification algorithm. Therefore, the learning algorithm is used to estimate the correct class assignment probability. By contrast, in the "learning from reductions" task, the distribution over labels is established a priori (from the labels provided by the raters and the specific reduction employed) and is then used to train a ML model, by minimizing the loss w.r.t. this prior distribution.

2.2. Introduction to uncertainty representation

Probability theory and related methods, such as information theory, are the most used frameworks to represent and manage uncertainty in ML, due to its solid theoretical foundations and the wide successful application in many fields. However, a variety of alternative or complementary uncertainty management frameworks have been proposed in the literature in order to cope with some shortcomings of probability theory in dealing with phenomena like representation of ignorance [37]. Among them, we will consider possibility theory based on fuzzy sets [49] and three-way decision [48].

Three-way decision (TWD), originally proposed by Yao [48], is a framework for uncertainty management, inspired by human decision-making and Rough Set Theory (RST) [47], that can be understood as an extension of standard decision theory by which objects in the domain of interest are divided in three categories: a positive, or acceptance, region; a negative, or rejection, region; and a further boundary, or non-commitment, category, which represents lack of knowledge or (temporary) abstention concerning the status of the objects placed in this latter region. With respect to the ML setting, according to TWD, every data point can be classified as belonging to the target class, not belonging to the target class, or being in the *boundary*, that is a region that represents lack of knowledge with respect to class assignment. Despite its apparent simplicity, TWD has been successfully applied in the ML literature for many tasks, such as classification under uncertain boundaries [31], ensemble construction [50], deep learning [29]. Most relevantly for our purposes, TWD has been applied successfully to manage semi-supervised (or, more generally, weakly supervised) learning: Miao et al. [34] studied a semi-supervised learning approach based on TWD; Min et al. [35] proposed a weakly supervised algorithm based on a tripartition of the instances into positive/negative and uncertain one; and Campagner et al. [7] proposed a TWD-based framework of algorithms for general weakly supervised learning tasks. Interestingly, although the considered tasks and approaches are quite distinct from that which we present in this article (see the previous Sections), TWD has also been applied to multi-observer tasks: Hu et al. [22] studied an approach to information aggregation based on TWD; Huang et al. [23] studied how to employ TWD for information fusion; and Sang et al. [40] employed TWD to manage multi-source information tables (a generalization of multi-observer decision tables). For an extensive review of application of TWD in ML, see [8].

Another popular uncertainty representation is possibility theory, which allows for the quantification of degrees of possibility on the basis of a fuzzy set [49]. A fuzzy set F can be seen as a function $\mu_F : X \mapsto [0,1]$, that is, a generalization of the characteristic function representation of classical sets. A possibility measure is a function $pos : 2^X \mapsto [0,1]$ such that

1. $pos(\varnothing) = 0 \wedge pos(X) = 1$, and
2. if $A \cap B = \varnothing$ then $pos(A \cup B) = max(pos(A), pos(B))$.

It can be easily seen that every possibility measure and distribution is induced by a fuzzy set F as $pos(A) = max_{x \in A} \mu_F(x)$. A possibility distribution is *normal* if $\exists x \in X. \mu_F(x) = 1$.

A possibility distribution μ_F induces the *necessity measure* $nec : 2^X \mapsto [0,1]$, defined as $nec(A) = 1 - pos(A^c)$ which is thus dual to the measure pos. A possibility measure can be interpreted as *imprecise probabilities* [12], that is as an imprecise specification of our belief given by a set of compatible probability distributions

$$Pr_\mu = \left\{ Pr | \forall A \subseteq X. nec_\mu(A) \leqslant Pr(A) \leqslant pos_\mu(A) \right\}. \tag{4}$$

This epistemic view [12] of fuzzy sets as possibility distributions has led to the active research areas of *statistics on fuzzy data* and *machine learning on fuzzy data*: see Hüllermeier [24] for an introduction to fuzzy ML using generalized loss functions and Couso et al. [11] for a recent survey of the area. In Section 4.2, we build on this literature to study how to employ possibility theory to manage uncertainty in multi-rater settings.

3. Reliability

The intuitive notion of reliability, specifically inter-rater reliability, relates to the extent to which we can trust a GT in making decisions, or better yet, for training models on, which support our decision making. More technically, the reliability of a given multi-rater labeling of a dataset takes the *agreement* among the different raters into account: that is, the precision of the labelings (from a metrological point of view). If all raters agree on every case, the reliability is maximal. Most reliability measures, such as the proportion of agreement and the k, which currently are those most widely employed in empirical studies, have limitations from a theoretical and conceptual point of view [27]. For instance, they do not take into account that agreement could arise by chance or they employ scales for score interpretation that are too permissive and arbitrary.

Because of such limitations, in this article we will consider a more robust reliability measure, known as Krippendorff's α [27]. This is defined as:

$$\alpha_K = 1 - \frac{D_o}{D_e}. \tag{5}$$

In the formula of α_K, D_o and D_e can respectively be defined as:

$$D_o = \frac{1}{|U||D|} \sum_{i,j \in V_t} \delta(i,j) \sum_{x \in U} |D| \frac{c_{i,j}}{\binom{|D|}{2}}, \tag{6}$$

$$D_e = \frac{1}{\binom{|U||D|}{2}} \sum_{i,j \in V_t} \delta(i,j) P_{i,j}, \tag{7}$$

where $\delta(i,j)$ is a distance function and $c_{i,j}$ is defined as:

$$c_{i,j} = |\{d_1, d_2 \in D : d_1(x) = i \wedge d_2(x) = j\}|, \tag{8}$$

and $P_{i,j}$ can be defined as:

$$P_{i,j} = \begin{cases} n_i * n_j & i \neq j \\ n_i * (n_i - 1) & i = j \end{cases}, \tag{9}$$

and $n_i = \sum_x |\{d \in D | d(x) = i\}|$.

In this formulation, the D_o represents the observed degree of disagreement, while D_e represents the disagreement due to chance and is thus employed as a correction factor essentially based on the observed total class proportions. For example, if we have a strong class imbalance, D_e would be small and the factor $\frac{D_o}{D_e}$ would be large, resulting in a low α_K value. It should also be noted that both D_o and D_e are defined on the range $[0,1]$ and thus the resulting α_K can yield values in the range of $[-\infty, 1]$ where $\alpha_K = 1$ corresponds to perfect agreement. The case of $\alpha_K = 0$ corresponds to perfect disagreement (that is, when there is no statistical association between the objects and their labels), while values $\alpha_K < 0$ corresponds to cases where the disagreements are systematic. A further analysis of the α_K measure, including a discussion of its robustness and its desirable properties as a measure of inter-rater agreement can be found in [27]. This publication discusses, among other topics, robust criteria for assess the reliability of a multi-rater decision table, and indicates that a dataset should be considered suf-

ficiently reliable if $\alpha_K \geqslant 80\%$. If needed, one can use a bootstrapping procedure to estimate the distribution of α_K and find a confidence interval $[\alpha_K^{min}, \alpha_K^{max}]$ which is then compared to the 80% threshold.

4. Methods

4.1. Reductions

In this section, we will discuss the concept of *reduction*: any computational procedure that transforms multi-rater decision tables into decision tables with structure-valued decision attributes [4]. We will describe a number of alternative reductions and discuss the epistemic interpretation that can be associated with each of these reductions.

We will assume a binary-valued target: that is, the target of the multi-rater decision table is expressed in terms of a m–dimensional vector over the set $V_t = \{0, 1\}$ (i.e., $D(x) \in V_t^m$), where $m = |D|$. We suppose that the target of the training sample is generated from $x \in S$ via a reduction $T : V_t^m \mapsto \mathscr{C}(V_t)$, where $\mathscr{C}(V_t)$ is a set of structures, in a general sense, over V_t as already defined in Section 2 above. In general, the reduction T involves an information loss because it is impossible to perfectly recover $D(x)$ by observing $T(D(x))$ in the case that $\mathscr{C}(Y) \neq V_t^m$ and $T \neq id_{V_t^m}$. This means that T implicitly defines an inverse set–valued map $L_T : \mathscr{C}(Y) \mapsto \mathscr{P}(V_t^m)$.

Note that we can interpret a reduction T through an epistemic lens: $T[D(x)]$ represents the belief as to which alternatives in V_t are the most plausible ones. We can provide a *quantitative counterpart* to this qualitative concept by considering the *information loss* of each reduction. This can be quantified by considering the inverse L_T of a reduction T, which defines the set of all possible $D(x)$ that satisfy a given requirement. This measure of information loss IL will be defined as

$$IL(D(x), T) = \frac{|L_T(T[D(x)])| - 1}{2^m - 1}. \tag{10}$$

The simplest and most commonly used reduction is the *majority-voting* reduction, or just *majority reduction*, which selects the *mode* from a set of labels. The majority reduction can be defined as $maj : V_D \mapsto V_t$ with

$$maj[D(x)] = argmax_{v \in \{0,1\}} |\{d \in D | d(x) = v\}|. \tag{11}$$

The epistemic stance justifying this reduction is that each rater $d \in D$ has a small error rate $\epsilon_d \ll \frac{1}{2}$ and thus, if $m = |D|$ is sufficiently large, the probability that $maj[D(x)] \neq t(x)$ is negligible. Note, however, that this assumption may be unjustified when the plurality choice has a small margin compared with the other alternatives, or whenever it is difficult to assess and bound the error rate of the raters. We can quantify the information loss of the majority reduction by taking

$$IL(D(x), maj) = O\left(\frac{\sum_{f^*=\lceil m/2 \rceil}^{m} \binom{m}{f^*}}{2^m - 1}\right). \tag{12}$$

In Eq. (12), the numerator term indicates that the only preserved information is the identity of the most frequent label according to the majority reduction.

As described in Section 2.1 above, most work in the multi-rater supervised setting has been devoted to establishing procedures that act as corrections to or improvements over simple majority voting, typically based on methods that estimate the error rate of the raters. We will denote all these methods with the term *Corrected Majority*. As anticipated above, these approaches present some limitations when they are applied to medicine or similarly critical domains. They usually require additional data, beyond the annotations' [21,38,45], and have generally yielded results that differed significantly from the simple majority only when either a large number of annotators or data instances is available. We conjecture that these limitations are due to these methods' focus on finding corrections to the problems of simple majority voting in a crowdsourcing setting, with a strong requirement of compatibility with the traditional supervised setting and learning algorithms.

To avoid these limitations, we consider four different reductions, first proposed in [4]. These attempt to preserve more information than the majority reduction, by adopting more expressive representations, based on standard *uncertainty representation frameworks*, namely probability theory, possibility theory and three-way decision.

The *probabilistic reduction* is defined as $prob : V_D \mapsto Pr(V_t)$, where $Pr(V_t)$ are the probability distributions over V_t, with

$$prob[D(x)] = \langle 0 : \frac{|D_0(x)|}{|D|}, 1 : \frac{|D_1(x)|}{|D|} \rangle, \tag{13}$$

where $D_v(x) = \{d \in D | d(x) = v\}$. Compared with the majority reduction, the probabilistic reduction is more *information-conservative*. Indeed, the probabilistic reduction preserves the frequency of occurrence of each alternative in V_t and forgets only which rater proposed a given label. The information loss of the probabilistic reduction is

$$IL(D(x), prob) = \frac{\binom{m}{|D_1(x)| - 1}}{2^m - 1}, \tag{14}$$

and it can easily be established that $IL(D(x), prob) < IL(D(x), maj)$.

As a "hybrid" reduction, which in terms of informativeness can be placed between the majority and probabilistic reductions, we also introduce the *overwhelming majority reduction*, which is defined as $over : V_D \mapsto Pr(V_t)$ with

$$over[D(x)] = \begin{cases} \langle maj[D(x)] : 1 \rangle & \text{if } max_{v \in \{0,1\}} |D_v| \geqslant \tau \\ sm(\mathbf{1} - \langle P_e(0), P_e(1) \rangle) & otherwise. \end{cases} \tag{15}$$

In Eq. (15) we have that τ is a parameter, $\mathbf{1}$ is the constant 1 vector, $sm : \mathbb{R}^k \mapsto [0,1]^k$ is the softmax function

$$sm(\langle p_1, \ldots, p_k \rangle) = \langle e^{p_1}, \ldots, e^{p_k} \rangle * \frac{1}{\sum_{i=1}^{k} e_i^p}, \tag{16}$$

and $P_e(v)$ is an estimate of the probability that v is not the real value of $t(x)$. If, for instance, we assume that the raters have a constant error rate ϵ, then in the binary case $V_t = \{0,1\}, P_e(v) = \epsilon^{D_v}(1-\epsilon)^{D_{1-v}}$.

The *fuzzy-possibilistic reduction* is defined as $fuzzy : V_D \mapsto \mathscr{F}(V_t)$, where $\mathscr{F}(V_t)$ is the collection of fuzzy sets definable over V_t defined as

$$fuzzy[D(x)] = \langle 0 : \frac{|D_0|}{|D_{v^*}|}, 1 : \frac{|D_1|}{|D_{v^*}|} \rangle, \tag{17}$$

where $v^* = argmax_{v \in V_t} |D_v|$.

The fuzzy reduction can be seen as an expression of the degrees of belief in terms of a relative preference among alternatives. Thus, by means of the fuzzy-possibilistic reduction, we transform a multi-rater labeled instance into *fuzzy data*.

Notice that, although the fuzzy-possibilistic and probabilistic reductions have different epistemic interpretations, their information loss is exactly the same:

Theorem 1. *IL(D(x), fuzzy) = IL(D(x), prob).*

Proof. Let $fuzzy[D(x)] = \langle f_0, f_1 \rangle$ and assume, without loss of generality, that f_0, f_1 are sorted in decreasing order and let $0 = argmax_{v \in V_t} |D_v|$. Noticing that $\sum_i f_i = \frac{\sum_{v \in V_t} |D_v|}{|D_0|} = \frac{m}{|D_0|}$ and we can thus obtain $|D_0| = \frac{m}{\sum_i f_i}$ from which we can obtain $\forall v \in V_t \; |D_v| = f_v * |D_0|$. Consequently we can obtain exactly the result of $prob[D(x)]$ using only the information that was originally available using $fuzzy[D(x)]$. □

Finally, as an alternative approach similar to, but more conservative than, the overwhelming majority reduction, we introduce the *three-way reduction*, which is defined as $tw : V_D \mapsto \mathscr{P}(V_t)$ with

$$tw[D(x)] = \begin{cases} \langle maj[D(x)] : 1 \rangle & \text{if } max_{v \in V_t} |D_v| \geqslant \tau \\ V_t^{\delta} = \{ v \in V_t | \tau \geqslant |D_v| \geqslant \delta \} & otherwise \end{cases}, \tag{18}$$

where τ and δ are two parameters controlling, respectively, the minimum probability threshold that would be required to select a single specific alternative, and the minimum probability threshold required to claim that an alternative would be acceptable. Thus, if there is any strong evidence, measured by τ, toward a specific alternative, then that specific alternative is selected. Otherwise, it is not possible to precisely claim which is the correct alternative. In this case, all of the alternatives supported by a sufficient amount of evidence, as measured by δ, are selected and all others are excluded. Thus, we have three alternatives: to accept 0; to accept 1, and to abstain, as it is typical of TWD. Note also that, according to this epistemic interpretation, $0 < \delta < \tau$ and $\frac{1}{|V_t|} \leqslant \tau$ should hold true.

The information loss for the three-way reduction can be defined as follows:

$$IL(D(x), tw) = \begin{cases} \frac{\sum_{k=\tau}^{m} \binom{m}{k}^{-1}}{2^m - 1} & \exists v \text{ s.t. } |D_v| \geqslant \tau \\ \frac{\sum_{k=\delta}^{\tau} \binom{m}{k}^{-1}}{2^m - 1} & otherwise \end{cases}. \tag{19}$$

In the following example we illustrate the results of the different reductions on a simple multi-rater decision table.

Example 1. Let S be a multi-rater decision table of three cases and

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

the respective labeling given by five observers.

Applying the *maj* reduction, we obtain

$$maj[D(S)] = [\,0 \quad 1 \quad 0\,], \tag{20}$$

and the information loss is

$$IL(D(S), maj) = [\,15/31 \quad 15/31 \quad 15/31\,]. \tag{21}$$

Then, for the probabilistic reduction we obtain

$$prob[D(S)] = \begin{bmatrix} (0:3/5, 1:2/5) \\ (0:1/5, 1:4/5) \\ (0:4/5, 1:1/5) \end{bmatrix}, \tag{22}$$

for which the information loss is

$$IL(D(S), prob) = [\,10/31 \quad 5/31 \quad 5/31\,]. \tag{23}$$

For the fuzzy reduction we get

$$fuzzy[D(S)] = \begin{bmatrix} (0:1, 1:2/3) \\ (0:1/4, 1:1) \\ (0:1, 1:1/4) \end{bmatrix}, \tag{24}$$

and the information loss is $IL(D(S), fuzzy) = IL(D(S), prob)$.

In the three-way reduction, we set $\tau = 4, \delta = 2$ giving

$$tw[D(S)] = [\,\{0,1\} \quad 1 \quad 0\,], \tag{25}$$

for which the information loss is

$$IL(D(S), tw) = [\,19/31 \quad 6/31 \quad 6/31\,]. \tag{26}$$

It can easily be seen that the majority, overwhelming majority, fuzzy-possibilistic (through the possibility-probability transformation, see Section 4.2) and three-way reductions can all be considered robust instances of the general probabilistic reduction. They, however, differ in how they assign probabilities p_0, p_1 to the two possible labels, applying different forms of penalization to control the noisiness of the raw frequencies employed by the probabilistic reduction:

- The majority reduction simply assigns probability equal to 1 to the label with maximum frequency;
- The overwhelming majority assigns exponentially more weight to the maximum frequency label: this probability is directly proportional to the error probability of each of the two labels;
- Similarly, the fuzzy-possibilistic approach assigns more weight to the maximum frequency label; however, this increase is smaller than for the overwhelming majority reduction and is proportional to $||D_0| - |D_1||$;
- The three-way reduction assigns probabilities according to a uniform distribution, after performing a thresholding of the label frequencies according to a threshold τ: the two labels are treated as indistinguishable if $max_{v\in\{0,1\}}|D_v| < \tau$.

Finally, it can be noted that when $max_{v\in\{0,1\}}|D_v| \geqslant \tau$ the results of the majority, overwhelming majority and three-way reduction coincide.

4.2. Learning from reductions

In this Section we define generalized learning paradigms, based on standard optimization and ML approaches, for the reductions defined in Section 4.1 above. We also report basic results with respect to learnability and error bounds from the perspective of Computational Learning Theory for the respective models. A summary of the different reductions and their advantages and limitations, including those of the associated learning setting and taking into account the results described in Section 4.3, is provided in Table 1.

4.2.1. Learning paradigms for reductions

Each reduction provides a transformation from the set of labels, that is, V_t, to a class of structures $\mathscr{C}(V_t)$ over V_t. Therefore it is evident that all reductions that do not result in a single label require a modification to the traditional supervised learning setting. We assume that, for a given reduction T, the optimal hypothesis $h^*(T(S))$ is induced by an algorithm $\mathscr{A}$ defined as a function $\mathscr{A} : \mathscr{P}(U \times \mathscr{C}(V_t)) \mapsto \mathscr{H}$, which represents a generalization of the standard framework that is defined by an algorithm $\mathscr{A} : \mathscr{P}(U \times V_t) \mapsto \mathscr{H}$.

In this work we will consider the following learning tasks, which can be instantiated from a multi-observer decision table by means of a reduction T:

Table 1
Summary of the reduction strategies.

Reduction	Advantages	Limitations
Majority	Simple; default ML models can be used	Not robust when few raters available or majority has low margin
Corrected Majority	More accurate; sample complexity bounds available	More complex implementation; require many annotators/instances; may require additional information
Probabilistic Overwhelming Maj. Fuzzy-Possibilistic Three-way	More accurate; sample complexity bounds available	Ad-hoc ML algorithms needed; exact learning algorithms have high time complexity

1. *Learning from noisy labels*: the employed reduction is the majority reduction, thus $\mathscr{A} : \mathscr{P}(U \times V_t) \mapsto \mathscr{H}$. Note that this case is not, in general, equivalent to a standard classification setting. Except in the trivial case for which $\forall x \in U, \forall d \in D. d(x) = t(x)$, each $d \in D$ has a non-zero error rate and thus the label induced by the majority reduction can be noisy.
2. *Noisy Superset Learning*: in this case the three-way reduction is employed, thus $\mathscr{A} : \mathscr{P}(U \times \mathscr{P}(V_t)) \mapsto \mathscr{H}$. In this case, the main assumption of Superset Learning ($t(x) \in three[D(x)]$) may not hold because there is no guarantee that the label $t(x)$ has a sufficient probability score attached, that is, it has been selected by a sufficient number of raters, and it is thus included in the result of the three-way reduction.
3. *Probabilistic Learning*: in this case, the probabilistic or overwhelming majority reductions are employed, thus $\mathscr{A} : \mathscr{P}(U \times Pr(V_t)) \mapsto \mathscr{H}$.
4. *Learning on fuzzy data* [24,11]: in this case the fuzzy reduction is employed, thus $\mathscr{A} : \mathscr{P}(U \times \mathscr{F}(V_t)) \mapsto \mathscr{H}$.

For each of the learning paradigms described above, we can clearly observe that empirical loss cannot be defined in terms of t, as the true labeling t is not observed in the multi-observer decision table, but only in terms of $T[D(S)]$:

$$EL_{T,S}(h) = \sum_{x_S} l_T(h, x). \tag{27}$$

Thus, compared with traditional supervised or superset learning, in learning from multi-observer tables we have another source of approximation because the real labels for the instances are in general not observed.

In learning from noisy labels, standard ML models, or simple modifications thereof, can be employed [1]. Moreover, *sample complexity* bounds have been established showing that, under reasonable constraints on the error rate, if class $\mathscr{H}$ is *probably approximately correct* (PAC) learnable [36] in the supervised setting, then $\mathscr{H}$ is also PAC learnable in the noisy label setting [1]. In our setting, assuming that each rater $d \in D$ has a constant error rate η_d, the probability that applying the majority reduction leads to an incorrect result (that is, the probability that $T(D(x)) \neq t(x)$) can be expressed via the *Poisson binomial distribution*. Thus, the probability that at least $\lceil \frac{m+1}{2} \rceil$ raters would make an error is given by

$$P(error) = \sum_{k=\lceil \frac{m+1}{2} \rceil}^{m} \sum_{B \in F_k} \Pi_{i \in B} \eta_i \Pi_{j \notin B} (1 - \eta_j), \tag{28}$$

where F_k is the family of all subsets of raters of size k. Applying the Chernoff bound, the inequality can be approximated as

$$P(error) \leqslant e^{-\frac{m+1}{2} log \frac{m+1}{2\mu}}, \tag{29}$$

where $\mu = \sum_i \eta_i$. Thus we can claim that the probability of error decreases exponentially with both increasing the number of raters, and decreasing the expected error rate per rater.

From this estimate, we can directly compute a condition for learnability of the hypothesis class $\mathscr{H}$, based on the results due to Angluin et al. [1].

Theorem 2. *Let $\mathscr{H}$ be a PAC-learnable hypothesis class with VC-dimension c. Then $\mathscr{H}$ is learnable with noisy samples, with probability $1 - \delta$ over the choice of the sample set S and maximum approximation error ϵ, when given $|S| > n_0$ samples with*

$$n_0 = \mathcal{O}\left(\frac{c \cdot log \frac{1}{\delta}}{\epsilon \left(1 - 2e^{-\frac{m+1}{2} log \frac{m+1}{2\mu}}\right)^2} \right). \tag{30}$$

Proof. From Angluin et al. [1] it holds that the sample complexity of learning with noisy examples is $n_0 = \mathcal{O}\left(\frac{c \cdot log \frac{1}{\delta}}{\epsilon(1-2\epsilon_r)^2}\right)$, where ϵ_r is the probability of errors in the labels. From the previous analysis we obtained that $\epsilon_r \leqslant e^{-\frac{m+1}{2} log \frac{m+1}{2\mu}}$, hence the result. □

We also note that the following inverse result was proven by Heinecke et al. [19], in a more general setting, and could be used to obtain a bound on the number of raters required to achieve a given threshold on the maximum error rate.

Theorem 3. [19] *Fix a sample S, a desired maximum error rate δ and an average rater error rate η_O. Then the number m of raters required to guarantee that $P(maj[D(S)] = t(S)) \geqslant 1 - \delta$ is*

$$m = \mathcal{O}\left(\frac{log\frac{|S|}{\delta}}{(1-2\eta_O)^2}\right). \tag{31}$$

In the next section, we will present two general algorithms that could be applied for other general reductions that do not necessarily return single labels.

4.2.2. Learning from non-majority reductions

For each of the overwhelming majority, probabilistic, fuzzy and three-way reductions, we consider two different learning meta-algorithms. The first is derived from a resampling-based ensemble strategy, and thus we will call the corresponding algorithms *probabilistic*, *fuzzy*, *three-way* and *overwhelming majority Resampling-based Reduction Learning* (RRL). This approach was first studied by Jin et al. [26] in the superset learning setting and we provide a generalization to all the settings considered in this paper.

The second learning strategy, which is a generalization of the approach proposed by Cour et al. [10] for the superset learning setting, is based on a weighted ensemble of models. We will call the respective algorithms based on this latter strategy *Score Weighting Reduction Learning* (SWRL).

In both cases, we assume the hypothesis class $\mathscr{H}$ to be composed of *scoring classifiers*, i.e., each $h \in \mathscr{H}$ is actually defined as the composition of two functions $s : X \mapsto \mathbb{R}^{V_t}$ and $dec : \mathbb{R}^{V_t} \mapsto V_t$. In this formulation we stipulate that $h(x) = dec(s(x))$, where *dec* is defined as

$$dec(s(x) = \langle s_1, \ldots, s_k \rangle) = argmax_{v_i \in V_t} s_i. \tag{32}$$

Thus, s maps every instance $x \in X$ to a vector $\langle s_1, \ldots, s_k \rangle$ where s_i is the score assigned to alternative $v_i \in V_t$, and *dec* simply selects the alternative v_i with maximum assigned score. We cast the learning problem in the standard optimization framework defined in Section 2.1 above; we will adopt an approach based on empirical loss minimization.

Consider first the overwhelming majority reduction or, equivalently the probabilistic reduction. In this case, the learning algorithm receives as input, for each instance $x \in S$, a probability distribution over (a subset of) all labels in V_t. According to the RRL meta-algorithm, the goal is to find a hypothesis h in $\mathscr{H}$ minimizing:

$$L_{prob,S}(h) = \frac{1}{|S|}\sum_{x \in S}\sum_{i=1}^{k} p_i l(h(x), v_i), \tag{33}$$

where $prob[D(x)] = \langle p_1 : v_1, \ldots, p_k : v_k \rangle$. The same form of the empirical loss can also be applied for the probabilistic reduction. The name RRL is used because a simple strategy to approximately minimize the empirical loss in this setting involves sampling from the distribution over the labels in order to obtain a set of instantiations of S. That is, we obtain a set of new decision tables $\{DT_i = \langle S_i, A_i, d_i \rangle\}_i$ where $\forall x \in S.\, d_i(x) \in D(x)$ and we then find a hypothesis h_i minimizing the empirical loss over DT_i. In detail, let $x \in S$ and let $T[D(x)] = \langle p_1 : v_1, \ldots, p_k : v_k \rangle$, where T is the probabilistic or overwhelming reduction. The RRL algorithm can be described as the following iterative algorithm:

1. For each iteration i
   (a) For each instance $x \in S$
      i. Sample a label $v \in over[D(x)]$ (resp. $prob[D(x)]$) according to the given probability distribution $\langle p_1, \ldots, p_k \rangle$
   (b) Find h_i minimizing the empirical loss with respect to the current instantiation $S \times over[D(S)]$.
2. Return the hypothesis given by $h(x) = argmax_{v \in V_t} |\{h_i : h_i(x) = v\}|$

The SWRL learning algorithm can be obtained from the RRL algorithm noting that, if the loss function is convex, we can apply Jensen's inequality to push the weights p_i inside the loss function: the hypothesis minimizing this new loss function has an empirical loss which is strictly less than that for the minimizer of Eq. (33). Formally, if $h = dec \circ s$, the goal in the SWRL meta-algorithm is to find the minimizer of the following optimization problem:

$$h^* = argmin_{h \in \mathscr{H}} \frac{1}{S}\sum_{x \in S} l_T\left(\sum_{i=1}^{k} p_i s_i(x), x\right), \tag{34}$$

where $s_i(x)$ is the score assigned for an instance x to the label $v_i \in V_t$.

These approaches can both be directly generalized to the case of the three-way reduction (and hence, to the noisy superset learning setting). Indeed, setting $\forall v_i \in tw[D(x)], p_i = \frac{1}{|tw[D(x)]|}$, the RRL algorithm reduces to the approach proposed by Jin et al. [26], while the SWRL algorithm reduces to the approach proposed by Cour et al. [10]. We also note that, although both are based on the same empirical loss minimization perspective, our proposed approach differs from that of Hüllermeier in [25]. The latter approach adopts an optimistic point of view with respect to the minimization of the loss function:

$$h^* = argmin_{h\in\mathscr{H}} \frac{1}{|S|}\sum_{x\in S} min_{i\in\{1,\dots,k\}} l(h_t(x), v_i). \tag{35}$$

Our approach, instead, adopts a *robust* point of view by requiring that the average of the losses (in the RRL algorithm), or the loss with respect to the average of the prediction scores (in the SWRL algorithm), is minimized. We argue that this approach better adapts to the noisy setting, as it reflects the uncertainty implicit in this learning task, given that the superset assumption could be violated (that is, the real label could be excluded from the value of $T[D(x)]$).

In regard to sample complexity bounds for RRL and SWRL in the binary classification setting, it can be verified that the probability of an error in a label, for the probabilistic reduction, can be upper bounded as

$$P_{prob}(error) = O\left(\sum_{k=\tau*m}^{m}\sum_{B\in F_k}\Pi_{i\in B}\eta_i\Pi_{j\notin B}(1-\eta_j) + \sum_{k=1}^{\tau*m-1}\sum_{B\in F_k}[p_{D_B}(\Pi_{i\in B}\eta_i\Pi_{j\notin B}(1-\eta_j)) + p_{D_{\setminus B}}(\Pi_{i\notin B}\eta_i\Pi_{j\in B}(1-\eta_j))]\right), \tag{36}$$

where p_{D_B} is the probability assigned to the decision expressed by the raters in B and $p_{D_{\setminus B}}$ is the complementary probability. Indeed, as in the general form of the probabilistic reduction $\tau = 1$, the first summand is the probability that all raters provided the wrong label while the second summand can be decomposed in two parts:

- The first part of the sum corresponds to the probability that the label selected by the raters in B is selected via resampling, weighted by the probability that the raters in B chose the wrong label.
- The second part of the sum is symmetric to the first part but with respect to the raters not in B.

Starting from this general error probability we can obtain analogous bounds for the overwhelming majority, $P_{over}(error)$, and three-way, $P_{tw}(error)$, reduction by setting

$$p_{D_B}^{over} = 1 - \frac{a}{a+b}, \tag{37}$$

$$p_{D_B}^{tw} = \frac{1}{2}, \tag{38}$$

where $a = \Pi_{i\in B}\eta_i\Pi_{j\notin B}(1-\eta_j)$ and $b = \Pi_{i\notin B}\eta_i\Pi_{j\in B}(1-\eta_j)$ and τ is set accordingly. Sample complexity bounds can then be obtained from the error bounds in Eq. (30).

Theorem 4. *Let $\mathscr{H}$ be a PAC-learnable hypothesis class with VC-dimension c. Then $\mathscr{H}$ is learnable using the reduction (red) and the respective RRL learning scheme, with probability $1-\delta$ over the choice of the sample set S and maximum approximation error ϵ, when given $|S| > n_0$ samples with*

$$n_0 = \mathcal{O}\left(\frac{c\cdot log\frac{1}{\delta}}{\epsilon(1-2P_{red}(error))^2}\right), \tag{39}$$

where $red \in \{prob, over, three\}$.

Proof. The theorem directly follows from Theorem 2 when setting $P_$red (error) accordingly. □

Moreover, it can be observed that, while the error estimate for these reductions is, in general, larger than the estimate for the majority reduction, these bounds are not tight. If the number of raters $|D|$ is large, then the first term of Eq. (36) can be made arbitrarily small by selecting an appropriately large value for τ. We also see that, for the overwhelming majority reduction, this condition has the effect of making the smallest sub-term of the second term vanishingly small. For the three-way reduction, if $P(error) \to 0$ (hence $t \in tw[D(S)]$ is guaranteed), then the learning problem reduces to superset learning and the sample complexity can be bounded per the result due to Liu et al. [33]. We can also observe that, when the error probability is vanishingly small and the reduction always returns a single labels, then the sample complexity bounds for standard (agnostic) PAC learning apply.

In regard to the fuzzy-possibilistic reduction, approaches to learning from fuzzy data have been proposed by both Hüllermeier [24], based on generalized loss functions and the optimistic empirical loss minimization previously described, and Denoeux [13], based on generalized maximum likelihood. Despite their conceptual simplicity, however, these methods are complex from a computational perspective, as they involve solving generalized loss minimization problems. See the survey by Couso et al. [11] for an overview of the field.

The approach that we propose recasts the learning from fuzzy data problem as a variant of the general probabilistic learning framework and is based on the work on possibility-probability transformations of Dubois et al. [14]. The correspondence between possibility measures and sets of probability distributions highlighted in Section 2.2 above. If we assume that a possibility distribution $fuzzy[D(x)] = \langle 1 = \pi_1 > \dots > \pi_k > \pi_{k+1} = 0\rangle$ represents an imprecise probabilistic distribution, and if we use the *principle of indifference* to select the least informative probability distribution compatible with π, then a probability distribution Pr_π can be obtained as

$$Pr_{\pi}(v_j) = \sum_{i=1}^{k} \frac{\pi_i - \pi_{i+1}}{|A_i|} \mu_{A_i}(v_j), \tag{40}$$

where $A_i = \left\{ v_j \in S | \frac{|D_{v_j}|}{|D_{v^*}|} \geqslant \pi_i \right\}$, i.e., A_i is the π_i-cut. Thus, we can simply implement the fuzzy-RRL and fuzzy-SWRL algorithms by substituting the probability values obtained through the above mentioned transformation into the corresponding probabilistic algorithm. Analogously, it is possible to obtain sample complexity bounds by plugging these values into the error probability estimate of Eq. (36).

From a complexity-theoretic point of view, we note that all the non-majority reduction-based learning approaches presented in this Section have a higher computational complexity than the majority reduction. This is because the associated loss minimization problem essentially iterates over all possible compatible labelings. While in the most general case the time complexity increase could be exponential, the computational costs can be reduced adopting an approximation scheme based on Monte Carlo sampling, as shown for the RRL algorithm. Thus, it is possible to implement the RRL and SWRL algorithms, renouncing to the optimality of the found solution, with a time complexity that is equivalent, up to a constant factor represented by the number of samples, to that of training a traditional model using empirical loss minimization.

4.3. Experimental evaluation

As mentioned, the goal of this article is to investigate the relationships among factors such as the number of raters and their expected expertise, measures of reliability, usage of different reductions and their effects on the accuracy and generalizability of the predictive models. Specifically, we aim to address the following questions:

1. Is the standard practice of considering a small number of raters (normally one or three) adequate for establishing a consensus to be employed as GT in ML applications? What is the influence of rater accuracy (modeled via their error rate η_i), and the influece of inter-rater agreement (quantified using Krippendorff's α), on the quality of the consensus? Ultimately, what is the influence of these factors on the performance of ML models?
2. Could better consensus strategies help alleviate any of the problems previously discussed? Could these strategies improve the accuracy and the generalizability/robustness of the induced models? We will investigate this while considering the influence of the number of raters, their reliability, their expertise and the selection of a specific reduction on the predictive accuracy and robustness of the models.
3. How does rater expertise (modeled using the error rate η_i of the raters) affect predictive performance of the model? Specifically, how is the potential robustness of the different reductions affected by a given proportion, n_c, of non-expert raters?

To address these research questions, we performed an experimental evaluation based on a collection of different datasets. The main challenge in setting up this experiment was the lack of publicly available multi-rater annotated datasets. To avoid this problem, we adopted a private medical dataset, collected by two authors of this study, and described in [42,43], which involved 11 raters. We also considered two standard medical datasets from the UCI repository, which were augmented using a synthetic procedure to obtain multi-rater annotations, and three synthetic datasets. We decided to employ synthetically generated datasets so that we could easily control the error rate of the virtual raters, a prerequisite for addressing our third research question, as listed above. In greater detail, the employed datasets were:

1. Circulating Tumor Cells (CTC) [42,43], 617 instances, 50x50 RGB images, labeled by 11 raters, identification of *Circulating Tumor Cells* from fluorescence microscopy;
2. Breast Cancer: 699 instances, 10 numerical features, single rater, identification of benign or malignant tumors;
3. Diabetes: 768 instances, 8 numerical features, single rater, identification of diabetic patients;
4. Synthetic Gaussians (SG): a synthetic dataset with two Gaussian distributed classes, 1000 instances, 20 features and 21 raters generated using a simple random label flipping procedure.[1] We vary the proportion of non-competent raters, n_c, with $n_c = 0.1, 0.25, 0.5$. A non-competent rater is simulated as a rater with accuracy of $\sim 50\%$. The remaining proportion $(1 - n_c)$ of the raters are considered experts, that is, with accuracy $\sim 95\%$. The three generated pairs of Synthetic Gaussian data with labels from rater populations with $n_c = 0.1, 0.25, 0.5$ will be denoted SG-0.1, SG-0.25 and SG-0.5 respectively.

As the Breast Cancer and Diabetes datasets were annotated by a single rater, we developed a synthetic procedure to simulate a multi-rater scenario by employing 11 Decision Tree classifiers with different hyperparameter settings, trained on boostrapped samples generated from the original dataset.

In the SG datasets, we considered only very accurate versus random raters so that we could easily control the percentage of non-expert raters and thus analyze the influence of this parameter on the performances of the induced models, while

[1] According to the described label flipping procedure, for each instance a given rater has probability p of retaining the correct label and probability $1 - p$ of flipping it.

keeping all other parameters constant. Hence, the average error rate η_O which, as shown in Section 4.2 above, is relevant in the resulting performances of the learning algorithms.

For each dataset, we considered four different reductions: the majority reduction, the overwhelming majority reduction, the three-way reduction and the fuzzy-possibilistic reduction. For the overwhelming majority and three-way reductions, we considered a threshold value of $\tau = 0.73$; this value was selected so that at least 8 raters would be needed to achieve significant majority in the datasets with 11 raters. In addition, for the three-way reduction we selected the confidence parameter $\delta = 0$, given that all considered datasets consist of binary classification problems. We did not evaluate the methods proposed in the crowdsourcing literature (described in the previous sections of the paper), as they were shown to significantly outperform the majority reduction only when large numbers of raters were available (usually > 50); we were primarily interested in studying ground truth and ML model quality when the number of available raters is small (~ 10), as it is typical in medical settings.

For all reductions, we employed the scikit-learn [1] implementation of the Random Forest algorithm with default hyperparameters: the number of estimators was set to 100. We chose Random Forest because it has been shown to be one of the most effective learning algorithms in general settings [9], and also because it is relatively easy to implement and tune. For the majority reduction, we trained a standard Random Forest, while for the other reductions we employed the RRL algorithm.

For the CTC dataset, we considered only data points where all 11 raters had a consensus as the GT to be used during testing, while for the Diabetes and Breast Cancer datasets we used the original labels as testing. In the SG datasets the correct label is defined during the data generation process (as instances are generated from one of two gaussian distributions).

To evaluate the previously described hypotheses for all datasets, we performed the following pipeline:

- For each number of raters $m \in \{1, 3, \ldots, |D| - 2\}$ do:
  1. Generate n samples of size m of $|D|$ where $n = min\left\{ \binom{|D|}{m}, 100 \right\}$. Note that this number of samples allows us to consider a large part of the population (which is the set of all possible combinations of raters, for a given number of raters). Specifically, when $\binom{|D|}{m} \leqslant 100$, we sampled the whole population. Given the large size of the considered samples, the obtained results are robust estimates of real performance.
  2. For each such sample S, train and test the ML model for each reduction using a fivefold cross-validation scheme. We note that training was performed using the reduced multi-rater labels whereas testing was performed using the correct labels;
  3. Compute the mean accuracy and its 95% confidence interval over the set of samples of n raters. Compute the mean Krippendorff α value across all generated samples of size n, to evaluate the average inter-rater agreement.

5. Results and discussion

The average accuracy and confidence intervals obtained for the different datasets and reductions are reported in Tables 2–7. The average α_K values, for the different datasets and numbers of raters, are reported in Table 9.

As a first observation, we can see the effects of the number of raters on the performance of the ML algorithms. As can be noted from the Tables, in all datasets and for all reductions, there is both an increase in average accuracy and a decrease in the width of confidence intervals when the number of raters increases. In all but two cases (the majority reduction on the CTC and SG-0.5 datasets), the difference between the performance of the algorithm trained on data provided by only one or three raters and the performance of the algorithm trained on data provided by the maximum number of raters was statistically significant. This can also be observed by looking at the confidence intervals that do not overlap in Figs. 1–6. These observations provide an answer to the first of our research questions: employing a low number of raters can severely undermine the ground truthing process, and consequently the performance and trustworthiness of ML models, in that low-quality GT labels have a detrimental effect on both the accuracy and the generalization capacity, here measured in terms of variability, of the models.

These observations can also be supplemented with a theoretical analysis based on the results reported in Section 4.2. Let us consider a dataset of size $|S| = 771$, which is the average of the number of instances in the considered datasets, and assume a value of $\delta = 0.05$: that is, we require a 95% probability that the labeling resulting from the majority reduction has no errors. Then, the relationship between the average error rate of the raters and the required number of raters needed to obtain perfect labeling with probability $\geqslant 1 - \delta$ is as shown in Fig. 7. The inflection point in the graph represents the fact that Eq. (31) provides only an upper bound on the number of sufficient raters and not an exact estimate, which is however available when the average rater accuracy is 100%.

The performance bounds for the various reductions, described by Theorems 2 and 4, have an inverse exponential dependency on the number of raters. This relationship provides a clear explanation for the previous observations. We can interpret the chart in Fig. 7 as a critique of the common practice in ML studies of considering the involvement of only one or three raters sufficient, or even adequate. In fact, to have any guarantee that the label quality is adequate with so few raters, we

[1] https://scikit-learn.org/stable/

Table 2
Mean Accuracy and 95% Confidence Intervals Obtained for the CTC Datasets.

Raters	maj	over	tw	fuzzy
1	0.80 ± 0.07	0.80 ± 0.07	0.80 ± 0.07	0.80 ± 0.07
3	0.86 ± 0.05	0.86 ± 0.04	0.88 ± 0.04	0.86 ± 0.04
5	0.87 ± 0.03	0.88 ± 0.03	0.89 ± 0.02	0.90 ± 0.02
7	0.87 ± 0.01	0.89 ± 0.02	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$
9	0.87 ± 0.01	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.00}$

Table 3
Average Accuracy and 95% confidence intervals obtained for the Breast Cancer Dataset.

Raters	maj	over	tw	fuzzy
1	0.79 ± 0.09	0.79 ± 0.09	0.79 ± 0.09	0.79 ± 0.09
3	0.82 ± 0.05	0.82 ± 0.06	0.82 ± 0.02	0.86 ± 0.02
5	0.89 ± 0.03	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.01
7	0.89 ± 0.01	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$
9	0.89 ± 0.01	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.93 \pm 0.00}$

Table 4
Mean Accuracy and 95% confidence intervals obtained for the Diabetes Dataset.

Raters	maj	over	tw	fuzzy
1	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04
3	0.71 ± 0.03	0.71 ± 0.02	0.73 ± 0.03	0.72 ± 0.03
5	0.71 ± 0.03	0.73 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
7	0.71 ± 0.03	$\mathbf{0.77 \pm 0.01}$	0.75 ± 0.01	0.73 ± 0.01
9	0.71 ± 0.03	$\mathbf{0.79 \pm 0.01}$	0.77 ± 0.01	0.78 ± 0.00

Table 5
Average accuracy and 95% confidence intervals obtained for the Synthetic Gaussians Dataset with $n_c = 0.1$.

Raters	maj	over	tw	fuzzy
1	0.82 ± 0.05	0.82 ± 0.05	0.82 ± 0.05	0.82 ± 0.05
3	0.84 ± 0.04	0.84 ± 0.03	0.87 ± 0.03	0.84 ± 0.02
5	0.86 ± 0.03	0.87 ± 0.01	0.87 ± 0.01	0.86 ± 0.02
7	0.86 ± 0.02	0.87 ± 0.01	$\mathbf{0.89 \pm 0.01}$	0.86 ± 0.02
9	0.87 ± 0.01	0.88 ± 0.00	$\mathbf{0.89 \pm 0.01}$	0.88 ± 0.01
11	0.88 ± 0.00	0.88 ± 0.00	$\mathbf{0.89 \pm 0.00}$	0.88 ± 0.01
13	0.88 ± 0.00	0.88 ± 0.00	$\mathbf{0.89 \pm 0.00}$	0.88 ± 0.01
15	0.88 ± 0.00	0.88 ± 0.00	$\mathbf{0.89 \pm 0.00}$	$\mathbf{0.89 \pm 0.00}$
17	0.88 ± 0.00	0.88 ± 0.00	$\mathbf{0.89 \pm 0.00}$	$\mathbf{0.89 \pm 0.00}$
19	0.88 ± 0.00	0.88 ± 0.00	$\mathbf{0.90 \pm 0.00}$	$\mathbf{0.89 \pm 0.00}$

Table 6
Average accuracy and 95% confidence intervals obtained for the Synthetic Gaussians Dataset With $n_c = 0.25$.

Raters	maj	over	tw	fuzzy
1	0.75 ± 0.06	0.75 ± 0.06	0.75 ± 0.06	0.75 ± 0.06
3	0.76 ± 0.06	0.76 ± 0.05	0.80 ± 0.04	0.8 ± 0.06
5	0.76 ± 0.04	0.81 ± 0.03	0.84 ± 0.03	0.85 ± 0.04
7	0.79 ± 0.03	0.82 ± 0.03	0.84 ± 0.02	0.85 ± 0.01
9	0.81 ± 0.03	0.82 ± 0.03	0.84 ± 0.02	0.85 ± 0.01
11	0.82 ± 0.03	0.83 ± 0.03	0.85 ± 0.02	0.86 ± 0.01
13	0.82 ± 0.01	0.85 ± 0.03	0.85 ± 0.02	0.86 ± 0.01
15	0.82 ± 0.01	0.85 ± 0.02	0.85 ± 0.02	0.87 ± 0.01
17	0.82 ± 0.00	0.86 ± 0.01	0.86 ± 0.01	0.87 ± 0.01
19	0.82 ± 0.00	0.86 ± 0.00	$\mathbf{0.87 \pm 0.00}$	$\mathbf{0.87 \pm 0.00}$

would need either a single perfect rater or a group of near perfect raters, with accuracy greater than 90%. It goes without saying that, in general, this is a requirement very difficult to satisfy in the real world.

Another line of reasoning, which reaches the same conclusions, would consider the α_K values. Increasing the number of raters results in a larger α_K value in all datasets, except SG-0.25 and SG-0.5. In general, this correlation confirms that increas-

Table 7
Mean Accuracy and 95% Confidence Intervals Obtained for the Synthetic Gaussians Dataset With $n_c = 0.5$.

Raters	maj	over	tw	fuzzy
1	0.52 ± 0.08	0.52 ± 0.08	0.52 ± 0.08	0.52 ± 0.08
3	0.56 ± 0.07	0.59 ± 0.07	0.59 ± 0.07	0.61 ± 0.07
5	0.58 ± 0.07	0.59 ± 0.07	0.59 ± 0.07	0.61 ± 0.07
7	0.59 ± 0.06	0.59 ± 0.07	0.59 ± 0.07	0.63 ± 0.07
9	0.59 ± 0.06	0.60 ± 0.07	0.60 ± 0.05	0.63 ± 0.05
11	0.60 ± 0.06	0.62 ± 0.05	0.60 ± 0.05	0.65 ± 0.05
13	0.60 ± 0.05	0.62 ± 0.05	0.62 ± 0.03	0.67 ± 0.05
15	0.61 ± 0.05	0.63 ± 0.05	0.63 ± 0.03	$\mathbf{0.69 \pm 0.03}$
17	0.61 ± 0.03	0.65 ± 0.03	0.65 ± 0.01	$\mathbf{0.71 \pm 0.02}$
19	0.61 ± 0.03	0.67 ± 0.01	0.65 ± 0.01	$\mathbf{0.71 \pm 0.01}$

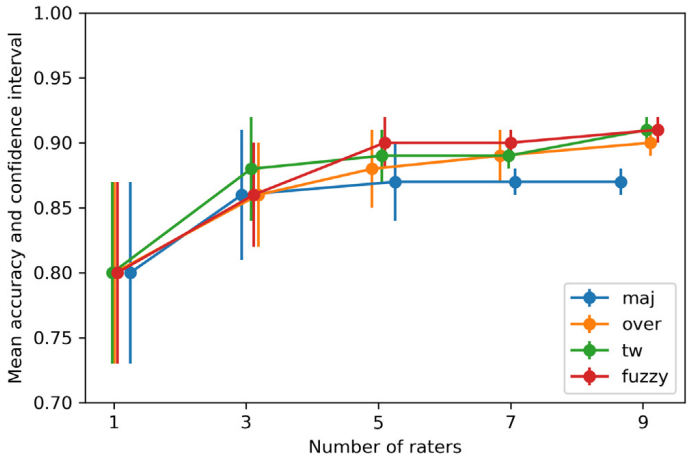


Fig. 1. Average accuracy and 95% confidence intervals for the CTC dataset. Horizontal jitter has been added to limit the overlap of the error bars.

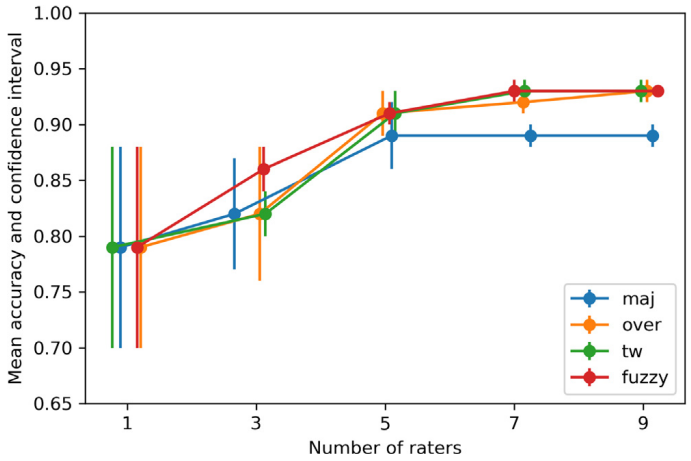


Fig. 2. Average accuracy and 95% confidence interval for the Breast Cancer dataset. Horizontal jitter has been added to limit the overlap of the error bars.

ing the number of raters results in an increase of labeling quality, as demonstrated in Theorems 2 and 4. The exceptions to this finding are datasets SG-0.25 and SG-0.5, where we can identify a number of raters n_T (15 for dataset SG-0.25 and 13 for dataset SG-0.5, see Table 9) beyond which the reliability begins to decrease. This decreasing reliability can be explained by the fact that a large proportion of the raters in these datasets essentially classified completely at random, a behavior heavily penalized by the α_K measure when the number of raters increases. Interestingly, although the α_K values drop when the number of raters for SG-0.25 and SG-0.5 increases, the ML models increase their accuracy and decrease their variability as the number of raters grows; this is especially so for the overwhelming majority, three-way and fuzzy-possibilistic reductions (see Tables 6 and 7). These findings also provide an affirmative answer to our second research question above: whether the usage of non-majority reductions could be helpful, especially on datasets affected by high uncertainty.

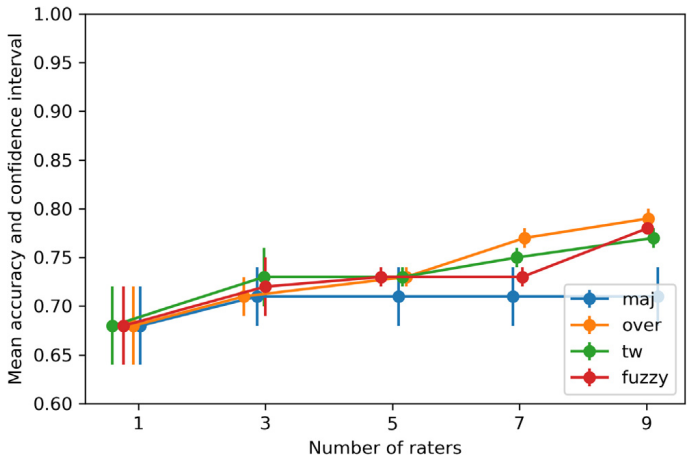


Fig. 3. Average accuracy and 95% confidence intervals for the Diabetes dataset. Horizontal jitter has been added to limit the overlap of the error bars.

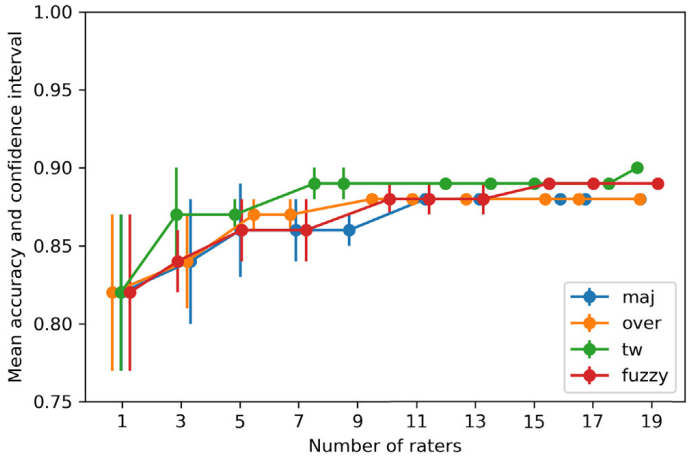


Fig. 4. Average accuracy and 95% confidence intervals for the Synthetic Gaussian with $n_c = 10\%$ dataset. Horizontal jitter has been added to limit the overlap of the error bars.

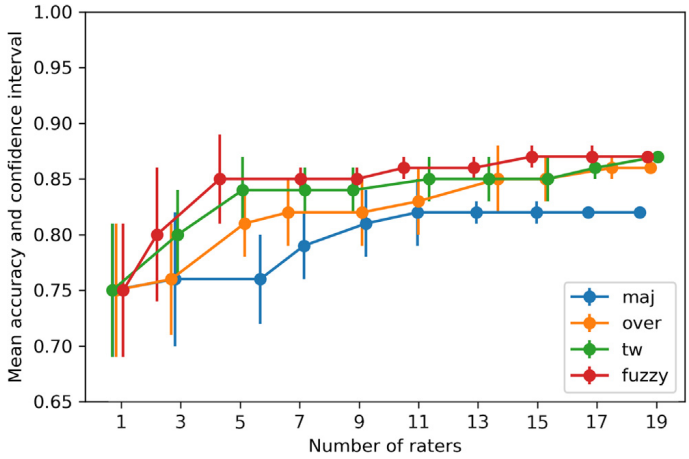


Fig. 5. Average accuracy and 95% confidence intervals for the Synthetic Gaussian with $n_c = 25\%$ dataset. Horizontal jitter has been added to limit the overlap of the error bars.

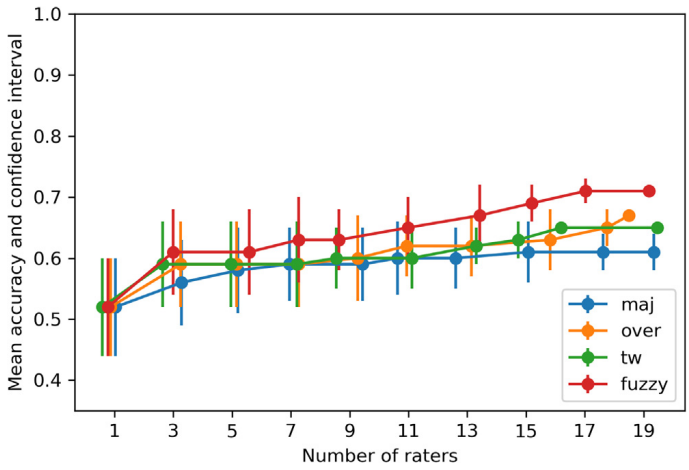


Fig. 6. Average accuracy and 95% confidence intervals for the Synthetic Gaussian with $n_c = 50\%$ dataset. Horizontal jitter has been added to limit the overlap of the error bars.

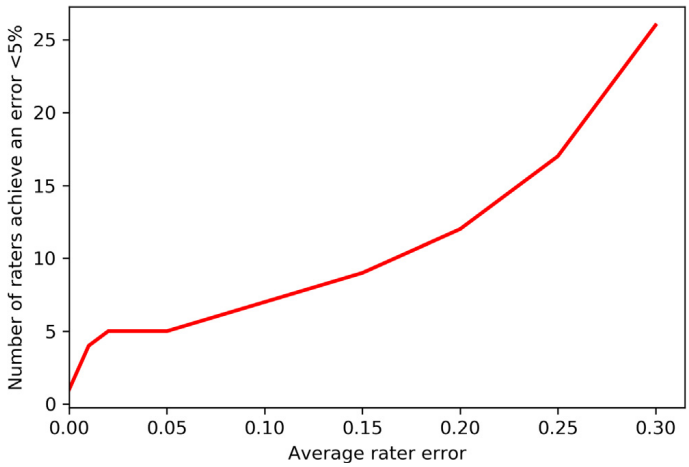


Fig. 7. Bound of the number of raters needed to obtain a labeling error $\delta \leqslant 0.05$ at a fixed average rater error rate on a dataset of size $|S| = 771$.

With reference to Figs. 1–6 and Tables 2–7, we can observe that for all datasets there exists a number of raters beyond which all non-majority reductions under consideration achieve a significantly better performance than the majority reduction. When considering the maximum number of raters (9 in the CTC, Diabetes and Breast Cancer dataset and 19 in the SG datasets), the difference in model accuracy associated with GT generated by the majority reduction versus that of GT associated with the non-majority reductions was always statistically significant.

With regard to the three non-majority reductions, we applied the Friedman test to detect any statistically significant difference among them. The average ranks are reported in Table 8. No statistically significant difference was found at $\alpha = 0.05$. Indeed, the summary statistics is $Q = 1.307$ and the critical value for $n = 6, k = 3, \alpha = 0.05$ is $Q_{crit} \sim 7$.

We can observe that, despite a lack of statistical significance, on average the three-way reduction is associated with better performance than the other reductions, because its Friedman rank (see Table 8) is lower. The fuzzy-possibilistic reduction, by contrast, had better performance on datasets affected by greater uncertainty, as can be observed with the Synthetic Gaussian dataset with $n_c = 50\%$ dataset.

The superior performance of all non-majority reductions compared with the majority reduction could be explained using a model-complexity argument. In fact, the learning algorithms for the non-majority reductions are based on model ensembling over a large number of possible instantiations of the original dataset, a learning paradigm that allows addressing effectively the bias-variance trade-off [36]. This effect is particularly evident with high uncertainty datasets, where the complex patterns concealed by the label noise cannot be reliably recovered using the majority reduction.

Finally, in response to our third research question (studying the influence of raters expertise and the proportion of expert raters on the performance of the ML models), we can observe that the average accuracy of ML models, for all the considered reductions, severely degrades when the proportion n_c of non-expert raters increases (see Tables 5–7 and Figs. 4–6). This degradation is most severe for the majority reduction: indeed when $n_c = 0.1 \rightarrow 0.5$, we get $acc = 0.88 \rightarrow 0.61$. This means that we move from a model accuracy higher than the average rater accuracy (81%), in SG-0.1, to a model accuracy lower

Table 8
Average Ranks of the three non-majority reductions over all datasets.

Reduction	over	tw	fuzzy
Average rank	2.33	1.67	2

Table 9
Krippendorff's Alpha Values For the Different Datasets and Number of Raters.

Raters	CTC	BC	Diabetes	SG 0.1	SG 0.25	SG 0.5
3	0.88	0.77	0.68	0.59	−0.16	−0.24
5	0.91	0.88	0.71	0.70	0.14	−0.21
7	0.94	0.94	0.75	0.76	0.23	−0.07
9	0.96	0.97	0.77	0.84	0.42	−0.05
11	0.96	0.99	0.78	0.86	0.64	0.04
13	–	–	–	0.86	0.70	0.12
15	–	–	–	0.86	0.79	−0.02
17	–	–	–	0.88	0.68	−0.21
19	–	–	–	0.88	0.62	−0.28
21	–	–	–	0.88	0.61	−0.35

than average rater accuracy (75%), in SG-0.5. The non-majority reductions, by contrast, were more robust to increasing, especially the fuzzy-possibilitic reduction: even in the SG-0.5 dataset, this reduction achieves a performance comparable to average performance of the raters.

Therefore, the non-majority reductions can be employed to effectively manage the uncertainty intrinsic to low-reliability contexts, as they exhibit a lesser degradation of performance compared with state-of-the-art approaches.

6. Conclusion

In this paper, we examined how *Ground Truth quality* may impact the performance of predictive models in multi-rater settings. To this end, we considered the importance of the number of the raters involved in ground truthing, the effect of their expertise and the effect of the degree of inter-rater agreement.

Our findings leads us to contest the meaningfulness of the common practice of training ML models on the basis of majority labeling obtained from small sets of annotators, or even single annotators. To study this problem, we introduced and studied the concept of *reduction* – that is, any computational procedure that manages the uncertainty of multiple labels and reduces the noise intrinsic in any multi-rater setting. We also proposed a set of reductions, based on *possibility theory* and *three-way decision*, and studied their theoretical properties. We then applied these reductions in a set of experiments on both real-world and synthetic medical datasets, with the following outcomes:

- We found evidence that the number and expertise of the raters involved in the annotation phase have a critical influence on both the accuracy and generalization capacity of the trained models.
- We provided proof that the proposed reductions better leverage multi-rater annotations and can be used to define a set of more expressive ML models that can better capture the patterns hidden behind the uncertainty and noise resulting from the disagreements and errors of the raters; these models can achieve higher classification accuracy and exhibit higher robustness when the accuracy of the raters decreases.
- Our results show that this increase in both accuracy and robustness can be achieved with fewer raters than those needed by state-of-the-art approaches which, since they were conceived for crowdsourcing settings, typically require large numbers of raters (usually > 50).

In conclusion, we can assert that the proposed methods represent an advancement over and a cost-effective alternative to the existing approaches. These insights are not only important in themselves, they also facilitate further research questions that are beyond the scope of this study. For example:

- As mentioned in Section 4.2.2, the sample complexity bounds obtained for the non-majority reductions were not tight and were based directly on a generalization of the known results for learning in a noisy sample scenario. It would be interesting to provide bounds (e.g., as a generalization of the results of Liu et al. [33]) taking into account the assumptions and peculiarities of each learning setting.
- Similarly, it would be interesting to provide *rater bounds* that are valid for the other reductions – that is, as a parallel to the results of Heinecke et al. [19] for the majority reduction. This can be done, for example, by determining the number of raters sufficient to guarantee that the result of the TWD reduction contains the correct label, with probability close to 1.0.

- In this article, we focused only on binary classification problems, both for our definition of reduction and for theoretical analysis of the associated learning paradigms. Our discussion can then be generalized to the context of multi-class and regression problems.
- While in this article we considered only the traditional case, in which each rater provides a single label, it would be interesting to also consider cases in which the raters are able to express more information – for example, by providing a ranking of the possible labels or expressing their confidence in the label that they propose. Given the similarity of these settings to the problems typically investigated in the field of computational social choice further research should investigate the preference aggregation approaches originally proposed in that context, as applied in multi-rater scenarios with more general, structured labeling representations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Angluin, P. Laird, Learning from noisy examples, Machine Learning 2 (1988) 343–370.
[2] N. Bien, P. Rajpurkar, R.L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B.N. Patel, K.W. Yeom, K. Shpanskaya, et al, Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet, PLoS Medicine 15 (2018) e1002699.
[3] F. Cabitza, A. Campagner, D. Albano, A. Aliprandi, A. Bruno, V. Chianca, A. Corazza, F. Di Pietto, A. Gambino, S. Gitto, et al, The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability, Applied Sciences 10 (2020) 4014.
[4] F. Cabitza, A. Campagner, D. Ciucci, New frontiers in explainable AI: Understanding the GI to interpret the GO, in: A. Holzinger, P. Kieseberg, A.M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2019, pp. 27–47.
[5] F. Cabitza, D. Ciucci, R. Rasoini, A giant with feet of clay: on the validity of the data that feed machine learning in medicine, in: Organizing for the Digital World, Springer, 2019, pp. 121–136.
[6] F. Cabitza, A. Locoro, C. Alderighi, R. Rasoini, D. Compagnone, P. Berjano, The elephant in the record: on the multiplicity of data recording work. Health Informatics Journal, 2019, p. 1460458218824705..
[7] A. Campagner, F. Cabitza, D. Ciucci, The three-way-in and three-way-out framework to treat and exploit ambiguity in data, International Journal of Approximate Reasoning 119 (2020) 292–312.
[8] A. Campagner, F. Cabitza, D. Ciucci, Three-way decision for handling uncertainty in machine learning: a narrative review, in: Proceedings of International Joint Conference on Rough Sets 2020, 2020, pp. 137–152, Springer volume 12179 of LNCS.
[9] R. Caruana, N. Karampatziakis, A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 96–103.
[10] T. Cour, B. Sapp, B. Taskar, Learning from partial labels, Journal of Machine Learning Research 12 (2011) 1501–1536.
[11] I. Couso, C. Borgelt, E. Hullermeier, R. Kruse, Fuzzy sets in data analysis: From statistical foundations to machine learning, IEEE Computational Intelligence Magazine 14 (2019) 31–44.
[12] I. Couso, D. Dubois, Statistical reasoning with set-valued information: Ontic vs. epistemic views, International Journal of Approximate Reasoning 55 (2014) 1502–1518.
[13] T. Denøeux, Maximum likelihood estimation from fuzzy data using the em algorithm, Fuzzy Sets and Systems 183 (2011) 72–91.
[14] D. Dubois, H. Prade, S. Sandri, On possibility/probability transformations, in: R. Lowen, M. Roubens (Eds.), Fuzzy Logic: State of the Art, Springer, Netherlands, Dordrecht, 1993, pp. 103–112.
[15] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115.
[16] D.S. Gierada, T.K. Pilgram, M. Ford, R.M. Fagerstrom, T.R. Church, H. Nath, K. Garg, D.C. Strollo, Lung cancer: Interobserver agreement on interpretation of pulmonary findings at low-dose ct screening, Radiology 246 (2008) 265–272.
[17] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A.B.H. Hassen, L. Thomas, A. Enk, et al, Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Annals of Oncology 29 (2018) 1836–1842.
[18] S.S. Han, G.H. Park, W. Lim, M.S. Kim, J. Im Na, I. Park, S.E. Chang, Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network, PloS one 13 (2018) e0191493.
[19] S. Heinecke, L. Reyzin, Crowdsourced PAC learning under classification noise, in: Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, 2019, pp. 41–49, AAAI volume 7.
[20] F. Hentschel, A.F. Jansen, M. Günther, R. Pauli, S. Lüth, Eosinophil counts in mucosal biopsies of the ileum and colon: Interobserver variance affects diagnostic accuracy, Pathology Research International, 2018, 2018..
[21] R. Hertwig, Tapping into the wisdom of the crowd–with confidence, Science 336 (2012) 303–304.
[22] B.Q. Hu, H. Wong, K.F.C. Yiu, The aggregation of multiple three-way decision spaces, Knowledge-Based Systems 98 (2016) 241–249.
[23] C. Huang, J. Li, C. Mei, W.-Z. Wu, Three-way concept learning based on cognitive operators: an information fusion viewpoint, International Journal of Approximate Reasoning 83 (2017) 218–242.
[24] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization, International Journal of Approximate Reasoning 55 (2014) 1519–1534.
[25] E. Hüllermeier, W. Cheng, Superset learning based on generalized loss minimization, in: A. Appice, P.P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, C. Soares (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2015, pp. 260–275.
[26] R. Jin, Z. Ghahramani, Learning with multiple labels, in: Advances in Neural Information Processing Systems, 2003, pp. 921–928..
[27] K. Krippendorff, Content analysis: An introduction to its methodology, Sage Publications, 2018.
[28] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, International Conference on Machine Learning, vol. 3, 2013..

[29] H. Li, L. Zhang, X. Zhou, et al, Cost-sensitive sequential three-way decision modeling using a deep neural network, International Journal of Approximate Reasoning 85 (2017) 68–78.
[30] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, L.-J. Li, Learning from noisy labels with distillation, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 1910–1918.
[31] Y. Li, L. Zhang, Y. Xu, Y. Yao, R.Y.K. Lau, Y. Wu, Enhancing binary classification by modeling uncertain boundary in three-way decisions, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 1438–1451.
[32] Z. Li, B. Ko, H.-J. Choi, Naive semi-supervised deep learning using pseudo-label, Peer-to-Peer Networking and Applications 12 (2019) 1358–1368.
[33] L.-P. Liu, T.G. Dietterich, Learnability of the superset label learning problem, in: Proceedings of ICML-2014 – Volume 32 ICML'14, 2014, pp. II–1629–II–1637, JMLR.org..
[34] D. Miao, C. Gao, N. Zhang, Three-way decisions-based semi-supervised learning, Theory and Applications of Three-way Decisions, 2012, pp. 17–33..
[35] F. Min, F.-L. Liu, L.-Y. Wen, et al, Tri-partition cost-sensitive active learning through knn, Soft Computing 23 (2019) 1557–1572.
[36] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, second ed., The MIT Press, 2018.
[37] S. Parsons, Qualitative Methods for Reasoning under Uncertainty, MIT Press, 2001.
[38] D. Prelec, H.S. Seung, J. McCoy, A solution to the single-question crowd wisdom problem, Nature 541 (2017) 532–535.
[39] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, Journal of Machine Learning Research 11 (2010) 1297–1322.
[40] B. Sang, Y. Guo, D. Shi, W. Xu, Decision-theoretic rough set model of multi-source decision systems, International Journal of Machine Learning and Cybernetics 9 (2018) 1941–1954.
[41] J. Shen, C.J. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S.Y. Wong, P.-H. Fang, W.-K. Ming, Artificial intelligence versus clinicians in disease diagnosis: Systematic review, JMIR Medical Informatics 7 (2019) e10010.
[42] Svensson C.-M., Hübler R., Figge M.T., Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance, Journal of Immunology Research 2015 (2015) Article ID:573165, 9.
[43] C.-M. Svensson, S. Krusekopf, J. Lücke, M.T. Figge, Automated detection of circulating tumor cells with naive bayesian classifiers, Cytometry Part A 85 (2014) 501–511.
[44] C.-M. Svensson, O. Shvydkiv, S. Dietrich, L. Mahler, T. Weber, M. Choudhary, M. Tovar, M.T. Figge, M. Roth, Coding of experimental conditions in microfluidic droplet assays using colored beads and machine learning supported image analysis, Small 15 (2019) 1802384.
[45] J. Whitehill, T. Wu, J. Bergsma, J.R. Movellan, P.L. Ruvolo, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in: Advances in Neural Information Processing Systems 22, Curran Associates Inc., 2009, pp. 2035–2043.
[46] Y. Yan, Z. Xu, I.W. Tsang, G. Long, Y. Yang, Robust semi-supervised learning through label aggregation, Thirtieth AAAI Conference on Artificial (2016), Intelligence.
[47] Y. Yao, Three-way decisions with probabilistic rough sets, Information Sciences 180 (2010) 341–353.
[48] Y. Yao, An outline of a theory of three-way decisions, in: J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra, L. Polkowski (Eds.), Rough Sets and Current Trends in Computing, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 1–17.
[49] L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) 3–28.
[50] Y. Zhang, D. Miao, J. Wang, Z. Zhang, A cost-sensitive three-way combination technique for ensemble learning in sentiment classification, International Journal of Approximate Reasoning 105 (2019) 85–97.

Seveso *et al. BMC Medical Informatics and Decision Making* 2020, **20**(Suppl 5):142
https://doi.org/10.1186/s12911-020-01152-8

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings

Andrea Seveso[1]*, Andrea Campagner[2], Davide Ciucci[1] and Federico Cabitza[1]

From The 16th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2019) Special Session on Machine Learning in Healthcare Informatics and Medical Biology
Bergamo, Italy. 04-06 September 2019

Abstract

Background: Despite the vagueness and uncertainty that is intrinsic in any medical act, interpretation and decision (including acts of data reporting and representation of relevant medical conditions), still little research has focused on how to explicitly take this uncertainty into account. In this paper, we focus on the representation of a general and wide-spread medical terminology, which is grounded on a traditional and well-established convention, to represent severity of health conditions (for instance, pain, visible signs), ranging from *Absent* to *Extreme*. Specifically, we will study how both potential patients and doctors perceive the different levels of the terminology in both quantitative and qualitative terms, and if the embedded user knowledge could improve the representation of ordinal values in the construction of machine learning models.

Methods: To this aim, we conducted a questionnaire-based research study involving a relatively large sample of 1,152 potential patients and 31 clinicians to represent numerically the perceived meaning of standard and widely-applied labels to describe health conditions. Using these collected values, we then present and discuss different possible fuzzy-set based representations that address the vagueness of medical interpretation by taking into account the perceptions of domain experts. We also apply the findings of this user study to evaluate the impact of different encodings on the predictive performance of common machine learning models in regard to a real-world medical prognostic task.

Results: We found significant differences in the perception of pain levels between the two user groups. We also show that the proposed encodings can improve the performances of specific classes of models, and discuss when this is the case.

Conclusions: In perspective, our hope is that the proposed techniques for ordinal scale representation and ordinal encoding may be useful to the research community, and also that our methodology will be applied to other widely used ordinal scales for improving validity of datasets and bettering the results of machine learning tasks.

Keywords: Ordinal scales, Machine learning, Fuzzy sets, Ground truth

*Correspondence: andrea.seveso@unimib.it
[1]Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The machine learning community seems to put particular emphasis on performance metrics and skill improvement. And rightly so, if this general attitude has pushed some models to perform equally or even better than humans in many tasks, especially with respect to pattern recognition [1, 2].

Much smaller attention and reflection has been paid so far in regard to the validity of data, both input (training) data and output data, that is, the predictions. With validity we do not mean just accuracy, as widely intended, but above all the extent to which a measurement is well-founded and corresponds to the real world phenomena that are to be rendered in symbolic terms [3]. In other terms, we intend the *validity of a data set* as the degree to which the data set represents the phenomena it is intended to.

In order to deal with the intrinsic uncertainty of the medical domain [4], a natural choice has always been to make use of fuzzy logic and fuzzy sets. Several surveys on this connection can be found in literature, for instance [5–8]. The main use of fuzzy logic in this context is to model rules in expert systems (for example [9]) or, often in combination with other approaches such as neural networks, for image processing . On the other hand, only a few attempts to deal with the vagueness of medical terms have been made. We recall here the pioneering work to represent medical terms [10], the fuzzy version of the Arden markup language [11] and several fuzzy ontology applications to medicine [12, 13]. More related to our work is the paper [14], as discussed later in this introduction. Further, even less efforts are available on how uncertainty influences the validity of medical datasets. The recent work by Zywica [15] goes in this direction, by using fuzzy sets for transforming heterogenous data in homogenous ones and to deal with the lack of knowledge.

In this light, we set out to investigate how a specific kind of ordinal features (that is, features whose values come from a categorical label set on which an order relation is defined. In what follows, we consider these ordered categories ordinal data *natively*.) can be transformed in order to improve the *internal validity* of the training set (in the sense above), as well as the validity of the model output (that is, accuracy).

In this article we will specifically address the problem of the representation of ordinal scales in quantitative terms (and vice-versa), and the usage of these representations to define user-informed encoding to be employed in machine learning tasks, by considering the specific case of a very common terminology to represent severity of health conditions and symptoms in medical documents, which has been recently adopted also by the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) [16] framework, that is, the most widely adopted standards framework for the representation of health data on the Internet and in digital health applications [17].

This terminology is used in many questionnaires (for instance, the EQ-5D-5L [18]) aimed at collecting *Patient Reported Outcome Measures* (PROMS), which are recognized [19] as a powerful tool to enable the monitoring of the actual safety and effectiveness of medical procedures and treatments, their continuous improvement, and what is called a *value-based health care* [20, 21].

According to this terminology, both patients and doctors are called to express the severity of health conditions and symptoms in medical documents in terms of five ordinal categories, namely: *Absent* (or *No Condition*), *Mild*, *Moderate*, *Severe* and *Very Severe* (or *Extreme*) conditions. Ordinal scales are very common in medicine [22, 23] and on their basis doctors can understand each other and make critical decisions despite their seeming arbitrariness and loosely defined semantics; ordinal values like those mentioned above are also extensively used to annotate medical records, and to some extent report a written interpretation of other medical data, like laboratory results and medical images. For this reason severity labels are increasingly used in *ground truthing*, that is the preparation of training and test data sets for the definition and evaluation of predictive models. This justifies our interest in investigating whether some knowledge on how these levels are interpreted by the actors involved can affect the performance of predictive models and decision making. Although these categories are used extensively and on a daily basis by most medical doctors around the world in most forms, charts and reports (even paper-based ones), their meaning has never been established univocally and, more importantly from the computational point of view, quantitatively [24]. As a matter of fact, no standardizing body nor single doctor can establish what, say, *Moderate* really means in objective terms [25], nor determine that the transition from a *Mild* condition to a *Moderate* one is like passing from a *Moderate* one to a *Severe* condition: a standard terminology to describe severity is just a set of available values, in which only a total order relation is defined. Of course all these terms are subject to personal views, contextual situations or interpretation of evidence: in a word, they are intrinsically *fuzzy*.

More specifically, the scope of the present work is twofold:

1. Firstly, to represent severity categories using fuzzy sets by means of a collective intelligence process: by collecting the different perceptions provided by interested users, both domain experts (that is, medical doctors) and potential patients;
2. Secondly, to assess the potential impact of these techniques to construct encoding techniques for ordinal data, based on the collective knowledge, to be fed to machine learning models.

As regards the first research question, we will consider these categories as so-called linguistic labels [26] and assign them different types of fuzzy sets with domain on numerical scales according to a human-centered study. In doing so, we can get both a representative, yet approximate, model to map ordinal categories to numerical values (on a scale $[0 - 100]$, where the lower bound represents absence of perceivable signs of the condition of medical interest and the upper bound its strongest expression), and *vice versa*. Also the work [14] deals with grades of questionnaire answers, however, in a different way and with a different scope with respect to us. Indeed, the aim of the authors in [14] is to define a formal logic that enables to describe the derivation of a "total" scores (typically, the average) from a set of degrees (the answers to a questionnaire). Thus, they do not address the problem of defining the total score, but, given the definition of a total score, how to describe it in a formal logic.

The data set we used to define this mapping is a collection of intervals or numerical values for each category/label, provided by both domain experts (that is, medical doctors) and potential patients by means of an ad-hoc Web-based questionnaire, administered during an online survey. We present and discuss several ways to aggregate these values in order to obtain some kind of *fuzzification* of the severity conditions.

This approach is different from existing approaches to fuzzify ordinal scales such as [27, 28], where the fuzzification process is done automatically by assigning a fuzzy number to each label and then applied to a case study. Here, our aim is to fuzzify the ordinal scale starting from the collected data and we will particularly be interested in ascertaining if the representations provided by the different respondent groups (that is doctors and potential patients) present significant differences.

As regards the second research question, the traditional approaches, adopted in the machine learning community, to deal with ordinal data in a training set [29] regard either transforming them into categorical, usually binary, values (such as one-hot encoding or rank-hot encoding), or into the rank index of the corresponding level, that is a number usually ranging from 0 to k.

As already introduced, we explore an alternative approach, that is encoding ordinal values in terms of scalar values on a continuous 100-point scale, according to the fuzzy set representation constructed from the subjective perceptions of the corresponding level on that scale. In doing so, we aim to embed some "true" structure into the dataset, in cases where the assumption that ordinal values are equally-distributed numbers (as in the rank index) does not hold, is ill-grounded or excessively weak.

Methods

Data collection

In order to build the different representations, we collected user data in three different settings, which will be discussed in this section.

First data collection: quantitative meaning for doctors

To collect data on the subjective perception of the quantitative meaning of the categories (each denoted by a specific label) of the severity Health Level 7 (HL7) ordinal scale, we first designed a closed-ended two-page questionnaire to be administered online in a Computer-Assisted-Web-self-Interview (CAWI) configuration. The first page of this questionnaire (depicted in Fig. 1) asked the respondents to express each level of severity of the original 5-item HL7 scale (that is, *Absent*, *Mild*, *Moderate*, *Severe*, and *Very Severe*) into a Visual Analogue Scale (VAS). A VAS is a measurement instrument that has been devised and introduced in health care to try to measure characteristics that appear or are easily perceived as continuous but that cannot be directly measured easily, like pain, and by which to overcome the intrinsically discrete nature of ordinal categorizations [30].

To this aim, we associated each item with a 2-cursor range slider control. By moving each of the two independent cursors the respondents could thus create an *inner interval*, comprised *within* the two cursors, encompassing all those numerical values that they felt could represent the ordinal category properly. The interface was designed so that initially the respondents would want to move the cursors to set the new intervals and, in doing so, "see" the overlap that they deem useful to report between the categories. This overlap was neither promoted nor prevented, as the cursors could be moved freely along each range slider with the only constraint that the 'lower' extreme cursor could never be moved to the right of the 'higher' extreme cursor, and vice versa. Moreover, the respondents could get only an approximate idea of the numerical values that were associated with the position of the cursors (and in fact this association was not mentioned in the task description, reported at the top of Fig. 1, but only in the help section), since the range was intended to be on a strict analogue scale, with no explicit nor numerical anchor. That notwithstanding, VASs are common representational tools most potential respondents were very familiar with for its wide adoption in clinical practice, as said above, and this suggests that respondents performed the task effortlessly. We also explicitly asked for a single number that the respondents could perceive as the most representative for each level: we call this number *Representative Point* (of each level, RP). The second page of the questionnaire was intended to collect a few data on the respondent's professional profile (which was intended to be anonymous), namely their medical specialty.

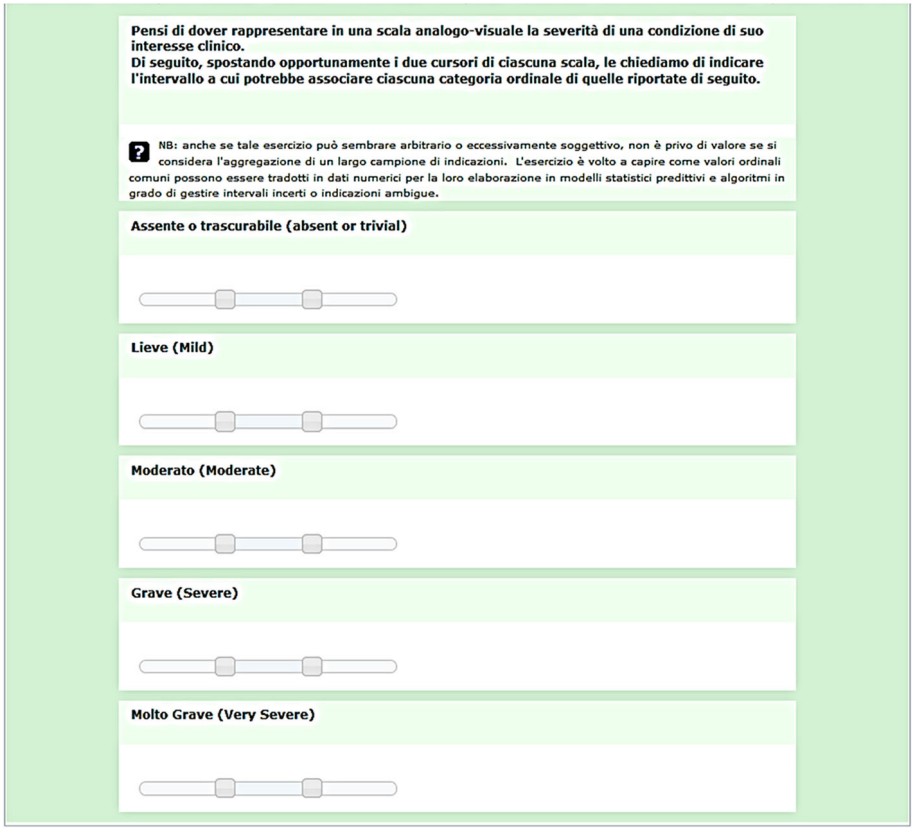


Fig. 1 The first page of the on-line questionnaire that we administered to the sample of clinicians to collect their perception on severity categories (original text in Italian). The translation of the question asked is as follows: *"Think of having to represent the severity of a condition of clinical interest on an analogue-visual scale. Below, by appropriately moving the two cursors of each scale, we ask you to indicate the range to which each ordinal category of the following could associate"*

At the end of November 2017, we invited 97 clinicians by email to fill in the two-page questionnaire. Most respondents worked as clinicians and surgeons at the Scientific Institute for Research, Hospitalization and Healthcare (IRCCS) Orthopedic Institute Galeazzi (IOG), which is one of the largest teaching hospitals in Italy specialized in the study and treatment of musculoskeletal disorders; at IOG almost 5,000 surgeries are performed yearly, mostly arthroplasty (hip and knee prosthetic surgery) and spine-related procedures. After two weeks since this first invitation we sent a gentle reminder and one week later we definitely closed the survey. Response rate was moderately high, especially in light of the very busy daily schedule of the involved prospective respondents, the anonymity of the survey and the lack of incentives: indeed slightly less than half of the potential respondents accepted the invitation and filled in the on-line questionnaire: thus we collected 42 questionnaires by as many respondents (Fig. 2). When we analyzed the responses, some questionnaires were found filled in with seemingly random data and were discarded: then the final dataset contained 298 data points, corresponding to 149 intervals (Fig. 3) by 31 different respondents. Moreover, the questionnaires completed in each and every item were 27. In doing so, we obtained an *Interval Extreme Distribution* (IED) for each severity item. The original doctor data set contained the lower and upper extremes of the five ordinal categories expressing increasing levels of severity for all of the survey respondents, that is a 31 x 10 matrix of data points on the severity dimension, ranging from 0 to 100. From this data set of coordinates of interval extremes we computed a new one, by computing the central points for each IED. An extract of this dataset is reported in Table 1. Both from Fig. 2 and Table 1, it can be seen that in the majority of cases, each level is represented as an interval, not just a coordinate point, and these intervals can overlap. Also, significant differences can exist between different doctors.

Second data collection: quantitative meaning for potential patients

In addition to the doctors of IRCCS Orthopedic Institute Galeazzi ($N_{doctors}$=31), the *doctor* sample of our data, we

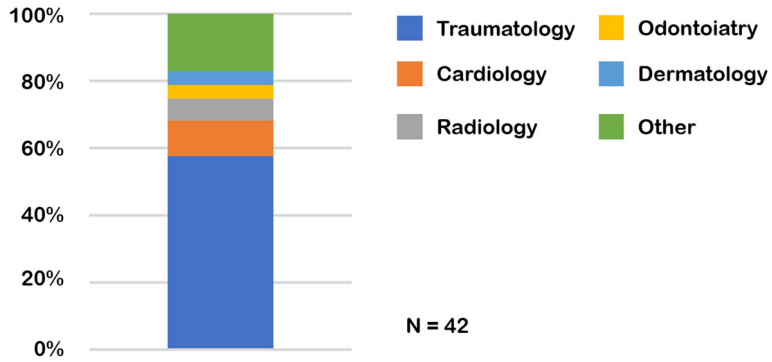


Fig. 2 Stacked bar chart representing the composition of the sample of respondents involved in this study. The majority of the sample were trauma and orthopedic surgeons, the rest of the sample is relatively varied, as also shown by the 'other' category, which is the second one for numerosity and encompasses (among the others) two neurologists, one endocrinologist and one rheumatologist. This suggests that, despite the relatively small sample, this is sufficiently heterogeneous not to consider the responses limited to a specific medical discipline

also involved the students enrolled in a computer science bachelor degree class in the 2018/2019 academic year and asked them to involve other potential respondents among their contacts ($N_{patients}$=1,152); students were given extra credits for participating in the survey and their responses provided the *laypeople* (seen as potential *patients*) sample in this study.

Students were asked to complete a questionnaire similar to the doctor's, as in the previous section. We then computed the Centroids of the IED (CoIED) for each level (that is, IEDs) in both strata. We also calculated the median, as the data appeared to be affected by noise and dirtiness and thus a more robust central tendency indicator would be more useful, RP of each level, for both doctors and patients.

Third data collection: qualitative meaning

Lastly, in order to collect data on the perception of the qualitative meaning of each category, we administered a short questionnaire to the students enrolled in the same computer science class in the following year and their acquaintances. For each questionnaire, a random value is generated a priori in a range from 1 to 99 with equiprobability. The following question is then asked:

"Imagine that you are a patient, and that you are given a scale from 0 to 100, which is often used to represent your health level in numerical form. Imagine that you want to mark on that scale that your health level today is {100 - *random value generated*}. If you had to express in words the same concept, coherently with this numerical value, what expression would you use between the following?"

The respondent is asked to select which category is the most appropriate for his value, from the list of severity categories from HL7. Users are also optionally asked for their sex and age range. We collected 1,257 responses between student and acquaintances. 265 (21%) answers had to be discarded due to an incomplete submission, meaning only 992 (79%) forms were complete and useful for our purposes. For each value in the numerical scale we had an average of 10 complete answers, with a standard deviation of 3.2. For visualization purposes and to enhance the clarity, we performed a binning of the value with granularity of 3, obtaining 33 different bins.

Dataset for regression analysis

In order to perform the regression analysis and test the effects, if any, of the proposed encodings we employed a further dataset. This dataset has been collected from

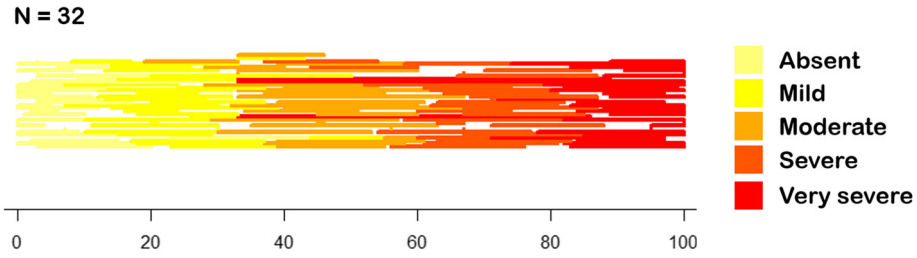


Fig. 3 Diagram showing the data set at a glance. Different questionnaires are represented along the vertical dimension; intervals related to different severity categories are represented in different hues along the horizontal 0-100 continuum

Table 1 An extract of the dataset, for each severity level the min and max values are shown, while the representative point, a scalar value ∈ [0, 100] for each level, is not shown for brevity

Absent	*Mild*	*Moderate*	*Severe*	*Extreme*
3–20	23–40	39–55	56–76	83–100
0–18	18–36	37–58	61–81	82–100
2–15	17–37	39–61	63–83	84–97
23–58	44–78	55–93	60–91	71–97
0–9	10–30	30–53	54–77	78–100
7–7	30–30	56–56	67–67	95–95

real patients who had undergone joint surgery in IRCCS Orthopedic Institute Galeazzi (IOG), one of the major Italian hospitals specialized in musculoskeletal disorders. Specifically, the dataset contains data about 336 patients, with particular reference to so-called Patient Recorded Outcomes (PROMs), that is data reported and collected by the patients (or the doctors) in the last 3 years. In order to measure the effect of the proposed encodings, we considered in particular as a target feature their improvement (on a physical function score) 6 months after joint surgery.

Representation of ordinal values using fuzzy sets

Starting from the collected data, we will define different techniques for representing ordinal scale level using fuzzy sets [31] and to transform the obtained fuzzy set representations into scalar (or vector) features, so to implement encodings of ordinal features.

From ordinal values to fuzzy sets

We will consider a linguistic variable [26] with values in $V = \{v_1, ..., v_k\}$ (in our specific context, the linguistic variable is *Severity condition* and $V = \{$*Absent, Mild, Moderate, Severe, Extreme*$\}$). In this section, we give a semantics to each term in V by means of a fuzzy set in the universe $U = [0, 100]$. The precise fuzzification technique that one can adopt, depends on the type of information specified by the involved respondents; indeed, as described in the "Data collection" section, we asked the respondents two different types of information with respect to the representation of ordinal levels in numeric terms: single numeric values (that is representative points), or whole intervals associated to a given level. In the first case, the fuzzification is straightforward: for each term v in V and each value x in the range $[0 - 100]$ we simply count how many times x has been associated to term v as a representative point. In the second case, two approaches can be adopted:

1 An indicator of central tendency of the single intervals (such as the centroid of the interval or its median) can be employed to convert each interval to a single numeric value. These values can then be employed straightforwardly to compute the fuzzy sets for each of the ordinal levels.
2 The whole interval can be used to construct the fuzzy set representation of the ordinal levels. In this case, given an interval $i = [l_i, u_i]$ reported by a respondent, where l_i (resp. u_i) is the lower (resp. upper) limit of the interval, each point in i is weighted by a factor $w_i = \frac{1}{u_i - l_i + 1}$. Then, for each term v and each value x we count how many times an interval i such that $x \in i$ has been associated to term v, weighed by factor w_i. Compared with the above mentioned technique, this second approach has the advantage that the whole interval information is explicitly considered in building the fuzzy set, however it has been noted in [31] that simply applying this technique on the raw data may result in too noisy distributions, hence binning techniques should be employed to reduce the granularity.

As a concluding note, we observe that, irrespective of the fuzzification technique adopted, the resulting fuzzy sets are not required to be *fuzzy numbers* [32].

From fuzzy representations to encodings

In order to make the fuzzy representations of the ordinal values, obtained by means of one of the techniques previously describer, usable by machine learning algorithms, we need to perform another transformation in order to map the informative but unstructured fuzzy set representation into standard scalar-valued (or vector-valued) features, in a manner which is similar to the traditional *defuzzification* step [33]. To this end, we will describe three different approaches, two of which produce single scalar-valued encodings and one which results in a vector-valued encoding. Let v be an ordinal term and $\mu_v : [0, 100] \mapsto [0, 1]$ the respective fuzzy set encoding. As regards the first approach, that we call *Centroids of the Interval Extreme Distribution* (CoIED) and is akin to the standard *center of gravity* defuzzification method [33], we simply compute the centroid of the membership function μ_v, that is:

$$CoIED(v) = \frac{1}{\sum_{x \in [0,100]} \mu_v(x)} \sum_{x \in [0,100]} x * \mu_v(x) \qquad (1)$$

Notice that this approach produces the same value for each instance of the v label and thus, if the centroids are order-preserving (that is $v_1 \leq v_2 \implies CoIED(v_1) \leq CoIED(v_2)$) this method always preserves the ordinality of the labels.

The second approach that we describe, and that we call *Weighted Sampling*, is based on a sampling method, similar to Monte Carlo approaches [34] and the sampling defuzzification techniques which can be employed for

generalized fuzzy sets [35]. Given the membership function μ_v of an ordinal term v, a probability distribution is computed as $p_v(x) = \frac{\mu_v(x)}{\sum_y \mu_v(y)}$. Then uniformly across the dataset a value x is sampled randomly according to $p_v(x)$ and each occurrence of v is mapped to x. Notice that, contrary to the CoIED method, this method can reverse or otherwise change the ordinality of the labels as it may happen that even if $v_1 \leq v_2$, for a given row, two values x_1, x_2 are sampled (respectively, from p_{v_1} and p_{v_2}) such that $x_2 \leq x_1$.

The third approach, which we call *Membership*, results in a vector-valued encoding and is based on a two-step method. Firstly, given a term v, the numeric value x_v which is most representative of it is selected, that is $x_v = argmax_{x \in [0,100]} \mu_v(x)$. Then x_v is mapped to the vector of its membership values in the different level-specific fuzzy sets, that is:

$$Membership(v) = \langle \mu_{v_1}(x_v), ..., \mu_{v_k}(x_v) \rangle \quad (2)$$

where, respectively, μ_{v_i} is the membership function associated to the ordinal term $v_i \in V$. It is easy to observe that this approach consists of a generalization of one-hot or rank-hot encodings which takes in consideration the inherent vagueness of the underlying ordinal scale: indeed, if the fuzzy sets of the different terms are completely disjoint (that is there does not exists $x \in [0, 100]$ and $v_1, v_2 \in V$ such that $v_1, v_2 \geq 0$) then the result of the membership encoding is equivalent to the above mentioned encodings.

Ordinal data in machine learning

The fuzzy set representations obtained with the quantitative data collection allow us to address two research questions. First: do doctors and potential patients perceive severity levels differently (on an equivalent 100-scale)? On the other hand, the resulting representations were used to address a second research question: does a user-centered encoding improve the validity of machine learning models on some regression tasks?

To this latter aim, we have compared the performance of 4 common machine learning models, namely Random Forests (RF) [36] and Support Vector Regressor (SVR) [37], whose performance is generally recognized as the best one in data-driven predictive tasks [38], and the k-Nearest Neighbour (k-NN) [39] and Least Absolute Shrinkage and Selection Operator (LASSO) [40] ones. These regression models were trained on the same dataset whereas, in one case, ordinal values had been encoded traditionally (that is, 0, 1, 2, 3, 4 respectively), and in the other we had applied the CoIED, Weighted Sampling and Membership encodings.

The regression predictive modeling was based on a set of 15 features (namely gender, age, type of intervention, 3 continuous scores and 9 ordinal features, which were all filled in by patients in pre-operative PROMs questionnaires) to predict the functional improvement 6 months after joint surgery, the models were compared with respect to the *Mean Absolute Error* (MAE) metric and *coefficient of determination* (R2). Comparisons among models were performed on the basis of the confidence intervals on 5-fold nested cross validation. In order to account for the randomness in the Weighted Sampling approach, for that encoding only we repeated the process 10 times and calculated average performances.

Results

In this section we briefly report the results of the statistical procedures conducted in our studies.

Visualization of quantitative meaning: differences between doctor and patient's perception

We performed a Kolmogorov-Smirnov test [41] to compare the shapes of the IEDs of doctors and laypeople (Fig. 4). We decided to employ the Kolmogorov-Smirnov test, in place of other goodness of fit such as the Cucconi test or the Anderson-Darling test, as it provides a conservative test for equality of distributions [42] with good quality implementations in standard statistical packages. We found a statistically significant difference in regard to the *Absent* condition and the two highest severity levels (*Absent*, P<0.001, *Severe*, P=0.038 and *Extreme*, P=0.021), while for the other levels the difference was not found significant, although the p-values are quite low (*Mild*, P=0.067 and *Moderate*, P=0.145).

We performed a Mann-Whitney U test [43] to compare the mean ranks of the patients IEDs (as a sort of hypothetical testing on the equality of their centroids, Table 2) and found significant differences in regard to *Absent*, *Severe* and *Extreme* (P<0.001 in all cases), while differences were not significant for the *Mild* and *Moderate* levels (P=0.425 and 0.105, respectively). We decided to adopt the above test, instead of the Student's t-test, because the main assumptions of this latter did not hold true, and because the Mann-Whitney test is more efficient than the t-test for non-normally distributed data, as well is generally less susceptible to outliers [44].

We also performed a Mann-Whitney U test to compare the mean ranks of the RP distributions and found the same significant differences, in regard to *Absent*, *Severe* and *Extreme* (P<0.001 in all cases), while differences were not significant for *Mild* and *Moderate* (P=0.425 and 0.105, respectively).

Visualization of qualitative meaning

We also investigated the inverse mapping, that is, how respondents mapped precise numerical values to ordinal

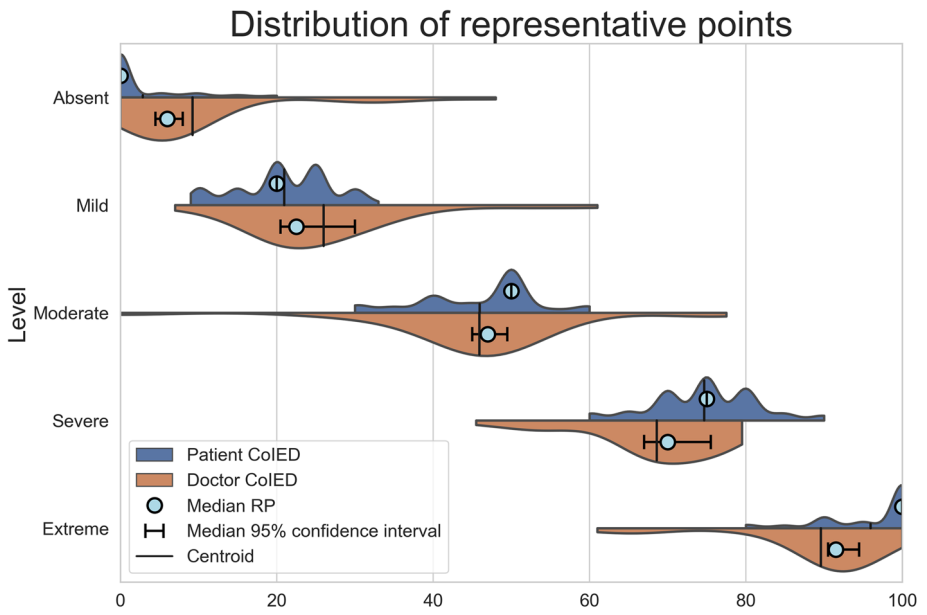


Fig. 4 Violin plots of the IEDs for each severity level (red for doctors, $N_{doctorIED}$=62; blue for patients, $N_{patientIED}$=[1970, 2155, 1971, 1944, 1670], respectively). The CoIED are indicated as a vertical lines in the violin plots. Small circles indicate the median RPs for each level and stratum (doctors $N_{doctors}$=31, patients $N_{patients}$=1,152)

labels from the Health Level 7 (HL7) terminology. A visualization of this mapping in terms of a stacked barchart is shown in Fig. 5.

Another way of visualizing this mapping is shown in Fig. 6. The hue represents the most common variable (red for *Absent*, blue for *Mild*, green for *Moderate*, purple for *Severe*, orange for *Extreme*), while transparency represent the prevalence: very light for superiority (mode), medium for majority (prevalence of the most common class > 50%), opaque for statistical majority (p-value < 0.05). Statistical majority has been calculated by the means of a χ^2 test between the most common class and the second most common.

Table 2 Findings from the user study on the perceptions (expressed in terms of CoIEDs and RPs) by doctors and laypeople of illness severity levels. Significance levels are computed through the Mann-Whitney *U* test

Level	Doctor CoIED 95% CI	Patient CoIED 95% CI	Doctor RP median 95% CI	Patient RP median 95% CI	Diff Doctor vs Patient
Absent	[4.74, 13.7]	[12.9, 14.6]	[4.5, 8.0]	[0.0, 0.0]	***
Mild	[22.2, 29.6]	[25.8, 27.3]	[20.5, 30.0]	[20.0, 20.0]	NS
Moderate	[42.6, 50.7]	[40.4, 42.2]	[45.0, 49.5]	[50.0, 50.0]	NS
Severe	[63.5, 71.5]	[56.83, 59.15]	[67.0, 75.5]	[75.0, 75.0]	***
Extreme	[83.5, 92.5]	[69.12, 72.51]	[90.5, 94.5]	[99.0, 100.0]	***

Results of proposed ordinal representations in machine learning

In Fig. 7 and Tables 3 and 4 we show the results of the comparative regression analysis, after having trained 4 common models on the dataset discussed in the "Methods" section, in order to predict their improvement (on a physical function score) 6 months after joint surgery.

Discussion

This paper addresses the fuzzification of a common terminology, which is also adopted by the Health Level 7 (HL7) framework in the digital health domain, that characterises health conditions, the appearance of medical signs and other expressions of medical relevance. We show how these are perceived by either the medical doctors or the patients themselves (for instance, in the so called Patient Reported Outcome Measures [19]) and the usage of these fuzzy representations to implement knowledge-based encodings to be used by machine learning algorithms.

Perception of HL7 terminology

As regards the perception of these terminologies for the two different respondent groups, as highlighted in the "Results" section, we found a statistically significant difference between the distributions obtained for the respondent groups. In particular, we found that patients tend to overestimate the severity of illness, when this is either serious or absent. We can conjecture that differences in

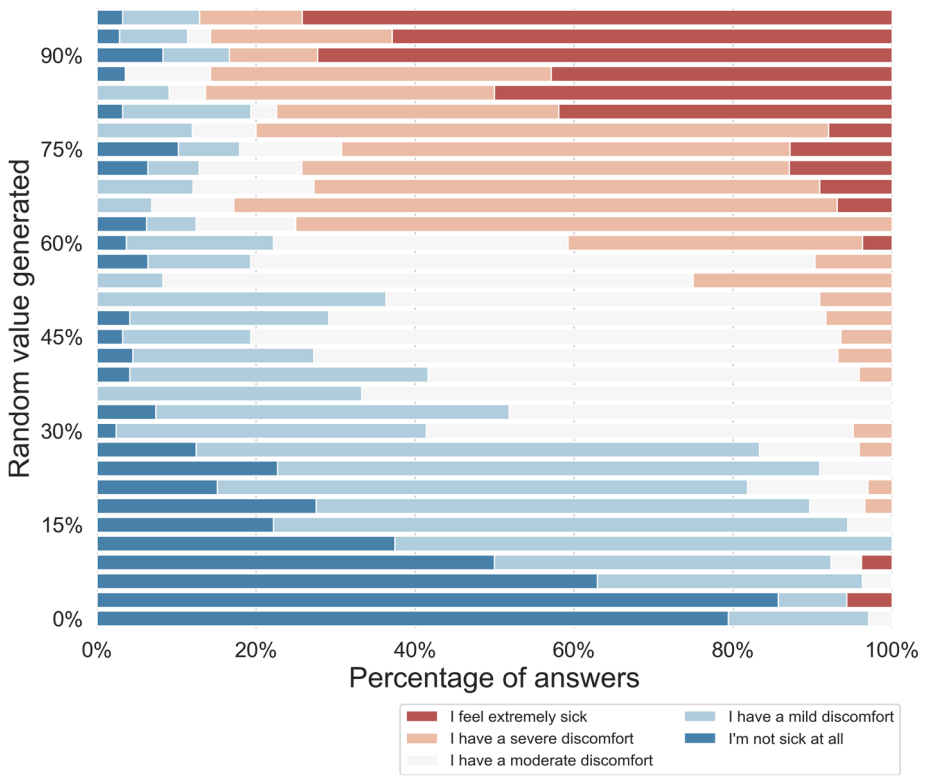


Fig. 5 The stacked bar charts indicate, for each bin, the percentage of respondents for each linguistic label. The original labels were in Italian, as shown in the legend, but they can be directly translated to the already discussed HL7 labels

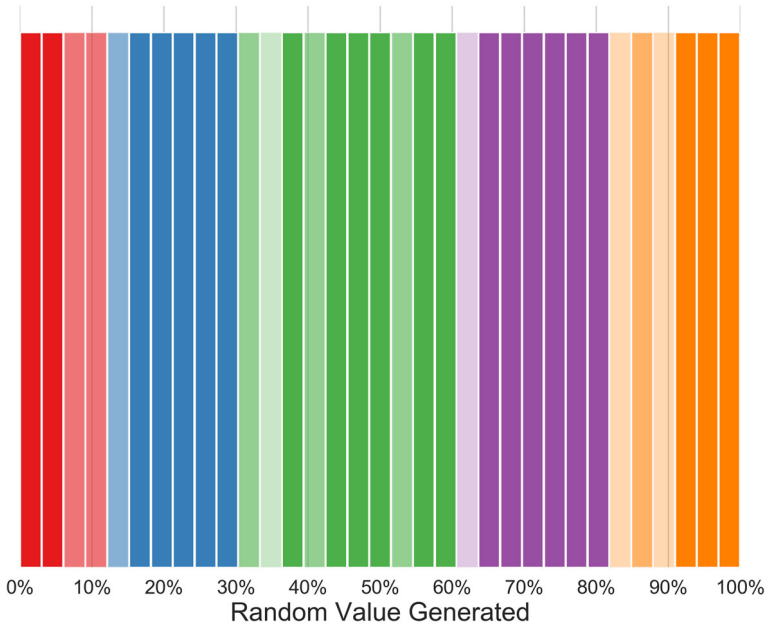


Fig. 6 These bar charts indicate, for each bin, the most common variable chosen by respondents and its prevalence

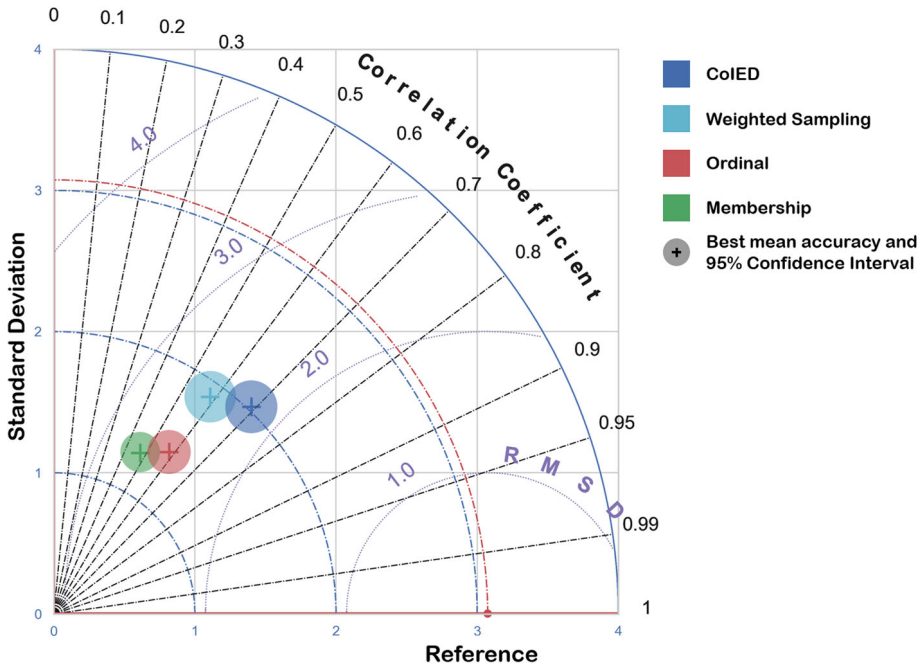


Fig. 7 Taylor diagram of the models' skills. Each point on the diagram indicates the mean performance of all model in a specific representation, with the circle indicating the RMSD 95% confidence interval. RMSD indicates the Root Mean Square Deviation (the lower the better, with 0 denoting a perfect fit to the data). If models denoted with the same glyphs get closer to the inner circle of RMSD and move downward in the diagram (that means that their predictions are more correlated with the true value), then their performance improves and gets better. Red glyphs indicate models with ordinal encodings, while blue glyphs the models with CoIED encodings

the higher part of severity spectrum could be related to the fact that laypeople experience illness in the first person, and hence see it as under a magnifying glass, while doctors have had experience of a much wider range of conditions, relatively few extremely serious and therefore can often scale the assessment lower than patients. By a weaker conjecture, we could see differences in the lower end of the scale as effect of a sort of suppression of the idea to be ill and fear of disease, that induces underestimating light symptoms. These findings, which confirm and are supported by similar findings in the clinical literature [45, 46], have relevant implications, especially as regards their potential impact on machine learning and Artificial Intelligence systems. Indeed, these observations draw attention to the importance of carefully considering the source of data (that is who annotated a specific ordinal value) as the underlying meaning of the same label, even from a standardized terminology as in the case that we considered, could be strongly dependent on who produced the said label. This means that using labels as univocal tokens in advanced statistical techniques, like the ones employed in machine/statistical learning and in the definition of predictive models, can be harmful. The same patient could be associated with a *Mild* label by a doctor, and a *Severe* label by another doctor, and this even if either doctors intend to characterize the very same condition, which could be represented by the same numerical value on a 0-100 continuum. This observation regards the phenomenon of inter-rater reliability that, although widely known in the medical ambit [47], is still little known and considered

Table 3 The regression performance of the 4 machine learning models considered in the comparative study in terms of Mean Absolute Error (MAE) and related confidence intervals (CIs, at a 95% Confidence Level): the lower the value, the better the performance. The first column presents the CIs of the MAE of models with the ordinal encoding; the second column the same accuracy indicators for the CoIED encoding

	Ordinal	CoIED	Membership	Weighted Sampling
RF	[1.458, 1.89]	[1.459, 1.89]	[1.467, 1.861]	[1.688, 1.825]
k-NN	[2.012, 2.277]	**[1.503, 1.813]**	[2.078, 2.321]	**[1.731, 1.854]**
LASSO	[1.586, 1.863]	[1.474, 1.736]	[2.121, 2.367]	[1.769, 1.902]
SVR (RBF kernel)	[1.985, 2.312]	**[1.268, 1.736]**	[2.047, 2.373]	**[1.654, 1.829]**

Table 4 The regression performance of the 4 machine learning models considered in the comparative study in terms of coefficient of determination (R2) and related confidence intervals (CIs, at a 95% Confidence Level): the higher the value, the better the performance. The first column presents the CIs of the R2 of models with the ordinal encoding; the second column the same accuracy indicators for the CoIED encoding

	Ordinal	CoIED	Membership	Weighted Sampling
RF	[0.275, 0.581]	[0.275, 0.58]	[0.291, 0.602]	[0.338, 0.435]
k-NN	[0.043, 0.275]	**[0.333, 0.567]**	[0.006, 0.229]	**[0.339, 0.426]**
LASSO	[0.324, 0.545]	[0.364, 0.637]	[0.0, 0.198]	[0.312, 0.416]
SVR (RBF kernel)	[0.017, 0.296]	[0.265, 0.665]	[0.01, 0.26]	**[0.303, 0.427]**

in most of the fields of applied computer science [3, 48]. For these reasons we argue that any method for properly representing ordinal scales in numerical terms should be grounded on an empirical and human-centered approach, that is, on the subjective perceptions of domain experts for whom the ordinal categories to be fuzzified are meaningful according to the context, right in virtue of their descriptive power and despite their ambiguity. It is noteworthy to say the fuzzification methods proposed and discussed in this paper have been applied to the traditional 5-item severity terminology only as a proof of the concept: we chose this terminology because it is common to many health conditions, used in most medical specialties, and it has also been recently adopted by the HL7 standard developing organization and hence it is nowadays widespread in most digital health applications. However, these fuzzification methods can be applied to *any* ordinal terminology, and not only to those specific of the medical domain.

A potential limitation with respect to this first part of the study, regards the fact that some respondents contacted us after doing the CAWI to warn us that they had found it difficult to move the cursors of the range slider controls on mobile and multi-touch devices like smart phones and tablets. Although we did not collect information on the device used during the CAWI, we can consider that several people could have tried to fill in questionnaire from their smart phones: this could account for some of the "dirtiness" we detected in the original data set (like improbable interval extremes and empty cells). In any case, to our knowledge no study has so far involved more than thirty domain experts to have them represent the quantitative "meaning" (onto a numerical 0-100 range) for the ordinal categories they use in their reports and records on a daily basis.

Machine learning with ordinal encodings

As regards our second research question, that is investigating the effects of the proposed encodings on the performance of the machine learning models, we recall Table 3: as the reader can easily see, the best performing method (in terms of average MAE) is the SVR algorithm with the CoIED encoding. When considering the confidence intervals, the SVR with CoIED encoding is not significantly better than other models on the same representation (RF, LASSO with Ordinal encoding, all algorithms with the CoIED encodings, RF with the Weighted Sampling encoding and SVR with the Membership encoding) however it has both a smaller lower bound and one of the smallest interval widths. In general, all algorithms except RF obtained a better performance using the CoIED encoding and in particular they were statistically significant for both *k*-NN and SVR. This suggest that, at least for specific model classes, the usage of user-informed encodings can significantly improve the predictive performance. Interestingly, the performance of RF using the Ordinal and CoIED encoding were almost exactly equivalent, the explanation for such a behavior resides in the specifics of the RF training algorithm [49]. Indeed, the construction of the Regression Trees embedded in the Random Forests requires the determination of threshold levels on the features and does not take in consideration the metric distance between the values of a feature but only their ordinality: this means that every feature transformation which is order-preserving, such as the CoIED encoding, results in the same exact trees.

As regards the Membership encoding, there were no statistically significant differences with the Ordinal encoding except for the LASSO algorithm, for which the Membership method had worse performance than the traditional Ordinal encoding.

As regards the LASSO algorithm, a possible explanation of the observed behavior is not completely straightforward. A possible explanation may consists in the fact that the Membership encoding replaces a single feature with a group of features which are mutually related, while this relationship is not taken in consideration when train the LASSO model. In this sense, a group LASSO [50] or sparse group LASSO [51] could be an appropriate choice to properly take into consideration the relations and constraints

between the level features introduced by the Membership encoding.

Interestingly, the Weighted Sampling encoding was found to be significantly better than the Ordinal Encoding for *k*-NN and SVR, although generally the CoIED encoding resulted in better average performance. This observation is especially interesting as we did not consider averaging techniques during model training, having just performed multiple samplings for performance evaluation. This suggests that further research should consider the combination of the Weighted Sampling encoding with probabilistic ensembling techniques [52] to assess if these could result in robust and effective methods.

This second part of our study has some limitations, mainly due to its exploratory nature. First, we are aware that performances, as we previously discussed in the case of Random Forests and the CoIED encoding, can vary depending on the match between different encodings, model families, and specific tasks. Even assuming that our encoding is more valid (that is truthful) than the traditional one, for many practical tasks the order information (hence, the Ordinal Encoding) can be as much predictive as the finer-grained one provided by a user-informed one. Although we adopted an approach similar to that applied in [53], we recognize that considering only one task could not be sufficient to draw definitive recommendations. That notwithstanding, we emphasize that we considered a regression task with actual prognostic value that is based on real-world PROMS and clinical data, and that has been integrated in a decision support system currently experimented in a large Orthopedic hospital with promising results.

We are also aware that the observed improvements, while in specific cases statistically significant, are relatively small. That notwithstanding, it is known that significant differences could be associated also to confidence intervals that overlap slightly [54], so our findings must be considered conservative; and most notably all the MAEs observed are lower than the *minimum clinically important difference* values found for the prognostic task at hand [55] (which are at least almost twice as big, if not much bigger).

Furthermore, we are aware that in the specialist literature some methods to encode ordinal variables in numerical terms exist (for instance, *rologit* [56]). For this reason, our future work will be devoted to integrate the knowledge about the user perceptions into these methods to achieve a good compromise between validity and generalization. Also a further validation of the incremental advantage due to the user-informed encoding on different predictive tasks is due.

Conclusion

In this paper we have provided elements to consider fuzzification as a convenient way to convert single ordinal labels, which are the representation of choice of many predictive models, into numbers by the means of a user-informed approach.

The advantage of this approach lies in the fact just mentioned above: the mapping is made on the basis of the perceptions of a heterogeneous sample of domain experts, in our case, clinicians. If perceptions are collected from the experts who annotated a ground truth data set, this mapping could optimally represent the implicit meaning that group of people, as a collective, attach to the annotation labels, and hence to the classes the machine learning have to work with. Even if the perceptions are not collected from the same group of people involved in the observations and the annotations, the opportune selection of the sample (for instance,through stratified random sampling) could guarantee a certain degree of representativeness and bring forth reasonable and meaningful mappings. We also observed that significant differences may exist in the representations provided by different user groups and argued that these should be taken into proper consideration when working with this type of information, as otherwise using naive encodings could be harmful: leading to noisy or wrong predictions or, perhaps even worse, deceitful or ill-founded conclusions.

We then showed how these novel user-based encoding techniques, and more specifically the CoIED encoding, could profitably be used to enhance the performance of standard classes of machine learning models. We also suggested potential areas of improvements and future research with respect the other two proposed encoding techniques.

In conclusion, we believe this paper contributes to the research line that, within the more general field of machine learning in medicine, aims to embed user-derived knowledge into feature engineering tasks (for instance, [31]), especially in regard to the encoding of ordinal features, which are very common in medical data sets, to improve the validity of predictions and of the data considered for medical decision making.

Our future work will be devoted to integrate the knowledge about the user perceptions into other methods to achieve a good compromise between validity and generalization. Also a further validation of the incremental advantage due to the user-informed encoding on different predictive tasks is due.

Abbreviations

CAWI: Computer-Assisted-Web-self-Interview; CoIED: Center of Interval Extreme Distribution; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7; IOG: Orthopedic Institute Galeazzi; IED: Interval Extreme Distribution; IRCCS: Scientific Institute for Research, Hospitalization and Healthcare; *k*-NN: *k*-Nearest Neighbour; LASSO: Least Absolute Shrinkage and Selection Operator; MAE: Mean Absolute Error; PROMS: Patient Reported

Outcome Measures; RF: Random Forests; RP: Representative Point; SVR: Support Vector Regressor; VAS: Visual Analogue Scale

Acknowledgments
The authors would like to thank Pietro de Simoni, a Master student of the Master Degree in Data Science, who has proposed an intuition for Fig. 3. The authors are also grateful to Prof. Giuseppe Banfi for advocating the survey at IOG and to all of the anonymous clinicians and students who spontaneously participated in the research by playing the game of reporting severity categories on a traditional VAS.

About this supplement
This article has been published as part of BMC Medical Informatics and Decision Making, Volume 20 Supplement 5, 2020: Selected articles from the CIBB 2019 Special Session on Machine Learning in Healthcare Informatics and Medical Biology. The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-5.

Authors' contributions
FC and AC provided the PROM data. FC and DC collected the ordinal data. FC designed the research. AS, AC and FC wrote the manuscript; FC and DC supervised and supported the research; DC and AC conceived the theoretical analysis; AS performed experimental analysis. FC and AS substantively revised the manuscript. All author(s) have read and approved the final manuscript.

Funding
The work has been published with the contribution of the Department of Informatics of the University of Milano-Bicocca.

Availability of data and materials
The user questionnaire data generated during the current study are available in the *BMC2020 Public Dataset* repository, located at https://github.com/AndreaSeveso/BMC-2020-Public-Dataset. On the other hand, patient data that support the findings of the machine learning results are property of IRCCS Orthopedic Institute Galeazzi, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the above Institute.

Ethics approval and consent to participate
The dataset used in the machine learning part of the study is anonymous and it was built within a research compliant with all relevant national regulations, institutional policies and in accordance with the tenets of the Helsinki Declaration (as revised in 2013), which was approved by the IRCCS Orthopedic Institute Galeazzi Institutional Review Board or equivalent committee. For the ordinal data collections, which do not contain personal nor medical data, approval by review boards or equivalent committees was not necessary at the time we conducted this study.

Consent for publication
We obtained the written consent to publish their clinical data from the patients in this study.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy. [2] IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi 4, 20161 Milan, Italy.

Received: 29 May 2020 Accepted: 8 June 2020 Published: 20 August 2020

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Cuadros J, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc. 2016;316(22):2402–10.
3. Cabitza F, Campagner A, Ciucci D. New frontiers in explainable AI: Understanding the GI to interpret the GO LNCS, volume 11713. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Cham: Springer; 2019. p. 27–47.
4. Fox RC. Medical uncertainty revisited. Handb Soc Stud Health Med. 2000;409:425.
5. Abbod MF, von Keyserlingk DG, Linkens DA, Mahfouf M. Survey of utilisation of fuzzy technology in medicine and healthcare. Fuzzy Sets Syst. 2001;120(2):331–49.
6. Ahmadi H, Gholamzadeh M, Shahmoradi L, Nilashi M, Rashvand P. Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. Comput Methods Prog Biomed. 2018;161:145–72.
7. Szczepaniak P, Lisboa P, Kacprzyk J, (eds). Fuzzy Systems in Medicine. Heidelberg: Springer; 2000.
8. Barro S, Marín R. Fuzzy Logic in Medicine. Heidelberg: Springer; 2002.
9. Godo L, de Mántaras RL, Puyol-Gruart J, Sierra C. Renoir, pneumon-ia and terap-ia: three medical applications based on fuzzy logic. Artif Intell Med. 2001;21(1-3):153–62.
10. Sanchez E. In: Jones A, Kaufmann A, Zimmermann H-J, editors. Medical Applications with Fuzzy Sets. Dordrecht: Springer; 1986, pp. 331–47.
11. Vetterlein T, Mandl H, Adlassnig K-P. Fuzzy Arden syntax: A fuzzy programming language for medicine. Artif Intell Med. 2010;49(1):1–10.
12. El-Sappagh S, Elmogy M. A fuzzy ontology modeling for case base knowledge in diabetes mellitus domain. Eng Sci Technol Int J. 2017;20(3):1025–40.
13. Lee C-S, Wang M-H, Hsu C-Y, Chen Z-W. Type-2 fuzzy set and fuzzy ontology for diet application. Stud Fuzziness Soft Comput. 2013;301:237–56.
14. Vetterlein T, Zamansky A. Reasoning with graded information: The case of diagnostic rating scales in healthcare. Fuzzy Sets Syst. 2016;298:207–21.
15. Zywica P. Modelling medical uncertainties with use of fuzzy sets and their extensions vol. 855. In: 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Cham: Springer; 2018.
16. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform. 2019;94:103188.
17. Valueset-condition-severity - FHIR V4.0.1. 2019. http://hl7.org/fhir/ValueSet/condition-severity. Accessed 01 Apr 2020.
18. Hernandez G, Garin O, Dima AL, Pont A, Pastor MM, Alonso J, Van Ganse E, Laforest L, de Bruin M, Mayoral K, Serra-Sutton V, Ferrer M. EuroQol (EQ-5D-5L) Validity in Assessing the Quality of Life in Adults With Asthma: Cross-Sectional Study. J Med Internet Res. 2019;21(1):10178.
19. Black N. Patient reported outcome measures could help transform healthcare. Br Med J. 2013;346:167.
20. Baumhauer JF, Bozic KJ. Value-based healthcare: patient-reported outcomes in clinical decision making. Clin Orthop Relat Res. 2016;474(6):1375–8.
21. Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: Issues about uptake and clinical utility. J Am Med Assoc. 2019;321(24):2405–406.
22. Forrest M, Andersen B. Ordinal scale and statistics in medical research. Br Med J Clin Res Ed. 1986;292(6519):537–8.
23. Jakobsson U. Statistical presentation and analysis of ordinal data in nursing research. Scand J Caring Sci. 2004;18(4):437–40.
24. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metrics. 2003;1(1):12.
25. Atkinson TM, Hay JL, Dueck AC, Mitchell SA, Mendoza TR, Rogak LJ, Minasian LM, Basch E. What do 'none,' 'mild,' 'moderate,''severe,' and 'very severe' mean to patients with cancer? Content validity of PRO-CTCAE response scales. J Pain Symptom Manag. 2018;55(3):3–6.
26. Zadeh LA. The concept of a linguistic variable and its application to approximate reasoning I. Inf Sci. 1975;8(3):199–249.
27. Li Q. A novel likert scale based on fuzzy sets theory. Expert Syst Appl. 2013;40(5):1609–18.
28. Vonglao P. Application of fuzzy logic to improve the likert scale to measure latent variables. Kasetsart J Soc Sci. 2017;38(3):337–44.
29. Coates A, Ng AY. The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011. p. 921–8. https://icml.cc/Conferences/2011/papers/485_bibfile.bib.
30. Crichton N. Visual analogue scale (VAS). J Clin Nurs. 2001;10(5):706–6.

31. Cabitza F, Ciucci D. Fuzzification of ordinal classes. The case of the HL7 severity grading LNCS, volume 11142. In: International Conference on Scalable Uncertainty Management. Cham: Springer; 2018. p. 64–77.
32. Dijkman JG, van Haeringen H, de Lange SJ. Fuzzy numbers. J Math Anal Appl. 1983;92(2):301–41.
33. Van Leekwijck W, Kerre EE. Defuzzification: criteria and classification. Fuzzy Sets Syst. 1999;108(2):159–78.
34. Kroese D, Taimre T, Botev Z. Handbook of Monte Carlo Methods. Hoboken: Wiley; 2011.
35. Greenfield S, Chiclana F, John R, Coupland S. The sampling method of defuzzification for type-2 fuzzy sets: Experimental evaluation. Inf Sci. 2012;189:77–92.
36. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
37. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14(3):199–222.
38. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(1):3133–81.
39. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
40. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
41. Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46(253):68–78.
42. Landoni E, Ambrogi F, Mariani L, Miceli R. Parametric and nonparametric two-sample tests for feature screening in class comparison: a simulation study. Epidemiol Biostat Public Health. 2016;13(2):.
43. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18(1):50–60.
44. Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or *t*-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv. 2010;4:1.
45. Boyle CM. Difference between patients' and doctors' interpretation of some common medical terms. Br Med J. 1970;2(5704):286–9.
46. Forrest M. Assessment of pain: a comparison between patients and doctors. Acta Anaesthesiol Scand. 1989;33(3):255–6.
47. Cabitza F, Locoro A, Laderighi C, Rasoini R, Compagnone D, Berjano P. The elephant in the record: on the multiplicity of data recording work. Health Inform J. 2019;25(3):475–90.
48. Cabitza F, Ciucci D, Rasoini R. A giant with feet of clay: on the validity of the data that feed machine learning in medicine. In: Cabitza F, Magni M, Batini C, editors. Organizing for the Digital World. Cham: Springer; 2019. p. 121–36.
49. Hastie T, Tibshirani R, Friedman J. Additive Models, Trees, and Related Methods. New York, NY: Springer; 2009, pp. 295–336.
50. Yuan M, Lin Y. J R Stat Soc Ser B Stat Methodol. 2006;68(1):49–67.
51. Puig AT, Wiesel A, Hero AO. A multidimensional shrinkage-thresholding operator. In: 2009 IEEE/SP 15th Workshop on Statistical Signal Processing. Cardiff: IEEE; 2009. p. 113–116.
52. Murphy KP. Machine Learning: a Probabilistic Perspective. Cambridge, Massachusetts: MIT press; 2012.
53. Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. Int J Comput Appl. 2017;175(4):7–9.
54. Ranstam J. Why the *p*-value culture is bad and confidence intervals a better alternative. Osteoarthr Cartil. 2012;20(8):805–8.
55. Hung M, Bounsanga J, Voss MW, Saltzman CL. Establishing minimum clinically important difference values for the patient-reported outcomes measurement information system physical function, hip disability and osteoarthritis outcome score for joint reconstruction, and knee injury and osteoarthritis outcome score for joint reconstruction in orthopaedics. World J Orthop. 2018;9(3):41.
56. Ophem HV, Stam P, Praag BV. Multichoice logit: modeling incomplete preference rankings of classical concerts. J Bus Econ Stat. 1999;17(1):117–28.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

Everything is Varied: The Surprising Impact of Individual Variation on ML Robustness in Medicine

Andrea Campagner[1], **Lorenzo Famiglini**[1], **Anna Carobene**[3], **Federico Cabitza**[1,2]

Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano-Bicocca, viale Sarca 336 – 20126 Milano, Italy

IRCCS Istituto Ortopedico Galeazzi, Milano, Italy

IRCCS Ospedale San Raffaele, Milano, Italy

Abstract

In medical settings, Individual Variation (IV) refers to variation that is due not to population differences or errors, but rather to within-subject variation, that is the intrinsic and characteristic patterns of variation pertaining to a given instance or the measurement process. While taking into account IV has been deemed critical for proper analysis of medical data, this source of uncertainty and its impact on robustness have so far been neglected in Machine Learning (ML). To fill this gap, we look at how IV affects ML performance and generalization and how its impact can be mitigated. Specifically, we provide a methodological contribution to formalize the problem of IV in the statistical learning framework and, through an experiment based on one of the largest real-world laboratory medicine datasets for the problem of COVID-19 diagnosis, we show that: 1) common state-of-the-art ML models are severely impacted by the presence of IV in data; and 2) advanced learning strategies, based on data augmentation and data imprecisiation, and proper study designs can be effective at improving robustness to IV. Our findings demonstrate the critical relevance of correctly accounting for IV to enable safe deployment of ML in clinical settings.

Introduction

In recent years, the interest toward the application of Machine Learning (ML) methods and systems to the development of decision support systems in clinical settings has been steadily increasing (Benjamens, Dhunnoo, and Meskó 2020). This interest has been mainly driven by the promising results obtained and reported by these systems in academic research for different tasks (Aggarwal et al. 2021; Yin, Ngiam, and Teo 2021)

Despite these promising results, the adoption of ML-based systems in real-world clinical settings has been lagging behind (Wilkinson et al. 2020), with these systems often failing to meet the expectations and requirements needed for safe deployment in clinical settings (Andaur Navarro et al. 2021; Futoma et al. 2020), a concept that has been termed the *last mile of implementation* (Coiera 2019) . While reasons behind the gaps in this "last mile" are numerous, among them we recall the inability of ML systems to reliably generalize in new contexts and settings (Beam, Manrai, and Ghassemi 2020; Christodoulou et al. 2019), as well as their lack of robustness and susceptibility to variation in data, leading to poorer performance in real settings (Li et al. 2019) and, ultimately, to what has been called the *replication crisis* of ML in medicine (Coiera et al. 2018).

In the ML literature, the notion of variation has usually been associated with variance in the population data distribution (that is, as it relates to either the larger reference population, or a sample taken from this latter), due to the presence of outliers or anomalies (Akoglu 2021), out-of-distribution instances (Adila and Kang 2022; Morteza and Li 2022) or concept/co-variate shifts and drifts (Liu et al. 2021; Rabanser, Günnemann, and Lipton 2019). While these forms of variation are certainly relevant, however they are not the only ones that can arise in real-world settings: indeed, another source of variation in data is the so-called *individual variation* (IV) (Fraser 2001), which is especially common in laboratory data or other physiological signals and biomarkers, and more generally in every phenomenon whose manifestations can exhibit time-varying patterns.

IV denotes between-subject variation that is not due to population differences or errors, but rather to the intrinsic and characteristic (that is, individual) patterns of variation pertaining to single instances, that is to within-subject variation (Plebani, Padoan, and Lippi 2015); and more specifically it relates to two possible sources of variation: either the feature values for a given subject or patient, what is called *biological variation* (BV) (Plebani, Padoan, and Lippi 2015); or the measurement process and instrument itself, i.e., what is called *analytical variation* (AV). The presence of IV entails (Badrick 2021) that for each individual one can identify a "subject average" or central tendency (*homeostatic point* (Fraser 2001)) arising from such factors as personal characteristic of the individuals themselves (e.g., genetic characteristics, age, phenotypic elements such as diet and physical activity) or of the measurement instrument (e.g., instrument calibration), as well as a distribution of possible values, whose uncertainty is represented by the extent of the IV: crucially, only a snapshot (i.e., a sample) from this distribution can be accessed at any moment.

While the potential impact of IV on computer-supported diagnosis has been known for a while (for instance, in (Spodick and Bishop 1997) authors reported that "computer interpretations of electrocardiograms recorded 1 minute apart were significantly (grossly) different in 4 of 10

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cases"), only conjectures have so far been produced to estimate its extent. Nonetheless, IV has two strong implications for ML applications. First, ML models trained on data affected by IV, even highly accurate ones, can fail to be robust and properly generalize not only to new patients, but also to the same patients observed in slightly different conditions: for example, an healthy patient could indeed be classified as healthy with respect to the features actually observed for them, while they could have been classified as non-healthy for a slightly different set of feature values, which nevertheless would still be totally compatible with the distribution due to IV[1]. Second, differently from distribution-related variation, collecting additional data samples, which has been considered a primary factor in the continued improvement of ML systems, can help only marginally in reducing the impact of IV, unless specific study designs are adopted that allow to capture multiple observations for each individuals across time (Aarsand et al. 2018; Bartlett et al. 2015).

Despite these apparently relevant characteristics, the phenomenon of IV has largely been overlooked in the ML literature: indeed, while recent works have started to apply ML techniques to analyze IV data, for example to cluster patients based on their IV profiles (Carobene et al. 2021) or to provide Bayesian models for IV (Aarsand et al. 2021), to our knowledge no previous work has investigated the impact of IV on ML systems, as well as possible techniques to improve robustness and manage this source of perturbations.

In this article, we attempt to bridge this gap in the specialized literature, by addressing two main research problems. To this aim, this paper will consist of two parts: in the first part we will address the research question "can individual variation significantly affect the accuracy, and hence the robustness, of a machine model on a diagnostic task grounding on laboratory medicine data" (H_1). Due to the pervasiveness of individual variation, proving this hypothesis could suggest that most ML models could be seriously affected by lack of robustness on real-world and external data. To this aim, we will apply a biologically-grounded, generative model to simulate the effects of IV on data, and we will show how commonly used classes of ML models fail to be robust to it. On the other hand, the second part of the paper will aim to build on the rubble left by the first part, and it will address the hypothesis whether more advanced learning and regularization methods (grounding on, either, data augmentation (Van Dyk and Meng 2001) or data imprecisiation (Lienen and Hüllermeier 2021b)) will achieve increased robustness in face of the same perturbations (H_2).

Background and Methods

As discussed in the previous section, the aim of this article is to evaluate and address the potential impact of IV on ML models' robustness. In this section, we first provide basic background on IV, its importance in clinical settings, and methods to compute it. Then, in the next sections, we will describe two different experiments: in the first experiment, we evaluate how commonly used ML models fare when dealing with data affected by IV; then, in the second experiment, we evaluate the application of more advanced ML approaches to improve robustness to IV.

Individual Variation in Medical Data

IV is considered one of the most important sources of uncertainty in clinical data (Plebani, Padoan, and Lippi 2015) and recent research has highlighted the need to take IV properly into account in any use of medical data (Badrick 2021; Fröhlich et al. 2018) . IV can be understood as encompassing three main components: pre-analytical variation, analytical variation and (within-subject) biological variation (Fraser 2001; Plebani, Padoan, and Lippi 2015).

Pre-analytical variation denotes uncertainty due to patients' preparation (e.g., fasting, physical activity, use of medicaments) or sample management (including, collection, transport, storage and treatment) (Ellervik and Vaught 2015); it is usually understood that pre-analytic variation can be controlled by means of careful laboratory practice (Fraser 2001). AV, by contrast, describes the un-eliminable uncertainty which is inherent to every measurement technique, and is characterized by both a random component (i.e., variance, that is the agreement between consecutive measurements taken with the same instrument); and a systematic component (i.e., bias, that is the differences in values reported by two different measurement instruments). Finally, BV describes the uncertainty arising from the fact that features or biomarkers can change through time, contributing to a variance in outcomes from the same individual that is independent of other forms of variation.

As already mentioned, IV can influence the interpretation and analysis of any clinical data: for this reason, quantifying IV, also in terms of its components, is of critical importance. However collecting reliable data about IV is not an easy task (Carobene et al. 2018; Haeckel, Carobene, and Wosniok 2021). To this aim, standardized methodologies have recently been proposed (Aarsand et al. 2018; Bartlett et al. 2015): intuitively, IV can be estimated (Aarsand et al. 2021; Carobene et al. 2018; Røraas, Petersen, and Sandberg 2012) by means of controlled experimental studies that monitor *reference individuals*[2] (Carobene et al. 2016) by collecting multiple samples over time.

Formally speaking, let us assume that a given feature of interest x has been monitored in n patients for m time steps. At each time step, $k \geq 2$ repeated measurements should be performed, so as to determine the AV component of IV. Then, the IV of feature x, for patient i, is estimated as $IV_i(x) = Variance(x^i)$, while the AV component is defined as $AV_i(x) = Variance(x^i_s)$, where x^i denotes the collection of values of x for patient i, and x^i_s denotes the collection of values of x for patient i at the s-

[1] As we show in the following, this setting is a generalization of the usual one adopted in ML theory (Shalev-Shwartz and Ben-David 2014): not only we assume that the best model could have less than perfect accuracy, but we also assume that any instance is represented as a distribution of vectors possibly lying in opposite sides of the decision boundary.

[2] The term reference individual denotes an individual that, for some reasons, can be considered representative of the population of interest (e.g., healthy patients).

th time step. Then, the BV component of IV is computed as $BV_i(x) = \sqrt{IV_i(x)^2 - AV_i(x)^2}$. Usually, IV, AV and BV are expressed in percent terms, defining the so-called coefficients of individual (resp., analytical, biological) variation, that is $CVT_i(x) = \frac{IV_i(x)}{\hat{x}^i}$, $CVA_i(x) = \frac{AV_i(x)}{\hat{x}^i}$ and $CVI_i(x) = \frac{BV_i(x)}{\hat{x}^i}$. The overall variations, finally, can be computed as the average of the coefficients of variation across the population of patients. The value of CVT, for a given set of features $x = (x_1, ..., x_d)$, can then be used to model the uncertainty about the observations obtained for any given patient i: indeed, any patient i, as a consequence of the uncertainty due to IV, can be represented by a d-dimensional Gaussian $\mathcal{N}_i(x^i, \Sigma^i)$, where x^i is a d-dimensional vector characteristic representation of patient i, called *value at the homeostatic point*, and Σ^i is the diagonal covariance matrix given by $\Sigma^i_{j,j} = CVT(x_j) * x^i_j$ (Fraser 2001). More generally, having observed a realization $\hat{x}^i$ of $\mathcal{N}_i(x^i, \Sigma^i)$ for patient i, its distribution can be estimated as $\mathcal{N}_i(\hat{x}^i, \hat{\Sigma}^i)$, where $\hat{\Sigma}^i_{j,j} = CVT(x_j) * \hat{x}^i_j$.

Due to the complexity of design studies to obtain reliable IV estimates, a few compiled sources of IV data, for healthy patients, are available: the largest existing repositories in this sense, are the data originating from the European Biological Variation Study (EuBIVAS) and the Biological Variation Database (BVD) (Aarsand et al. 2020; Sandberg, Carobene, and Aarsand 2022), both encompassing data about commonly used laboratory biomarkers. In the following sections, we will rely on data available from these sources in the definition of our experiments.

Individual Variation and Statistical Learning

One of the most simple yet remarkable results in Statistical Learning Theory (SLT) is the *error decomposition theorem* (Shalev-Shwartz and Ben-David 2014) (also called bias-variance tradeoff, or bias-complexity tradeoff), which states that the true risk $L_D(h)$ of a function h from a family H w.r.t. to a distribution D on the instance space $Z = X \times Y$ can be decomposed as:

$$L_D(h) = \epsilon^{Bayes} + \epsilon^{Bias} + \epsilon^{Est} \tag{1}$$

where $\epsilon^{Bayes} = min_{f \in F} L_D(f)$ is the *Bayes error*, i.e. the minimum error achievable by any measurable function; $\epsilon^{Bias} = min_{h' \in H} L_D(h') - min_{f \in F} L_D(f)$ is the *bias*, i.e. the gap between the Bayes error and the minimum error achievable in class H; $\epsilon^{Est} = L_D(h) - min_{h' \in H} L_D(h')$ is the *estimation error*, i.e. the gap between the error achieved by h and the minimum error achievable in H.

A striking consequence of IV for ML tasks regards a generalization of the error decomposition theorem due to the impossibility of accessing the true distributional-valued representation of instances but only a sample drawn from the respective distributions. To formalize this notion, as in the previous section, denote with $f_i = \mathcal{N}(x^i, \Sigma^i)$ the distributional representation due to IV for instance i. Then, the learning task can be formalized through the definition of a *random measure* (Herlau, Schmidt, and Mørup 2016) η defined over the Borel σ-algebra $(Z, \mathcal{B})$ on the instance space $Z = X \times Y$, which associates to each instance (x, y) a probability measure $\mathcal{N}(x, \Sigma) \times \delta_y$, where δ_y is the Dirac measure at $y \in Y$. A training set $S = \{(x_1, y_i), \ldots, (x_m, y_m)\}$ is then obtained by first sampling m random measures $f_1, \ldots, f_m$ from η^m, and then, for each i, by sampling a random element $(x_i, y_i) \sim f_i$. Then, the IV-induced generalization of the error decomposition theorem can be formulated as:

$$L_\eta(h) = \epsilon^{Bayes}_\eta + \epsilon^{Bias}_\eta + \epsilon^{Est}_\eta + \epsilon^{IV}_\eta \tag{2}$$

Indeed, the true error of h w.r.t. η can be expressed as $L_\eta(h) = E_{F \sim \eta^m}\left[\frac{1}{m}\sum_{f_i \in F} E_{(x_i,y_i)\sim f_i} l(h, (x_i, y_i))\right]$. Letting D be the probability measure over $X \times Y$ obtained as the *intensity measure* (Kallenberg 2017) of η, and $L_D(h) = E_{S \sim D^m} L_S(h)$ be the expected error of h w.r.t. to the sampling of a training set S from the product measure D^m, then the above expression can be derived by setting $\epsilon^{Bayes}_\eta = min_{f \in F} L_\eta(f)$, $\epsilon^{Bias}_\eta = min_{h' \in H} L_\eta(h') - min_{f \in F} L_\eta(f)$, $\epsilon^{Est}_\eta = L_D(h) - min_{h' \in H} L_\eta(h')$ and $\epsilon^{IV}_\eta = E_{F \sim \eta^m, S \sim D}\left[\frac{1}{m}\sum_i E_{(x_i,y_i)\sim f_i} l(h, (x_i, y_i)) - l(h, (x'_i, y'_i)\right]$.

Thus, compared with Eq (1), Eq (2) includes an additional error term ϵ^{IV} which measures the gap in performance due to the inability to use the IV-induced distributional representation of the instances, bur rather only a single instantiation of such distributions. This aspect is also reflected in the estimation error component in which the reference $min_{h' \in H} L_\eta(h')$ is compared not with the true error $L_\eta(h)$ but rather with the expected error over all possible instantiations $L_D(h)$. In the following sections, we will show, through an experimental study, that the impact of IV can be significant and lead to an overestimation of any ML algorithm's performance and robustness.

Measuring the Impact of Individual Variation on Machine Learning Models

In order to study whether and how the performance of a ML model could be impacted by IV, we designed an experiment through which we evaluated several commonly adopted ML models in the task of COVID-19 diagnosis from routine laboratory blood exams, using a public benchmark dataset. Aside from its practical relevance (Cabitza et al. 2021), we selected this task for three additional reasons. First, blood exams are considered one of the most stable panels of exams (Coskun et al. 2020): this allows us to evaluate the impact of IV in a conservative scenario where the features of interest are affected by relatively low levels of variability. Second, validated data about IV for healthy patients who underwent blood exams are available in the specialized literature (Buoro et al. 2017a,b, 2018) and these exams have high predictive power for the task of COVID-19 diagnosis (Chen et al. 2021b). Third, the selected dataset was associated with a companion longitudinal study (authors a) that has been used to estimate IV data for the COVID-19 positive patients: we believe this to be particularly relevant since, even though IV data are available for healthy patients, no information of this kind is usually available for non-healthy patients, due to the complexity of designing studies for the

collection of IV data, which could exhibit disease-specific patterns. Although the estimation of IV is of paramount importance, both in medicine and other safety-critical domains, the striking lack of datasets presenting information to assess IV makes it a priority to devote further efforts and initiatives to make such resources available to the ML research community to make their models more robust and reliable.

To this purpose, we used a dataset of patients who were admitted at the emergency department of the IRCCS Ospedale San Raffaele and underwent a COVID-19 test (authors b). The dataset was collected between February and May 2020 and encompasses 18 continuous features and 3 binary features (including the target). Since the dataset was affected by missing data, in order to limit the bias due to data imputation, we discarded all instances having more than 25% missing values: the resulting dataset encompasses 1422 instances and is described in Table A1 in Appendix A.

To evaluate the impact of IV, we used a biologically-informed generative model whose aim was to simulate the effect of biological and analytical variation on the measured features of the patients in the dataset. More in detail, based on the definition and computation of IV described previously, the generative model is defined by a case-dependent, class-conditional, multi-variate Gaussian distribution $N(x, \Sigma^{x,y})$, where we recall $\Sigma^{x,y} = diag(\langle x, \sqrt{CVA^2 + CVI_y^2} \rangle)$. We note that, even though the assumptions of normality and independence of variables may be considered strong, they are widely adopted in the specialized IV literature (Fraser 2001) as well as implicitly in the release format of the available IV data sources. Nonetheless, we believe that further work should be devoted at exploring more general models of IV that may take into account dependencies among features.

More in particular, for CVA and $CVI_{y=0}$ we considered values previously reported in the literature (Buoro et al. 2017b,a, 2018), while the values of $CVI_{y=1}$ were estimated from the longitudinal observation of the COVID-19 positive patients considered in this study (authors a), using the same methodology as described in the previous section.

We considered 7 different ML models, commonly used in medical settings on tabular data, namely: Support Vector Machine (with RBF kernel) (SVM), Logistic Regression (LR), k-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (GB), ExtraTrees (ET). We evaluated, in particular, the scikit-learn implementations of the previous models, with default hyper-parameters. Further information on implementation details is in Appendix C. We did not evaluate deep learning models as such models often require extensive hyper-parameter optimization and are usually out-performed by other models on tabular data (Grinsztajn, Oyallon, and Varoquaux 2022). The impact of IV on the performance of the above mentioned ML models was evaluated by means of a repeated cross-validation evaluation procedure: for a total of 100 iterations, a 3-fold cross-validation procedure was applied. More in detail, in each 3-fold cross-validation the two training folds were used to train the ML models, while the test fold Te was used to obtain a perturbed fold Te_p as follows: for each instance $(x, y) \in Te$, a perturbed instance (x_p, y) was obtained to simulate the effect of individual variation, by sampling x_p from $N(x, \Sigma^{x,y})$. The trained ML model was then evaluated on both Te and Te_p to measure the impact of individual variation, if any, by comparing the distribution of average performance on the original test folds with that of the perturbed test folds. In terms of performance metrics, we considered the accuracy, the AUC and the F1 score. The robustness of the ML models to IV was evaluated by comparing the average performance on the non-perturbed and IV perturbed data: in particular, we considered a model to be robust to IV if the 95% confidence intervals for the above mentioned quantities overlapped (equivalently, the confidence interval of the difference included the value 0).

Results First of all, we assessed whether the IV perturbed data obtained by means of the considered generative model was statistically significantly different from the original data: ideally, to be realistic, IV-based perturbations of data should not influence too much the overall data distribution (Fraser 2009). To this purpose, we considered a subset of 4 predictive features (namely LY, WBC, NE and AST), which were previously shown to be among the most predictive features for the considered task (Chen et al. 2021b). We compared the distributions of the above mentioned features before and after the IV perturbations, by means of the Kolmogorov-Smirnov test with $\alpha = 0.01$. The obtained p-values were, respectively, 1 (for LY, WBC and NE) and 0.104 (for AST): thus, for all of the considered features, the null hypothesis of equal distributions for the IV perturbed and non-perturbed data could not be rejected.

The impact of IV on the ML models is reported in Figure 1. The difference in performance (baseline vs perturbed) was significant for all algorithms: indeed, for all algorithms, the confidence intervals on the baseline and IV perturbed data did not overlap. The best algorithms on the non-perturbed data were RF and ET, w.r.t. all considered metrics (AUC: 0.87, Accuracy: 0.8, F1: 0.8); while the best algorithms on the the IV perturbed data were SVM (w.r.t. AUC: 0.69, and Accuracy: 0.5) and GB (w.r.t. F1: 0.5).

These results highlight how, even though the distributions of highly predictive feature were not significantly affected by IV, IV nonetheless had a significant impact on the performance of the considered ML algorithms, that were therefore not robust to IV-related uncertainty. Algorithms, however, were not equal in their robustness (or lack thereof) w.r.t. IV: in particular, the more robust models were SVM (w.r.t. Accuracy, with average performance decrease 0.25, and AUC, with average decrease 0.12) and GB (w.r.t. F1 score, with average performance decrease 0.28), with all other models being significantly less robust (that is, having a significantly larger difference between baseline and IV perturbed performances). While this latter observation can be given a learning theoretical justification based on the notion of margin[3],

[3]Both SVM and GB are margin-based classifiers (Grønlund, Kamma, and Green Larsen 2020; Hanneke and Kontorovich 2021). It is not hard to see that the existence of a large margin on the non-perturbed data is a necessary (but not sufficient) condition for robustness to IV.

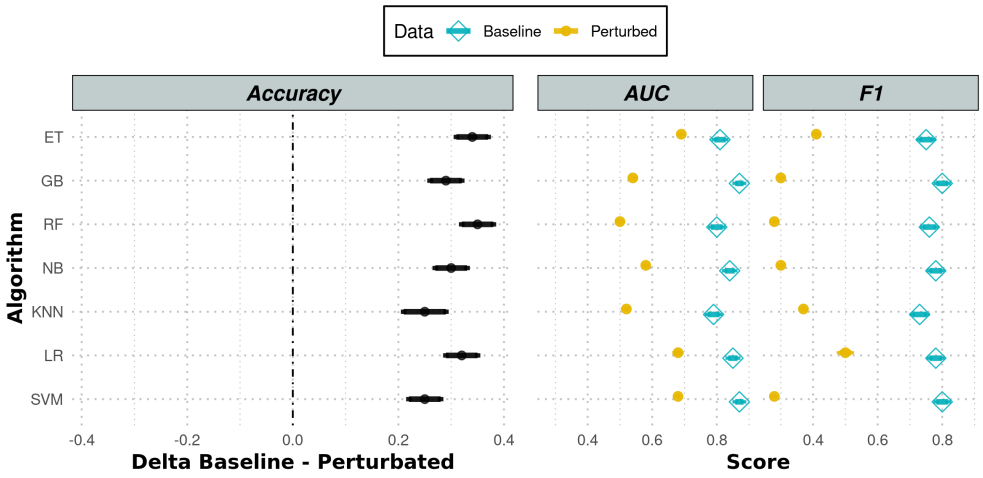


Figure 1: Results of the experiments for measuring the impact of IV on the performance of standard ML models. For each algorithm and metrics, we report the average and 95% confidence interval for both baseline (that is, non-perturbed) and IV perturbed data.

we note that even SVM and GB reported a significant decrease in performance on the IV-perturbed data: thus, even models that are usually considered to be robust to noise can nevertheless be strongly affected by IV.

Data Augmentation and Imprecisiation Methods to Manage Individual Variation

In light of the results reported in the previous section, which show the lack of robustness of standard ML models w.r.t. IV, in this section we investigate the application of more advanced methods that attempt to directly address the representation of IV in data and hence tackle the error decomposition show in Eq (2). In particular, we consider approaches based either on *data augmentation* or *data imprecisiation*. In both cases, we adopted the same experimental protocol described in the previous section (see below).

Data augmentation (Chen et al. 2021a; Van Dyk and Meng 2001) refers to regularization techniques that aim to increase the stability and robustness of a ML model by enriching the training set with new instances. In our setting, the idea is to inject further information related to the IV distribution within the model to improve generalization.

Since in the considered setting a generative model of IV was available, this latter was used to generate synthetic data points to augment the original training set. Basically, for each instance (x, y) in the training folds, we generated $n = 100$ new samples from the distribution $N(x, \Sigma^{x,y})$, so as to simulate the effect of having multiple observations, perturbed by IV, for each patient. We considered, in particular, the application of the above mentioned basic data augmentation strategy to the SVM (denoted as ACS) and Gradient Boosting (denoted as ACG) ML models, since these latter two were shown to be more robust to IV (see previous section). The pseudo-code for evaluating the data augmentation models is reported in Algorithm 1.

By contrast, data imprecisiation (Hüllermeier 2014; Lienen and Hüllermeier 2021b) refers to ML techniques by which data affected by some form of uncertainty are transformed into imprecise observations, that is distributions over possible instances, which are then used to train specialized ML algorithms. Formally speaking, an *imprecisiation scheme* is a function $i : X \times Y \mapsto [0, 1]^{X \times Y}$, where X is the feature space. In the experiments, we considered two commonly adopted imprecisiation schemes grounding on, respectively, probability theory and possibility the-

Algorithm 1: The procedure to evaluate the impact of IV on the data augmentation-based ML models.

procedure DATA_AUGMENTATION_EVAL(h: ML model, S: dataset, M: metric, n : number of augmented instances)
  for all iterations $i = 1$ to 100 **do**
    Split S in 3 class-stratified folds
    for all Tr: training fold, Te : test fold **do**
      $Tr_a = \emptyset$
      for all $(x, y) \in Tr$ **do**
        for all iteration $j = 1$ to n **do**
          Add to Tr_a (x_p, y), $x_p \sim N(x, \Sigma^{x,y})$
        end for
      end for
      $Te_p = \emptyset$
      for all $(x, y) \in Te$ **do**
        Add to Te_p (x_p, y), $x_p \sim N(x, \Sigma^{x,y})$
      end for
      Train h on Tr_a
      Eval h on Te $(M(h, Te))$, Te_p $(M(h, Te_p))$
    end for
  end for
  return The distributions of $M(h, Te)$ and $M(h, Te_p)$
end procedure

ory (Denœux, Dubois, and Prade 2020), namely:

$$i_{prob} : (x, y) \mapsto (N(x, \Sigma^{x,y}), y) \quad (3)$$

$$i_{poss} : (x, y) \mapsto (Gauss(x, \Sigma^{x,y}), y) \quad (4)$$

where $Gauss(a, b)$ denotes the Gaussian fuzzy vector, whose j-component is defined as $Gauss(a, b)_j(x) = e^{\frac{(x-a)^2}{b^2}}$. Intuitively, i_{prob} represents each instance affected by IV as a Gaussian probability distribution over possible instances, while i_{poss} represents each instance affected by IV as a Gaussian possibility distribution (equivalently, a Gaussian fuzzy set) over possible instances. Thus, the general idea of applying data imprecisiation (and corresponding ML algorithms) in our setting is to model the uncertainty due to IV by representing each instance as a cloud of points in the feature space whose distribution is determined by the IV parameters, as a form of regularization.

We considered three ML algorithms proposed in the learning from imprecise data literature, namely: k-Nearest Distributions (KND, also called Generalized kNN) (Zheng, Fung, and Zhou 2010), Support Measure Machine (SMM) (Muandet et al. 2017), Weighted re-Sampling Forest (WSF) (Seveso et al. 2020). See also Appendix C for hyper-parameter settings for the considered models. KND denotes the generalization of kNN to distribution-valued instances, namely we used the i_{prob} scheme[4] and Mahalanobis distance:

$$(x_1 - x_2)^T \frac{\Sigma^{x_1,y_1 -1} + \Sigma^{x_2,y_2 -1}}{2} (x_1 - x_2) \quad (5)$$

SMM, by contrast, refers to the generalization of SVM to instances represented as probability distributions (thus, only the i_{prob} imprecisiation scheme was considered). The SMM model grounds on the notion of a *kernel mean embedding* (Muandet et al. 2017), that is a generalization of the notion of kernel in ML to the the space of probability distributions, which could thus be seen as a measure of similarity between two imprecise instances. For computational complexity reasons, we considered the RBF kernel, which for normally distributed imprecise instances can be expressed in closed form as (Muandet et al. 2017):

$$RBF_{\gamma, i_{prob}} = \frac{e^{-\frac{(x_1-x_2)^T(\Sigma^{x_1,y_1}+\Sigma^{x_2,y_2}+\frac{1}{\gamma}I)^{-1}(x_1-x_2)}{2}}}{\sqrt{det(\gamma\Sigma^{x_1,y_1} + \gamma\Sigma^{x_2,y_2} + I)}} \quad (6)$$

Finally, the WSF model is an approximation algorithm to solve the generalized risk minimization problem (Hüllermeier 2014), a commonly adopted (Lienen and Hüllermeier 2021a,b) approach to deal with imprecise. WSF is based on a generalization of bootstrapped tree ensembles to instances represented as possibility distributions (thus, only the i_{poss} imprecisiation scheme was considered): in addition to the randomization w.r.t. the split point selection and the bootstrap re-sampling of the instances, an additional randomization on the feature values is considered. Specifically, for each tree in the ensemble, each imprecise instance $i_{poss}(x, y)$ in the corresponding bootstrap set is used to sample an instance (x', y'), by means of a two-step procedure (Dubois, Prade, and Sandri 1993): first, a number $\alpha \in [0, 1]$ is selected uniformly at random, then a random value is drawn from the α-cut $i_{poss}(x, y)^\alpha = \{(x', y') \in X \times Y : i_{poss}(x, y)(x', y') \geq \alpha\}$. A pseudo-code description of WSF, along with an analysis of its computational complexity and generalization error, is in Appendix B.

The KND, SMM and WSF models were implemented in Python, and evaluated in a setup similar to the one adopted for the data augmentation-based ML models, as shown in Algorithm 2. The full code for the algorithms and evaluation procedures is available on GitHub at anonymizedurl.

Algorithm 2: The procedure to evaluate the impact of IV on the data imprecisiation-based ML models.

procedure DATA_IMPRECISIATION_EVAL(h: ML model, S: dataset, M: metric, i: imprecisiation scheme)
  for all iterations $t = 1$ to 100 **do**
    Split S in 3 class-stratified folds
    for all Tr: training fold, Te : test fold **do**
      $Tr_a = \emptyset$; $Te_b = \emptyset$; $Te_p = \emptyset$
      for all $(x, y) \in Tr$ **do**
        $Tr_a.append(i((x, y)))$
      end for
      for all $(x, y) \in Te$ **do**
        $Te_b.append(i((x, y)))$
        Sample $(x_p, y) \sim N(x, \Sigma^{x,y})$
        $Te_p.append(i((x_p, y)))$
      end for
      Train h on Tr_a
      Eval h on Te_b ($M(h, Te_b)$), Te_p ($M(h, Te_p)$)
    end for
  end for
  return The distributions of $M(h, Te_b)$ and $M(h, Te_p)$
end procedure

Results The results for data augmentation and imprecisiation-based ML models are reported in Figure 2. For all models except SMM, the difference in performance on baseline and IV perturbed data was not significant. The best models on the non-perturbed data were SMM, WSF (w.r.t. AUC: 0.87) and WSF, ACG (w.r.t. Accuracy: 0.8, F1: 0.81), while the best models on the IV perturbed data were ACG and WSF (AUC: 0.86, Accuracy: 0.79, F1: 0.8). Comparing these results with those shown in the previous section, it is easy to observe that both data augmentation and data imprecisiation-based ML models were much more robust to IV perturbations than the standard ML models. Indeed, the most robust models (w.r.t. AUC: WSF and ACS, with average difference 0.003; w.r.t. Accuracy and F1: WSF and ACG, with average difference 0.006) were hardly impacted by IV. Even the least robust model (i.e., SMM) was much more robust than the standard ML models (average differences w.r.t. AUC: 0.08; w.r.t. Accuracy: 0.09, w.r.t. F1: 0.09).

In light of these results, we claim that data augmentation and imprecisiation can be helpful to improve robustness un-

[4]Since Mahalanobis' distance takes into account only the mean and scale, using i_{poss} scheme would result in the same algorithm.

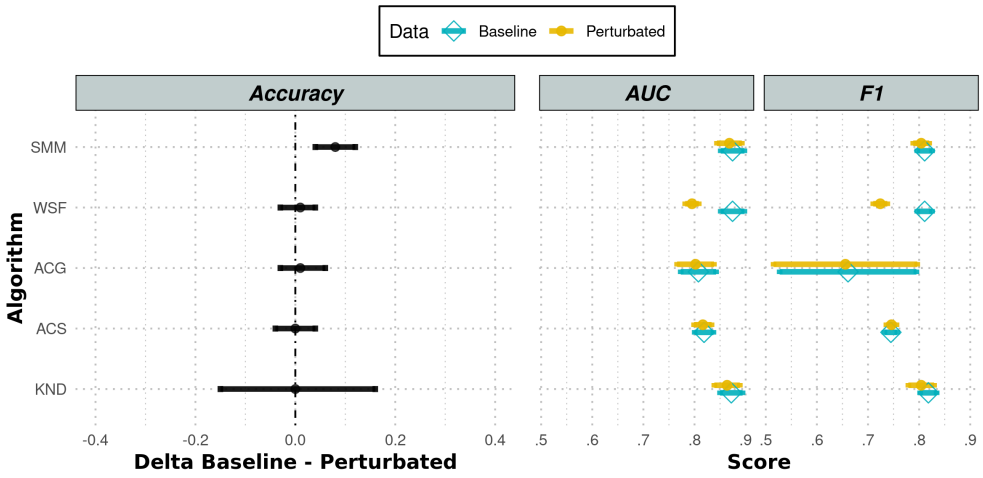


Figure 2: Results of the experiments for measuring the impact of IV on the performance of data augmentation-based and data imprecisiation-based ML models. For each algorithm and metrics, we report the average and 95% confidence interval for both baseline (that is, non-perturbed) and IV perturbed data.

der IV perturbations. We conjecture this to be due to directly taking into account information about IV in data representation and model training, which allows to strike a trade-off among the various components of the generalized error decomposition shown in Eq. (2). We note that these two approaches, while performing similarly in terms of accuracy and robustness, have different characteristics that may influence their suitability in practical scenarios. Data augmentation methods allow to use out-of-the-box ML models, since IV management is implemented as a pre-processing step: this is not the case for data imprecisiation-based approaches, which require specialized ML algorithms. By contrast, imprecisiation-based approaches have lower computational complexity and may thus scale better on larger datasets: e.g., if m is the training set size, d the number of features, n the number of ensembled models, and r the number of augmented instances then, the time complexities of SMM and WSF are, respectively, $O(m^2 d^3)$ and $O(n\,d\,m\log(m))$; by contrast, the complexities of ACS and ACG are, respectively, $O(m^2 r^2)$ and $O(n\,d\,m\,r\log(mr))$.

Conclusion

In this article we studied the impact of IV, an oft neglected type of uncertainty affecting data, on the performance and robustness of ML models. Crucially, through a realistic experiment on COVID-19 diagnosis, we showed that standard ML algorithms can be strongly impacted by the presence of IV, failing to generalize properly. Such an issue can severely limit the applicability and safety of ML methods in tasks where data are expected to be affected by IV, that is most applications in clinical settings and more generally in real-world domains where the manifestations of the phenomena of interest could exhibit time-varying patterns. Our results then imply that out-of-the-box methods cannot be naively applied in such domains. Nonetheless, every cloud has a silver lining, and we showed that more advanced learning methods, grounding on data augmentation and data imprecisiation, can achieve better robustness w.r.t. IV: this highlights the need to employ models that take into account the *generative history* underlying the data acquisition process, including the uncertainty due to IV, in their learning algorithms. Furthermore, we believe that our results highlight the importance of adopting proper algorithmic and experimental designs for ML studies in medicine: due to the potential impact of IV on the performance of ML models, data collection studies should be designed so as to enable the estimation of IV values which could then be used in the ML development phase. Thus, increasing emphasis should be placed on longitudinal studies, or otherwise studies in which multiple samples are collected for each involved patients under controlled conditions, so as to allow precise and reliable estimation of IV. We believe that these results could pave the way for the investigation of IV and its effects on the safety of ML models deployed in real-world clinical settings. Thus, we think that the following open problems could be of interest:

- In our experiments we assumed the IV distributions to be Gaussian with diagonal covariance. While this model is commonly adopted in the literature, we believe that further research should explore the relaxation of this assumption, by considering more general models of IV accounting for causal relationships among features;
- While we focused on the impact of IV in ML and briefly discussed IV in SLT, we believe the theoretical side of this issue merits further study: even though the problem of learning from distributional data has recently been investigated in SLT (Campagner 2021; Ma et al. 2021; Muandet et al. 2017), this area is still in its infancy;
- In this work we showed the impact of IV in the setting of COVID-19 diagnosis from blood tests. Future work should generalize our analysis to a broader spectrum of applications.

References

Aarsand, A. K.; Røraas, T.; Fernandez-Calle, P.; et al. 2018. The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation. *Clinical chemistry* 64(3):501–514.

Aarsand, A. K.; Fernandez-Calle, P.; Webster, C.; et al. 2020. The EFLM biological variation database.

Aarsand, A. K.; Kristoffersen, A. H.; Sandberg, S.; et al. 2021. The european biological variation study (EuBIVAS): Biological variation data for coagulation markers estimated by a bayesian model. *Clinical Chemistry* 67(9):1259–1270.

Adila, D., and Kang, D. 2022. Understanding out-of-distribution: A perspective of data dynamics. In *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*, 1–8. PMLR.

Aggarwal, R.; Sounderajah, V.; Martin, G.; et al. 2021. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ digital medicine* 4(1):1–23.

Akoglu, L. 2021. Anomaly mining: Past, present and future. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1–2.

Andaur Navarro, C. L.; Damen, J. A.; Takada, T.; et al. 2021. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj* 375:n2281.

Arratia, R., and Gordon, L. 1989. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology* 51(1):125–131.

authors, A. Anonymized publication.

authors, A. Anonymized publication.

Badrick, T. 2021. Biological variation: Understanding why it is so important? *Practical Laboratory Medicine* 23:e00199.

Bartlett, W. A.; Braga, F.; Carobene, A.; et al. 2015. A checklist for critical appraisal of studies of biological variation. *Clinical Chemistry and Laboratory Medicine (CCLM)* 53(6):879–885.

Beam, A. L.; Manrai, A. K.; and Ghassemi, M. 2020. Challenges to the reproducibility of machine learning models in health care. *Jama* 323(4):305–306.

Benjamens, S.; Dhunnoo, P.; and Meskó, B. 2020. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine* 3(1):1–8.

Buoro, S.; Carobene, A.; Seghezzi, M.; et al. 2017a. Short-and medium-term biological variation estimates of leukocytes extended to differential count and morphology-structural parameters (cell population data) in blood samples obtained from healthy people. *Clinica Chimica Acta* 473:147–156.

Buoro, S.; Seghezzi, M.; Manenti, B.; et al. 2017b. Biological variation of platelet parameters determined by the Sysmex XN hematology analyzer. *Clinica Chimica Acta* 470:125–132.

Buoro, S.; Carobene, A.; Seghezzi, M.; et al. 2018. Short-and medium-term biological variation estimates of red blood cell and reticulocyte parameters in healthy subjects. *Clinical Chemistry and Laboratory Medicine (CCLM)* 56(6):954–963.

Cabitza, F.; Campagner, A.; Ferrari, D.; et al. 2021. Development, evaluation, and validation of machine learning models for covid-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)* 59(2):421–431.

Campagner, A. 2021. Learnability in "learning from fuzzy labels". In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. IEEE.

Carobene, A.; Strollo, M.; Jonker, N.; et al. 2016. Sample collections from healthy volunteers for biological variation estimates' update: a new project undertaken by the working group on biological variation established by the european federation of clinical chemistry and laboratory medicine. *Clinical Chemistry and Laboratory Medicine (CCLM)* 54(10):1599–1608.

Carobene, A.; Guerra, E.; Locatelli, M.; et al. 2018. Providing correct estimates of biological variation—not an easy task. the example of s100-β protein and neuron-specific enolase. *Clinical Chemistry* 64(10):1537–1539.

Carobene, A.; Campagner, A.; Uccheddu, C.; et al. 2021. The multicenter european biological variation study (EuBIVAS): a new glance provided by the principal component analysis (PCA), a machine learning unsupervised algorithms, based on the basic metabolic panel linked measurands. *Clinical Chemistry and Laboratory Medicine (CCLM)*.

Chen, R. J.; Lu, M. Y.; Chen, T. Y.; et al. 2021a. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5(6):493–497.

Chen, Z.; Xu, W.; Ma, W.; et al. 2021b. Clinical laboratory evaluation of covid-19. *Clinica Chimica Acta* 519:172–182.

Christodoulou, E.; Ma, J.; Collins, G. S.; et al. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* 110:12–22.

Coiera, E.; Ammenwerth, E.; Georgiou, A.; et al. 2018. Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association* 25(8):963–968.

Coiera, E. 2019. The last mile: where artificial intelligence meets reality. *Journal of Medical Internet Research* 21(11):e16323.

Coskun, A.; Braga, F.; Carobene, A.; Ganduxe, X. T.; Aarsand, A. K.; Fernández-Calle, P.; Marco, J. D.-G.; Bartlett, W.; Jonker, N.; Aslan, B.; et al. 2020. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. *Clinical Chemistry and Laboratory Medicine (CCLM)* 58(1):25–32.

Denœux, T.; Dubois, D.; and Prade, H. 2020. Representations of uncertainty in artificial intelligence: Probability and

possibility. In *A Guided Tour of Artificial Intelligence Research*. Springer. 69–117.

Dubois, D.; Prade, H.; and Sandri, S. 1993. On possibility/probability transformations. In *Fuzzy logic*. Springer. 103–112.

Ellervik, C., and Vaught, J. 2015. Preanalytical variables affecting the integrity of human biospecimens in biobanking. *Clinical chemistry* 61(7):914–934.

Fraser, C. G. 2001. *Biological variation: from principles to practice*. American Association for Clinical Chemistry.

Fraser, C. G. 2009. Reference change values: the way forward in monitoring. *Annals of clinical biochemistry* 46(3):264–265.

Fröhlich, H.; Balling, R.; Beerenwinkel, N.; et al. 2018. From hype to reality: data science enabling personalized medicine. *BMC medicine* 16(1):1–15.

Futoma, J.; Simons, M.; Panch, T.; et al. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2(9):e489–e492.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.

Grønlund, A.; Kamma, L.; and Green Larsen, K. 2020. Margins are insufficient for explaining gradient boosting. *Advances in Neural Information Processing Systems* 33:1902–1912.

Haeckel, R.; Carobene, A.; and Wosniok, W. 2021. Problems with estimating reference change values (critical differences). *Clinica Chimica Acta* 523:437–440.

Hanneke, S., and Kontorovich, A. 2021. Stable sample compression schemes: New applications and an optimal svm margin bound. In *Algorithmic Learning Theory*, 697–721. PMLR.

Herlau, T.; Schmidt, M. N.; and Mørup, M. 2016. Completely random measures for modelling block-structured sparse networks. *Advances in Neural Information Processing Systems* 29.

Hotelling, H. 1992. The generalization of student's ratio. In *Breakthroughs in statistics*. Springer. 54–65.

Hüllermeier, E. 2014. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning* 55(7):1519–1534.

Kallenberg, O. 2017. *Random measures, Theory and applications*. Springer.

Li, X.; Zhang, S.; Zhang, Q.; et al. 2019. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology* 20(2):193–201.

Lienen, J., and Hüllermeier, E. 2021a. From label smoothing to label relaxation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8583–8591.

Lienen, J., and Hüllermeier, E. 2021b. Instance weighting through data imprecisiation. *International Journal of Approximate Reasoning* 134:1–14.

Liu, J.; Shen, Z.; Cui, P.; Zhou, L.; Kuang, K.; Li, B.; and Lin, Y. 2021. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8662–8670.

Ma, G.; Liu, F.; Zhang, G.; et al. 2021. Learning from imprecise observations: An estimation error bound based on fuzzy random variables. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. IEEE.

Morteza, P., and Li, Y. 2022. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8.

Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* 10(1-2):1–141.

Plebani, M.; Padoan, A.; and Lippi, G. 2015. Biological variation: back to basics. *Clinical Chemistry and Laboratory Medicine (CCLM)* 53(2):155–156.

Rabanser, S.; Günnemann, S.; and Lipton, Z. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems* 32.

Røraas, T.; Petersen, P. H.; and Sandberg, S. 2012. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clinical chemistry* 58(9):1306–1313.

Sandberg, S.; Carobene, A.; and Aarsand, A. K. 2022. Biological variation–eight years after the 1st strategic conference of EFLM. *Clinical Chemistry and Laboratory Medicine (CCLM)*.

Seveso, A.; Campagner, A.; Ciucci, D.; et al. 2020. Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings. *BMC Medical Informatics and Decision Making* 20(5):1–14.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Spodick, D. H., and Bishop, R. L. 1997. Computer treason: intraobserver variability of an electrocardiographic computer system. *The American journal of cardiology* 80(1):102–103.

Van Dyk, D. A., and Meng, X.-L. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10(1):1–50.

Wilkinson, J.; Arnold, K. F.; Murray, E. J.; et al. 2020. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health* 2(12):e677–e680.

Yin, J.; Ngiam, K. Y.; and Teo, H. H. 2021. Role of artificial intelligence applications in real-life clinical practice: systematic review. *Journal of medical Internet research* 23(4):e25759.

Zheng, K.; Fung, P. C.; and Zhou, X. 2010. K-nearest neighbor search for fuzzy objects. In *Proceedings of the 2010 ACM SIGMOD international conference on Management of data*, 699–710.

Appendix A: Data Characteristics

Descriptive statistics for the considered dataset are reported in Table A1.

Table A1: The list of features, along with the target. Mean and standard deviation are reported for continuous features, distribution of values is reported for discrete feature. For the discrete features we report the distribution of values. For the laboratory blood data, we also report the analytical (CVA) and biological (CVI) variation, differentiated by healthy vs non-healthy patients, and missing rate.

Features	Acronym	Units	Mean	Std	Missing	CVA	$CVI_{y=0}$	$CVI_{y=1}$
Alanine Transaminase	ALT	U/L	39.87	42.26	0.07	0.04	0.093	0.051
Aspartate Transaminase	AST	U/L	46.90	51.90	0.14	0.04	0.095	0.52
Alkaline Phosphatase	ALP	U/L	88.61	72.09	16.24	0.05	0.054	0.045
Gamma Glutamyl Transferase	GGT	U/L	67.48	140.52	17.09	0.035	0.089	0.036
Lactate Dehydrogenase	LDH	U/L	332.52	218.43	8.02	0.03	0.052	0.024
Creatine Kinase	CK	U/L	184.47	382.02	56.19	0.05	0.145	0.062
Calcium	CA	mg/dL	2.20	0.17	0.84	0.03	0.018	0.018
Glucosium	GLU	mg/dL	119.12	55.80	0.42	0.028	0.047	0.026
Urea	UREA	mg/dL	48.64	42.69	31.01	0.03	0.141	0.035
Creatinine	CREA	mg/dL	1.19	1.01	0.07	0.025	0.044	0.022
Leukocytes	WBC	10^9/L	8.65	4.77	0.00	0.019	0.111	0.033
Erythrocytes	RBC	10^{12}/L	4.55	0.72	0.00	0.009	0.018	0.010
Hematocrit	HCT	%	39.47	5.57	0.00	0.018	0.024	0.019
Neutrophils	NE	%	72.48	13.35	8.51	0.03	0.146	0.014
Lymphocytes	LY	%	18.58	11.11	8.51	0.036	0.11	0.043
Monocytes	MO	%	7.76	3.86	8.51	0.063	0.134	0.033
Eosinophils	EO	%	0.82	1.59	8.51	0.079	0.156	0.098
Basophils	BA	%	0.34	0.27	8.51	0.031	0.128	0.056
Sex	-	Female Male	42% 58%	-	-	-	-	-
Age	-	Years	61.19	18.89	-	-	-	-
Target	-	Positive Negative	53% 47%	-	-	-	-	-

Complete Blood Count data (i.e. features WBC, RBC, HCT, NE, LY, MO, EO, BA) was obtained by analysis of whole blood samples by means of a Sysmex XE 2100 haematology automated analyser. Biochemical data (ALT, AST, ALP, GGT, LDH, CK, CA, GLU, UREA, CREA) was obtained by analysis of serum samples by means of a Cobas 6000 Roche automated analyser. For each of the considered patients, COVID-19 positivity was determined based on the result of the molecular test for SARS-CoV-2 performed by RT-PCR on nasopharyngeal swabs: on a set of 165 cases for which the RT-PCR reported uncertain results, chest radiography and X-rays were also used to improve over the sensitivity of the RT-PCR test by combination testing.

Appendix B: The WSF Algorithm

Pseudo-code for the WSF algorithm is reported in Algorithm A1. As described in the main text, the computational complexity of WSF is $O(n\,d\,|S|\log(|S|))$ where d is the dimensionality of the input space.

In regard to the generalization error of WSF w.r.t. data generating random measure, for each base model h_i, let

$$L_S(h_i) = \sum_{(x,y)\in S} \mathbb{E}_{(x',y)\sim i_{poss}(x,y)} \left[\mathbb{1}_{h(x')\neq y}\right]$$

and $L_D(h_i) = \mathbb{E}_{S\sim D^m} L_S(h)$, where D is the intensity measure describe in Section "Individual Variation and Statistical Learning". Assume further, that for all $h \in \mathcal{H}$, with probability larger than $1-\delta$ if $(x-x')\Sigma^x(x-x') \leq T^2_{d,|S|-d}(1-\delta)$ it holds that $h(x) = h(x')$, where T is Hotelling's T-squared distribution (Hotelling 1992). Intuitively, this latter condition can be understood as a strong form of regularity for models in $\mathcal{H}$: if two instantiations likely come from the same distribution due to IV, then with high probability they will be classified in the same way by each $h \in \mathcal{H}$. Then, letting V_i be the out-of-bag sample for model h_i, by Hoeffding's inequality and above assumptions it follows that, with probability $1-\delta$, $L_D(h_i) \leq L_{V_i}(h_i) + \sqrt{\frac{\log(2|V_i|/\delta)}{2|V_i|}}$. Let $p = \sum_i L_{V_i}(h_i) + \sqrt{\frac{\log(2|V_i|/\delta)}{2|V_i|}} \leq \frac{1}{2}$. Then, assuming the h_i err independently of each other, and noting that WSF errs on an instance x iff at least n/2 base models err, with probability greater than $1-\frac{\delta}{n}$ the generalization error of WSF can be upper bounded through an application of Chernoff's bound for binomial distributions (Arratia and Gordon 1989) by $e^{-n\cdot KL(\frac{1}{2}||p)}$, where $KL(a||b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ is the Kullback-Leibler divergence.

Algorithm A1: The WSF algorithm.

procedure WSF(S: dataset, n: ensemble size, $\mathcal{H}$ model class)
  $Ensemble \leftarrow \emptyset$
  for all iterations $i = 1$ to n **do**
    Draw a boostrap sample S' from S
    $Tr_i \leftarrow \emptyset$
    for all $(x, y) \in S'$ **do**
      Sample $\alpha \sim U[0, 1]$
      Add $(x', y') \sim i_{poss}(x, y)^\alpha$ to Tr_i
    end for
    Add base model $h_i \in \mathcal{H}$ trained on Tr_i to $Ensemble$
  end for
  return $Ensemble$
end procedure

Appendix C: Implementation Details and Hyper-parameter Settings

All code was implemented in Python v. 3.10.4, using numpy v. 1.23.0, scikit-learn v. 1.1.1 and scikit-weak v. 0.2.0. For the standard ML models, we considered the default hyper-parameter values as defined in scikit-learn v. 1.1.1, with the exception of the random_state seed, which was set to 99 for all evaluated models to ensure reproducibility, and the max_depth hyperparameters for Random Forest and Gradient Boosting, which were set to 10 to avoid over-fitting and reduce the running time. For the data augmentation models we set the number of augmentation rounds to 100: for

ACS we used as base model a SVC with rbf kernel and default hyper-parameters, while for ACG we used a GradientBoostingClassifier with max_depth set to 0 and random_state set to 99 for consistency with the classical case. For SMM we used as kernel the RBF kernel defined in (6) with $\gamma = \frac{1}{\text{num. features}}$, while for WSF we used ExtraTreeClassifier as base classifier, we set the number of ensembled models to 100 and the random_state seed to 99. Finally, for KND we set the number of neighbors k to 5.

Part II

Dealing with Imprecision in the Output: Cautious Inference

The focus of the second part of this work will be on the handling of imprecision in the output of a ML model, that is the issue of how to employ imprecise set-valued predictions as an uncertainty quantification mechanism, to implement cautious inference algorithms that are able to denote and communicate their uncertainty about the issued classifications. These kinds of approaches have become widely relevant within the uncertainty quantification literature [133], due to their claimed simplicity of interpretation compared with techniques based on more complex uncertainty representation formalisms, as well as due to the availability of very effective, computationally efficient, and theoretically robust methods. Among such methods, in particular, this chapter will focus on three main classes of approaches that have attracted some attention in the recent years, namely selective prediction, decision-theoretic methods and conformal prediction. The aim of this part, then, will be to study two main theoretical questions related to these three families of techniques.

In the first chapter, the main objective will be to investigate the relationships among the above three mentioned cautious inference approaches, with the aim of addressing research question **P2.1** and thus providing a characterizations of the conditions under which (specific instantiations of) the corresponding learning algorithm could be thought as equivalent, or providing equivalent results. Focusing on a specific decision-theoretic family of approaches, namely *three-way decision*, this is a general post-hoc cautious inference method that stems from a direct generalization to set-valued prediction of the decision-theoretic expected utility principle, its relationship with selective and conformal prediction will be investigated. In particular, in the first section the learnability properties of three-way decision learning algorithms will be investigated within a generalization of the PAC learning framework, by which it will be shown that, under weak assumptions with respect to the selected learning algorithm, decision-theoretic methods generalize selective prediction ones. The following section, on the other hand, will focus on the relationship between three-way decision and conformal prediction, with the aim of providing a finer characterization of the learnability and validity properties for the former class of methods: to this end, the main theoretical contribution will be the proposal of

two procedures, to transform a three-way decision-based classifier into a conformal predictor and vice-versa, which will lead to a characterization of the conditions under which a three-way decision-based classifier can be isomorphically associated with a corresponding conformal predictor, with their set-valued predictions being generally equivalent. Complementing the above mentioned theoretical contributions, the first chapter of this part also explores three empirical contributions. First, showing, on a wide set of benchmark datasets, that the above mentioned constructions for transforming a three-way decision-based classifier into a conformal predictor and vice-versa can be applied to improve the accuracy of any given cautious classifiers with a minimal cost in terms of efficiency. Second, that the same constructions can be used to provide a generalization of conformal prediction to the setting of learning from imprecise data. Third, to illustrate a first initial attempt to investigate the user-oriented perspective on the impact of cautious inference methods on the socio-technical systems in which they are embedded. To this purpose, the results of a pilot study concerning the application of cautious inference to a complex real-world medical problem will be discussed, showing that cautious inference methods can provide an improvement not only terms of accuracy or robustness [12, 32], but also in terms of user-perceived usefulness.

The second chapter, instead, will focus on research question **P2.2** and investigate the issue of ensemble methods for cautious classifiers, from two different perspectives. In the first section, the application of cautious classifiers as base models to improve the generalization of standard ensemble techniques will be investigated by means of a large-scale comparison of several ensemble methods, encompassing both standard ensembling techniques, methods based on the combination of cautious predictors as well as methods based on other uncertainty representation formalisms. The main empirical contribution will be to show that ensemble methods based on the combination of cautious predictors, specifically so three-way decision-based models, can improve the robustness to noise as well as the generalization compared to standard ensemble techniques. In the second section, on the other hand, the application of ensemble methods to address the validity-efficiency trade-off in cautious inference

will be investigated, focusing on the study of different ensembling approaches in the framework of conformal prediction. The main theoretical contribution will be the introduction of a general methodology, drawing from the relationships between conformal prediction and possibility theory, to define and study the properties of methods to ensemble conformal predictors, as well as a characterization of the conditions for ensuring the validity of the resulting ensemble methods and investigation of their other properties, including efficiency, based on a general approach, grounding on copula theory, that relaxes assumptions (e.g. independence) commonly held in the existing literature. The main empirical contribution, on the other hand, will be the application of the above mentioned ensembling methods for cautious inference to the setting of multi-variate time series classification, showing that such techniques can improve the performance of state-of-the-art precise classification models as well as of state-of-the-art cautious inference methods.

Chapter 5

Cautious Inference Methods

As mentioned in the Introduction, cautious inference refers to a generalization of the standard supervised learning setting in which a ML model is allowed to be imprecise, in the sense that its output is not necessarily assumed to be a single class label but rather a set of possible labels, associated with an epistemic semantics: one of the the labels in the set is likely to be the correct one, according to the cautious predictor, but no further, more precise, information is provided.

In its most general form, cautious inference can be understood as a generalization of both standard supervised learning, which allows models of the form $h : X \to Y$, and learning with rejection [66], which allows models of the form $h : X \to Y \cup \{\perp\}$ where $\perp$ denotes an act of rejection from the model. To this purpose, cautious inference approaches consider models of the form $h : X \to 2^Y$. For any given instance x, the event $|h(x)| > 1$ denotes the presence of some degree of imprecision in the output of the model h, as a way to trade-off a reduced precision with an improvement in terms of accuracy [114, 122], whenever there is some uncertainty as to which label Y would be the correct one to predict. Indeed, as $|h(x)|$ increases, the precision of h decreases but, at the same time, the probability that the correct label y is in $h(x)$ grows, making h itself more robust and less prone to risky classifications albeit less directly informative.

To make this intuitive formulation of cautious inference more concrete, several approaches have been proposed in the literature [125, 141], starting from the pio-

neering work of Chow on optimal rejection rules [66]: such approaches encompass non-deterministic classifiers [84, 168], thresholding methods [45, 105, 108], decision-theoretic methods [45, 166, 173, 267], methods based on imprecise probabilities [72, 172, 271], conformal prediction [7, 14, 251] and selective prediction [265, 111, 192].

The aim of this chapter will be to address research question **P2.1** and thus investigate the theoretical properties of and relationships among three popular cautious inference frameworks among the above mentioned ones, namely decision-theoretic methods, conformal prediction and selective prediction, as representative examples of commonly used cautious inference approaches in practice. The main focus, in particular, will be on the study of theoretical properties of decision-theoretic methods, a term which refers to a general family of post-hoc cautious inference approaches that generalize to the context of set-valued imprecise predictions, the expected utility criterion. In the standard supervised setting, this latter can be expressed as:

$$y^*(x) = \arg\min_{y \in Y} \sum_{y' \neq y \in Y} p_h^{y'}(x) \cdot \epsilon_{y,y'} \tag{5.1}$$

$$= \arg\max_{y \in Y} \sum_{y' \in Y} p_h^{y'}(x) \cdot u_{y,y'} \tag{5.2}$$

where $p_h^{y'}(x)$ is the probability score assigned to class label y', for instance x, by the ML model h, $\epsilon_{y,y'}$ is the cost of making an error associated with predicting class label y when the true class label would be y' and $u_{y,y'}$ is the utility associated with predicting class label y when the true class label would be y'. Thus, the first formulation of the expected utility criterion amounts to selection of the class label that minimizes the expected mis-classification cost, while the second formulation amounts to selecting the class label that maximizes the expected utility: the two formulations are obviously equivalent as long as $\epsilon_{y,y'}$ and $u_{y,y'}$ are negatively, affinely correlated. The appeal of this decision rule in the ML setting is that, if the probability estimates p_h given by the model h are calibrated, then the above mentioned rule is equivalent to adopting the optimal Bayes classifier [283]. Thus, the generalization of this criterion aims to extend this association to the setting of cautious inference: such a generalization is usually achieved by requiring a generalization of the cost

function ϵ (equivalently, of the utility function u) to set-valued prediction, i.e. to functions of the form $c : 2^Y \times Y \to \mathbb{R}$. Several approaches have been proposed to address this generalization, among them the following sections will focus on a general formulation based on the theory of three-way decisions [45, 47, 267], an approach to decision-making with imprecision originally formulated in the context of Rough Set theory [268]. In its most general formulation [45], three-way decision adopts an error-based formulation based on the generalization of the error function ϵ as obtained by a decomposition of the form $c(T, y') = \epsilon_{T,y'} + \alpha(|T|)$, where $\epsilon : 2^Y \times Y \mapsto \mathbb{R}$ is a (normalized) *error function* such that $\max_{T \in 2^Y, y' \in Y} \epsilon_{T,y'} = 1$ and $\alpha : \mathbb{N} \to \mathbb{R}$ is an *imprecision penalty function* that is monotonically increasing and such that $\alpha(1) = 0$. Intuitively, while the ϵ term penalizes incorrect predictions as in the standard supervised setting, the α term further penalizes imprecise predictions, so as to allow the modeling in the cost function c of a trade-off principle between validity and efficiency of the set-valued predictions. The utility maximization principle is then generalized as:

$$T^*(x) = \arg\min_{T \in 2^Y} \sum_{y \in Y} \epsilon_{T,y} p_h^y(x) + \alpha(|T|) \tag{5.3}$$

The rationale to focus on this specific instantiation of the decision-theoretic approach stems from the above mentioned relationship with optimal Bayes classifiers, from its increasingly wide adoption in ML [47], as well as from its simplicity and flexibility. Indeed, even though three-way decision requires only a generalization of the cost function, in contrast with other approaches that also require the adoption of uncertainty representation frameworks which are more general and complex than probability theory [73, 166, 181, 264, 271], it is flexible enough to encompass also the previously mentioned methods as well as thresholding ones [45, 173], by appropriate selection of the cost function $c(T, y')$.

Despite these intuitively appealing characteristics, however, the theoretical properties of three-way decision, and decision-theoretic cautious inference methods more in general, have scarcely been considered in the literature, mostly focusing on their empirical evaluation or providing theoretical results resting on strong calibration

guarantees [67, 173] which are in general hard to satisfy based on finite samples. To address these limitations, then, this chapter will investigate the theoretical properties of three-way decision-based cautious inference, by relating this approach with two other cautious inference methods that have been widely studied from the theoretical perspective, namely selective prediction and conformal prediction.

Section 5.1, in particular, will be devoted to the study of PAC learning-style robustness guarantees for a specific approach to three-way decision-based cautious inference, based on the generalization of the empirical risk minimization paradigm to cautious inference, and its relationship with the selective prediction framework as generally formulated by El-Yaniv et al. [111, 257, 265]. This latter is a framework for learning with rejection in which a pair of classifiers $f : X \to Y, g : X \to [0, 1]$ is chosen so as to minimize the selective risk:

$$R(f, g) = \frac{\int_Z l(y, h(x))g(x)dz}{\int_Z g(x)dz},$$

which is an efficiency-reweighted version of the true risk in Eq. (1.1). Intuitively, g represents a *rejection* function that describes the probability of the selective predictor $h = (f, g)$ to abstain on any given instance x: in the specific case where $\forall x \in X, g(x) \in \{0, 1\}$, then $g(x) = 0$ implies that x is an uncertain instance for which the output prediction should be maximally imprecise. The selective risk, then, represents a trade-off between accuracy and efficiency, similarly to the cost function formulation adopted in three-way decision. One of the seminal results in selective prediction theory, proven in [257], shows that, given a class of classifiers $\mathcal{H}$ and a finite sample S, the above expression can be minimized with high probability guarantees (thus providing PAC learning-style bounds on both accuracy and efficiency) by a version space-based [171] algorithm defined as:

$$f(x) = ERM(\mathcal{H}, S) \tag{5.4}$$

$$g(x) = 1 \iff \forall h \in \mathcal{H} \text{ s.t. } L_S(h) \leq L_S(f) + \epsilon, f(x) = h(x). \tag{5.5}$$

The main theoretical contribution of Section 5.1 will be to show that the above mentioned PAC learning-style guarantees for three-way decision enable to draw a correspondence with the version space-based formulation of selective prediction, showing

in particular that, under a weak realizability assumption [128], this latter can be understood as a special case of three-way decision.

Section 5.2, by contrast, will be devoted to a more general theoretical characterization of three-way decision in its post-hoc formulation described previously, by exploring its connection with conformal prediction [251]. This latter is a post-hoc cautious inference method inspired by non-parametric statistics [126] and the theory of algorithmic complexity [150], in which imprecise predictions are generated by means of a *non-conformity measure* $m : Z^* \times Z \to \mathbb{R}$ which describes the similarity of a new instance (x, y) to a training set S. In particular, in the classification setting, where the non-conformity measure often depends on an underlying classifier h, a cautious inference method is obtained by a non-parametric, confidence set-style correction procedure that transforms h into a set-valued predictor by associating p-values $p_x(y)$ from a non-parametric permutation procedure to the class labels and then obtain a confidence set over Y:

$$p_x(y) = \frac{|\{(x_i, y_i) \in S : m_h(S, (x, y)) \leq m_h(S_i \cup \{(x, y)\}, (x_i, y_i))\}| + 1}{|S| + 1} \tag{5.6}$$

$$T^{\epsilon}_m(x) = \{y \in Y : p_x(y) > \epsilon\} \tag{5.7}$$

where S_i refers to i-deleted version of the training set S, in which the instance (x_i, y_i) has been removed from S. Intuitively, T^{ϵ} select all class label which, according to the underlying model h, would be not too dissimilar from the observations in the training set: the values $p_x(y)$, in particular, represent p-values. In the recent years, conformal prediction has attracted increasing interest due to its post-hoc nature, which similarly to three-way decision allows application to any existing ML algorithm [7], as well as due to its appealing theoretical properties, which are much stronger than typical PAC learning-style ones and establish that, under very weak conditions on the data generating distribution $\mathcal{D}$, the following result holds [251]:

Theorem 4. *Let m be a non-conformity measure, and assume that the instances in S has been sampled i.i.d. from $\mathcal{D}$. Let (x, y) be a new instance drawn from $\mathcal{D}$, $\epsilon \in [0, 1]$. Then $Pr_{(x,y)\in\mathcal{D}}(y \notin T^{\epsilon}_m(x)) \leq \epsilon$*

The main contribution of Section 5.2, then, will be the proposal of a novel procedure to transform any three-way decision cautious classifier into a conformal predictor, as well as of two reverse algorithms to transform a conformal predictor into a decision-theoretic classifier. By means of these algorithms, the relationships among these two cautious inference paradigms are investigated, providing general, robust learning bounds for three-way decision and general decision-theoretic algorithms based on results analogous to Theorem 4. These theoretical results will also be complemented with an experimental investigation of the proposed transformation algorithms showing that their successive application can significantly improve the accuracy of an underlying three-way decision cautious classifier. Finally, the chapter will conclude with an application of the above mentioned approaches in a user study in which a cautious inference algorithm is compared against a standard supervised learning model in terms of ecological utility and user satisfaction in a medical decision making problem, showing promising results in terms of improved user-perceived utility and satisfaction.

Preproceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 243–246

Three-way Learnability: A Learning Theoretic Perspective on Three-way Decision

Andrea Campagner, Davide Ciucci
Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, Viale Sarca 336/14, 20126 Milano, Italy

***Abstract*—In this article we study the theoretical properties of Three-way Decision (TWD) based Machine Learning, from the perspective of Computational Learning Theory, as a first attempt to bridge the gap between Machine Learning theory and Uncertainty Representation theory. Drawing on the mathematical theory of orthopairs, we provide a generalization of the PAC learning framework to the TWD setting, and we use this framework to prove a generalization of the Fundamental Theorem of Statistical Learning. We then show, by means of our main result, a connection between TWD and selective prediction.**

I. Introduction

In the recent years, there has been an increasing interest toward exploring the connections between learning theory and different uncertainty representation theories: This trend includes both the generalization of standard learning-theoretic tools and techniques to settings that involve representation formalisms that are more general than probability theory [1], [2], as well as the theoretical study of algorithms inspired by uncertainty representation [3], [4].

Among other uncertainty representation theories, Three-way decision (TWD) is an emerging computational paradigm, first proposed by Yao in Rough Set Theory [5], based on the simple idea of *thinking in three "dimensions"* (rather than in binary terms) when representing and managing computational objects [6]: in the Machine Learning (ML) [7] setting, this notion is usually declined in terms of allowing ML models to *abstain*. This approach attracted a large interest, also justified by promising empirical results in different ML tasks such as active learning [8], [9], cost-sensitive classification [10], clustering [11], [12], [9]. Despite these promising empirical results, the theoretical foundations of TWD-based ML received so far little attention [13], [14]. Indeed, even though, in the recent years, there has been an increasing interest toward generalizing computation learning theory (CLT) to cautious inference methods such as *selective prediction* [15] or the KWIK (*Knows what it Knows*) framework [16], such results cannot be easily applied to the TWD setting: While in the TWD setting abstention is a property of single classifiers; in the latter two frameworks abstention is usually achieved by consensus voting.

In this article, we study the generalization of a standard CLT mathematical framework, the so-called *Probably Approximately Correct* (PAC) learning framework, to the TWD setting: In particular, we will provide a generalization of the *Fundamental Theorem of Learning* to the TWD setting, and we show that our result generalizes previous results in the selective prediction setting. More in detail, the rest of this article is structured as follows: In Section II we provide the necessary mathematical background on TWD (in Section II-A) and CLT (in Section II-B); in Section III we describe the generalization of the PAC learning framework to the TWD setting and we prove our main result; finally, in Section IV, we summarise our contribution and describe possible research directions.

II. Background

A. Three-way Decision and Orthopairs

In this work we will refer to the formalization of TWD-based ML models (in the following, TW Classifiers) as *orthopairs*:

Definition 1. *An* orthopair *[17] over the universe X (which represents the instance space) is a pair of sets $O = (P, N)$ such that $P, N \subseteq X$ and $P \cap N = \emptyset$, with P and N standing, respectively, for* positive *and* negative. *The* boundary *is defined as $Bnd = (P \cup N)^c$.*

An orthopair represents an uncertain concept: Specifically, the status of the elements in the boundary is uncertain (i.e., it is not known whether they belong to the concept). Thus, a given orthopair stands as an approximation for a collection of consistent concepts:

Definition 2. *We say that an orthopair $O = (P, N)$ is* consistent *with a concept $C \subset X$ if $x \in P \implies x \in C$ and $x \in N \implies x \notin C$.*

Finally, we remark that it is possible to define different orderings between orthopairs: In particular, O_2 is *less informative* than O_1, denoted $O_2 \leq_I O_1$ if $P_2 \subseteq P_1$ and $N_2 \subseteq N_1$.

B. Computational Learning Theory

Computational Learning Theory [18] (CLT) refers to the branch of Machine Learning and Theoretical Computer Science focusing on the theoretical study of learning algorithms. Various mathematical formalisms have been proposed toward this goal, in this article we will refer to the PAC (probably approximately correct) learning framework, first proposed in [19]. Formally, let X be the instance space and Y be the target space, in this article we will focus on the *binary classification* setting, that is $Y = \{0, 1\}$. We assume that the observable data is generated i.i.d. according to an unknown probability distribution $\mathcal{D}$ over $X \times Y$. Let $\mathcal{H}$ be a hypothesis

©2022

class, that is a collection of functions $h : X \mapsto Y$, we define the *true risk* of h w.r.t. $\mathcal{D}$ as:

$$\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}\left[l(h(x), y)\right] = \int_{X\times y} l(h(x), y) d\mathcal{D}(x, y) \quad (1)$$

where $l : Y^2 \mapsto \mathbb{R}^+$ is a loss function. Since $\mathcal{D}$ is unknown, the true risk cannot be computed: It is usually approximated through the so-called *empirical risk* based on a sample, called *training set*, $S = (\langle x_1, y_1\rangle, \ldots, \langle x_m, y_m\rangle)$:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i) \quad (2)$$

Given a training set S, we denote by S_X the tuple $S_X = (x_1, ..., x_m)$, and by S_Y the tuple $S_Y = (y_1, ..., y_m)$. The *Empirical Risk Minimization* w.r.t. the hypothesis class $\mathcal{H}$ is the family of algorithms $ERM_{\mathcal{H},m} : (X \times Y)^m \mapsto \mathcal{H}$ s.t. $ERM_{\mathcal{H},m}(S) \in argmin_{h\in\mathcal{H}} L_S(h)$, where $S = (\langle x_1, y_1\rangle, \ldots, \langle x_m, y_m\rangle)$ is the training set.

The *Fundamental Theorem of Learning* [20] establishes a relation between the true risk and empirical risk for the ERM algorithm w.r.t. a hypothesis class $\mathcal{H}$ which depends only on the so-called VC dimension, a combinatorial dimension of the complexity of $\mathcal{H}$.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis class with VC dimension d. For each $\epsilon, \delta \in (0, 1)$ and distribution $\mathcal{D}$, then if $ERM_{\mathcal{H}}$ is given a dataset S of size $m \geq n_0$, with*

$$n_0 = O(\frac{d + ln(\frac{1}{\delta})}{\epsilon^2}) \quad (3)$$

with probability greater than $1 - \delta$, it holds that $|L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) - L_S(ERM_{\mathcal{H}}(S))| \leq \epsilon$. If, further, the realizability[1] *assumption holds, then, if S is a dataset of size $m \geq n_1$, with*

$$n_1 = O(\frac{d + ln(\frac{1}{\delta})}{\epsilon}) \quad (4)$$

with probability greater than $1 - \delta$, it holds that $L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) \leq \epsilon$.

Few works have studied the generalization of CLT results to hypothesis that can be described as orthopairs (that is, classifiers that can abstain on selected instances), mainly under the framework of *selective prediction* [21]: In this setting, the goal is to design learning algorithms $\mathcal{A}_{\mathcal{H},m} : (X \times Y)^m \mapsto \mathcal{O}_{\mathcal{H}}$, where $\mathcal{O}_{\mathcal{H}} \subseteq TW(\mathcal{H})$ (see Eq. (15)), s.t. $L_{\mathcal{D}}(A(S)) = 0$ but $A(S)$ is allowed to abstain on certain instances. This abstention is usually achieved either by the combination of a standard hypothesis $h : X \to Y$ with a rejection function $r : X \to \{\bot, \top\}$, or, equivalently, by consensus voting based on a version space $V \subseteq \mathcal{H}$ [21]. As we show in the following sections (specifically, in Section III-A) the setting we consider is a proper generalization of selective prediction. More recently, the application of orthopairs in CLT has been studied in the setting of adversarial machine learning [22], as well as to characterize the generalization capacity of hypothesis classes under generative assumptions [23]. We note, however, that even though the above mentioned work and the framework we study in this article rely on the representation formalism of orthopairs, the aims of these three frameworks are essentially orthogonal, also in terms of the mathematical techniques adopted: Indeed, while the three-way learning framework we study relies on a generalization of the ERM paradigm, the frameworks studied in [23], [22] rely on a transductive learning approach.

[1]Here realizability means that $\exists h \in \mathcal{H}$ s.t. $L_D(h) = 0$.

III. Three-way Learning

In this Section, we provide a first study of a generalization of standard Computational Learning Theory to the setting of TW Classifiers. As hinted in Section II-A, we will represent a TW Classifier as an orthopair O; then, a hypothesis space of TW Classifier will be represented as a collection $\mathcal{O}$ of orthopairs over X. In the TWD literature, the risk of a TW Classifier is usually evaluated by means of a cost-sensitive generalization of the 0-1 loss: $l_{TW}(O(x), y) = \begin{cases} 1 & O(x)\bot y \\ \lambda_a & x \in Bnd_O \\ 0 & otherwise \end{cases}$, where $\lambda_a \in [0, 0.5)$ is the cost of abstention, and $O(x)\bot y$ is the error case, that is $(x \in P_O \wedge y = 0) \vee (x \in N_O \wedge y = 1)$. Compared to the standard definition of risk adopted in the TWD literature we assume that the cost of error is always 1. Based on the loss function l_{TW} we can define both the true risk $\mathcal{L}^{TW}_{\mathcal{D}}$ and the empirical risk L^{TW}_S. Evidently, the risk of O can be decomposed as the sum of two functions:

$$\begin{aligned} \mathcal{L}^{TW}_{\mathcal{D}}(O) &= \mathbb{E}_{\mathcal{D}}\left[l_{TW}(O(x), y)\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{O(x)\bot y}\right] + \lambda_a \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{x\in Bnd_O}\right] \\ &= Pr_{x\sim\mathcal{D}}(O(x)\bot y) \\ &+ \lambda_a \cdot Pr_{x\sim\mathcal{D}}(x \in Bnd_O) \end{aligned} \quad (5)$$

The same decomposition can be similarly applied for the empirical risk. Let $\mathcal{E}_{\mathcal{D}}(O) = Pr_{x\sim\mathcal{D}}(O(x)\bot y)$ and $\mathcal{A}_{\mathcal{D}}(O) = \lambda_a \cdot Pr_{x\sim\mathcal{D}}(x \in Bnd_O)$. We denote with $\mathcal{O}^{OPT} = \{O \in \mathcal{O} : \mathcal{E}_{\mathcal{D}}(O) = min_{O'\in\mathcal{O}}\mathcal{E}_{\mathcal{D}}(O')\}$. We say that $\mathcal{D}$ is *weakly realizable* w.r.t. $\mathcal{O}$ if $\forall O^* \in \mathcal{O}^{OPT}$ it holds that $\mathcal{E}_{\mathcal{D}}(O^*) = 0$. If, furthermore, $\exists O^* \in \mathcal{O}^{OPT}$ s.t. $\mathcal{A}_{\mathcal{D}}(O^*) = 0$, then we say that $\mathcal{D}$ is *strongly realizable*. Through this article, we will assume only weak realizability. Compared to the realizability assumption, weak realizability assumption is indeed much weaker. As an example if the vacuous TW classifier $O_\bot = (\emptyset, \emptyset) \in \mathcal{O}$, then every distribution $\mathcal{D}$ is trivially weakly realizable w.r.t $\mathcal{O}$, while it is clearly not strongly realizable.

Let $\epsilon \in (0, 1)$, $\alpha \in (0, \lambda_a)$, then $O \in \mathcal{O}$ makes an (ϵ, α)-failure if one of the following holds:

$$\mathcal{E}_{\mathcal{D}}(O) > \epsilon, \quad \mathcal{A}_{\mathcal{D}}(O) > min_{O\in\mathcal{O}^{OPT}}\mathcal{A}_{\mathcal{D}}(O) + \alpha \quad (6)$$

Thus, O (ϵ, α)-fails if either its error is greater than ϵ, or if its abstention rate is greater, by a margin of at least α, than the lowest abstention rate among those TW Classifiers that make no error. We thus define the notion of *Three-way learnability*:

Definition 3. *$\mathcal{O}$ is* Three-way learnable *if exists an algorithm $C_m : (X \times Y)^m \mapsto \mathcal{O}$ and $m_{\mathcal{O}} : (0, 1)^2 \times (0, \lambda_a) \mapsto \mathbb{N}$ such*

that, for each distribution $\mathcal{D}$, $\epsilon \in (0,1), \delta \in (0,1), \alpha \in (0, \lambda_a)$ $\forall m \geq m_{\mathcal{O}}(\epsilon, \delta, \alpha)$, and given $S \sim \mathcal{D}^m$, C returns $O \in \mathcal{O}$, s.t. O (ϵ, α)-fails with probability lower than δ

We then want to provide a characterization for TW learnability, similar to Theorem 1. For this purpose, we first define a generalization of the ERM algorithm to the TWD setting, that we call Three-way Risk Minimization (TW-RM):

Definition 4. *Let $S \in (X \times Y)^m$. Then,*

$$\begin{aligned} TWRM(S) = argmax_{O \in \mathcal{O}} \mathcal{A}_{X \setminus S_X}(O) \text{ s.t.} \\ \mathcal{E}_S(O) = min_{O' \in \mathcal{O}} \mathcal{E}_S(O') \\ \mathcal{A}_S(O) = min_{O' \in \mathcal{O}^{OPT}} \mathcal{A}_S(O') \end{aligned} \tag{7}$$

Thus, the TWRM algorithm selects, among those TW classifiers with minimal empirical risk, the TW classifier with maximal abstention rate on the non-observed instances (that is, the instances in $X \setminus S_X$). This has the goal of minimizing errors on non-observed instances, and is analogous to the *maximum margin* principle, and the *disagreement coefficient* in version space learning, active learning and selective prediction [15].

In order to characterize TW learnability, given hypothesis class $\mathcal{O}$ (i.e. a collection of orthopairs), we define two derived hypothesis classes. Given any orthopair $O \in \mathcal{O}$ we can define a classifier $h_O : X \mapsto \{0,1\}$, as: $h_O(x) = \begin{cases} 1 & x \in Bnd_O \\ 0 & otherwise \end{cases}$.

We denote the collection of such binary classifiers as $\mathcal{H}_{\mathcal{O}} = \{h_O : O \in \mathcal{O}\}$. Thus, given $\mathcal{O}$, the derived hypothesis class $\mathcal{H}_{\mathcal{O}}$ describes the abstention capacity of $\mathcal{O}$: In the classical setting $\mathcal{H}_{\mathcal{O}} = \{h_0\}$, where $\forall x \in X, \quad h_0(x) = 0$, as no classifier in $\mathcal{O}$ is able to abstain: For all $O \in \mathcal{O}$, $Bnd_O = \emptyset$.

In regard to the second derived hypothesis class, we observe that the order $\leq_I$ defined in Section II-A defines a meet semi-lattice [17] on $\mathcal{O}$ with minimal element $O_\perp = (\emptyset, \emptyset)$. Then, we denote with $\mathcal{O}^\top = \{O \in \mathcal{O} : \nexists O' \in \mathcal{O} \text{ s.t } O \leq_I O'\}$, i.e. $\mathcal{O}^\top$ is the anti-chain of maximally informative elements of $\mathcal{O}$.

We now prove a generalization of Theorem 1 to the TWD setting, through which we show that the TW learnability of a hypothesis class $\mathcal{O}$, using the TWRM algorithm, can be characterized in terms of the derived hypothesis classes $\mathcal{H}_{\mathcal{O}}$ and $\mathcal{O}^\top$. In order to do so, we consider the VC dimension of the two derived hypothesis classes $\mathcal{H}_{\mathcal{O}}$ and $\mathcal{O}^\top$ as follows:

$$AVC(\mathcal{O}) = VC(\mathcal{H}_{\mathcal{O}}) \tag{8}$$

$$\begin{aligned} EVC(\mathcal{O}) = \sup\{|S| : S \subseteq X \wedge \forall C \subseteq S. \exists O \in \mathcal{O} \\ \text{s.t. } C = (P_O \cap S) \wedge (Bnd_O \cap S) = \emptyset\} \end{aligned} \tag{9}$$

Then, the following result holds:

Theorem 2. *Let $\mathcal{O}$ be s.t. $AVC(\mathcal{O}) = d_a$ and $EVC(\mathcal{O}) = d_e$. Then, for any distribution $\mathcal{D}$ weakly realizable w.r.t $\mathcal{O}$, $\epsilon, \delta \in (0,1), \alpha \in (0, \lambda_a)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m larger than :*

$$O\left(max\left\{\frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right), \left(\frac{\lambda_a}{\alpha}\right)^2\left(d_a + ln\frac{1}{\delta}\right)\right\}\right) \tag{10}$$

then, $TWRM_{\mathcal{O}}(S)$ (ϵ, α)-fails with probability lower than δ.

Proof. We want to guarantee that the following bound holds:

$$\begin{aligned} Pr_{fail} = P(S : \exists O \in \mathcal{O} \quad \wedge \\ |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon \quad \vee \\ |\mathcal{A}_D(O) - \mathcal{A}_S(O)| > \alpha) < \delta \end{aligned} \tag{11}$$

Then, the results would follow by uniform convergence. By the union bound, it holds that:

$$\begin{aligned} Pr_{fail} \leq Pr(S : \exists O \in \mathcal{O}, |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon) \\ + Pr(S : \exists O \in \mathcal{O}, |\mathcal{A}_D(O) - \mathcal{A}_S(O)| > \alpha), \end{aligned} \tag{12}$$

thus, it is sufficient to jointly upper bound the two summands by $\frac{\delta}{2}$. As regards the error rate (i.e $\mathcal{E}$) bound, we note that:

$$\begin{aligned} Pr(S : \exists O \in \mathcal{O}, |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon) \\ Pr(S : \exists O \in \mathcal{O}^\top, \mathcal{E}_D(O) > \epsilon) \end{aligned} \tag{13}$$

Since $O^\top$ is a binary hypothesis class, then, by Theorem 1, the above bound holds with probability greater than $1 - \delta$ as long as $|S| \geq \frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right)$. Furthermore, by uniform convergence this holds, in particular, for $TWRM_{\mathcal{O}}(S)$.

For the abstention part, the same line of reasoning can be applied, however, as we only assume weak realizability, only the result in Theorem 1 that applies to agnostic learning can be used. Then, as long as $|S| \geq \left(\frac{\lambda_a}{\alpha}\right)^2\left(d_a + ln\frac{1}{\delta}\right)$ it holds that $|\mathcal{A}_D(O) - \mathcal{A}_S(O)| < \alpha$ with probability greater than $1 - \delta$. This holds, in particular for $TWRM_{\mathcal{O}}(S)$, and thus the theorem follows by uniform convergence and Eq. (12). □

As a simple corollary, in the strong realizable setting, it can be easily verified that:

Corollary 1. *Let $\mathcal{O}$ be s.t. $AVC(\mathcal{O}) = d_a$ and $EVC(\mathcal{O}) = d_e$. Then, for any distribution $\mathcal{D}$ strongly realizable w.r.t $\mathcal{O}$, $\epsilon, \delta \in (0,1), \alpha \in (0, \lambda_a)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m larger than :*

$$O\left(max\left\{\frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right), \frac{\lambda_a}{\alpha}\left(d_a + ln\frac{1}{\delta}\right)\right\}\right) \tag{14}$$

then, $TWRM_{\mathcal{O}}(S)$ (ϵ, α)-fails with probability lower than δ.

Note that, if $|\mathcal{O}| < \infty$, then it can be easily shown that $AVC(\mathcal{O}) \leq log_2(\mathcal{H}_{\mathcal{O}})$. Furthermore, it also holds that $EVC(\mathcal{O}) \leq log_2(\mathcal{O}^\top)$, as if O satisfies Eq. (8), then it obviously holds that $Bnd_O = \emptyset$ and hence $O \in \mathcal{O}^\top$.

A. Three-way Learning and Selective Prediction

Finally, we show that the proposed mathematical framework and the obtained results can be used to establish a connection between TWD and *selective prediction*. This result relies on the connection between version space theory and orthopairs [17], and allows us to derive a generalization bound, originally proven by El-Yaniv et al. [21], for selective prediction: This shows that the latter setting can be understood as a special case of TWD. Let $\mathcal{H}$ be a hypothesis class of binary classifiers, we call the Three-way Closure of $\mathcal{H}$, denoted as $TW(\mathcal{H})$, the hypothesis space obtained as:

$$TW(\mathcal{H}) = \bigcup \{O_H : H \subseteq \mathcal{H}\} \tag{15}$$

where, for each $H \subseteq \mathcal{H}$, $O_H = (\{x : \forall h \in H.h(x) = 1\}, \{x : \forall h \in H.h(x) = 0\})$. Basically, we associate with each possible version space H in $\mathcal{H}$ a corresponding orthopair O_H which abstains on every instance for which the hypotheses in H disagree [17]. Then we can prove the following result:

Corollary 2. *Let $|\mathcal{H}| < \infty$, let $\mathcal{O} = TW(\mathcal{H})$ the Three-way Closure of $\mathcal{H}$, and let $\lambda_a = 1$. Then, for any distribution $\mathcal{D}$ strongly realizable w.r.t $\mathcal{O}$, and for any $\delta \in (0,1)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m, then:*

1) *With probability 1 it holds that $\mathcal{E}_{\mathcal{D}}(TWRM_{\mathcal{O}}(S)) = 0$;*
2) *With probability greater than $1-\delta$ it holds that:*

$$\mathcal{A}_{\mathcal{D}}(TWRM_{\mathcal{O}}(S)) \leq O\left(\frac{1}{m} ln\left(\frac{|\mathcal{H}_{\mathcal{O}}|}{\delta}\right)\right) \quad (16)$$

$$= O\left(\frac{1}{m}\left(|\mathcal{H}| + ln\frac{1}{\delta}\right)\right) \quad (17)$$

Proof. The first equality easily follows from strong realizability and by noting that, by definition of $TW(\mathcal{H})$, $x \notin Bnd_{TWRM_{\mathcal{O}}(S)}$ iff $(x \in S_X \vee \exists v \in \{0,1\}.\forall h \in \{h' \in \mathcal{H} : \mathcal{E}_S(h) = 0\},\ h(x) = v)$. In regard to the second statement, the first inequality follows by standard algebraic manipulations. The equality, on the other hand, follows by noting that $|\mathcal{H}_{\mathcal{O}}| = 2^{\mathcal{H}}$ (as $TW(\mathcal{H})$ contains a TW classifier for each possible subset of hypotheses in $\mathcal{H}$). □

IV. Conclusion

In this article, we aimed at providing an initial study on the generalization of CLT results to the TWD setting. To this purpose, we first proposed an extension of the standard PAC learning framework to the TWD setting, that we called Three-way Learning and showed that our results generalize the previously known results in the selective prediction literature. As our results represent only a first direction in the theoretical study of TWD as applied to Machine Learning, we believe that the following questions would be of particular interest:

- Our analysis in Theorem 2 relies on a generalization of the VC dimension to the TWD setting. Tighter bounds can usually be obtained by relying on concepts such as Rademacher complexities or covering numbers [18]. How can these be generalized to TWD?
- In Corollary 2 we proved that, in the realizable case, selective prediction can be understood as a special case of TWD learning. Does this analysis also applies to the agnostic (i.e. non-realizable) setting [15]?
- PAC-Bayes bounds [24] study generalization bounds that apply when a probability distribution is defined over the hypothesis space. How can the PAC-Bayes framework be generalized to TWD? Interestingly, a very similar open problem has recently been posed also in Belief Function Theory (BFT) [25]. Due to the connection with random sets, a belief function can be seen as a probability distribution over orthopairs [26]: Then, the generalization of the PAC-Bayes framework to TWD would also enable studying the relationships between TWD and BFT.

References

[1] E. Hüllermeier, "Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization," *International Journal of Approximate Reasoning*, vol. 55, no. 7, pp. 1519–1534, 2014.

[2] G. Ma, F. Liu, G. Zhang, and J. Lu, "Learning from imprecise observations: An estimation error bound based on fuzzy random variables," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–8.

[3] S. Abbaszadeh and E. Hullermeier, "Machine learning with the sugeno integral: The case of binary classification," *IEEE Transactions on Fuzzy Systems*, 2020.

[4] E. Hüllermeier and A. F. Tehrani, "On the vc-dimension of the choquet integral," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2012, pp. 42–50.

[5] Y. Yao, "Three-way decision: an interpretation of rules in rough set theory," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2009, pp. 642–649.

[6] M. Ma, "Advances in three-way decisions and granular computing," *Knowl.-Based Syst*, vol. 91, pp. 1–3, 2016.

[7] M. Hu, "Three-way bayesian confirmation in classifications," *Cognitive Computation*, pp. 1–20, 2021.

[8] F. Min, S.-M. Zhang, D. Ciucci, and M. Wang, "Three-way active learning through clustering selection," *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2020.

[9] H. Yu, X. Wang, G. Wang, and X. Zeng, "An active three-way clustering method via low-rank matrices for multi-view data," *Information Sciences*, vol. 507, pp. 823–839, 2020.

[10] H. Li, L. Zhang, X. Zhou, and B. Huang, "Cost-sensitive sequential three-way decision modeling using a deep neural network," *International Journal of Approximate Reasoning*, vol. 85, pp. 68–78, 2017.

[11] M. K. Afridi, N. Azam, and J. Yao, "Variance based three-way clustering approaches for handling overlapping clustering," *International Journal of Approximate Reasoning*, vol. 118, pp. 47–63, 2020.

[12] P. Wang and Y. Yao, "Ce3: A three-way clustering method based on mathematical morphology," *Knowledge-based systems*, vol. 155, pp. 54–65, 2018.

[13] A. Campagner, F. Cabitza, P. Berjano, and D. Ciucci, "Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches," *Information Sciences*, vol. 579, pp. 347–367, 2021.

[14] A. Campagner and D. Ciucci, "A formal learning theory for three-way clustering," in *International Conference on Scalable Uncertainty Management*. Springer, 2020, pp. 128–140.

[15] R. Gelbhart and R. El-Yaniv, "The relationship between agnostic selective classification, active learning and the disagreement coefficient." *J. Mach. Learn. Res.*, vol. 20, no. 33, pp. 1–38, 2019.

[16] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl, "Knows what it knows: a framework for self-aware learning," *Machine learning*, vol. 82, no. 3, pp. 399–443, 2011.

[17] D. Ciucci, "Orthopairs and granular computing," *Granular Computing*, vol. 1, no. 3, pp. 159–170, 2016.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[19] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[20] V. Vapnik, "On the uniform convergence of relative frequencies of events to their probabilities," in *Doklady Akademii Nauk USSR*, vol. 181, no. 4, 1968, pp. 781–787.

[21] R. El-Yaniv *et al.*, "On the foundations of noise-free selective classification." *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[22] S. Goldwasser, A. T. Kalai, Y. Kalai, and O. Montasser, "Beyond perturbations: Learning guarantees with arbitrary adversarial test examples," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 859–15 870, 2020.

[23] N. Alon, S. Hanneke, R. Holzman, and S. Moran, "A theory of pac learnability of partial concept classes," *arXiv preprint arXiv:2107.08444*, 2021.

[24] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor, "Pac-bayes analysis beyond the usual bounds," *arXiv preprint arXiv:2006.13057*, 2020.

[25] F. Cuzzolin, *The geometry of uncertainty*. Springer, 2017.

[26] Y. Yao and P. Lingras, "Interpretations of belief functions in the theory of rough sets," *Information sciences*, vol. 104, no. 1-2, pp. 81–106, 1998.

Information Sciences 579 (2021) 347–367

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches

Andrea Campagner [a,*], Federico Cabitza [a], Pedro Berjano [b], Davide Ciucci [a]

[a] Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano–Bicocca, viale Sarca 336, 2012620126 Milano, Italy
[b] IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi 4, 20161 Milano, Italy

ARTICLE INFO

Article history:
Received 16 April 2021
Received in revised form 9 July 2021
Accepted 2 August 2021
Available online 4 August 2021

Keywords:
Three-way decision
Cautious learning
Conformal prediction
Set-valued prediction
Decision support

ABSTRACT

The aim of this article is to study the relationship between two popular Cautious Learning approaches, namely: *Three-way decision* (TWD) and *conformal prediction* (CP). Based on the novel proposal of a technique to transform three-way decision classifiers into conformal predictors, and vice versa, we provide conditions for the equivalence between TWD and CP. These theoretical results provide error-bound guarantees for TWD, together with a formal construction to define cost-sensitive cautious classifiers based on CP. The proposed techniques are then applied and evaluated on a collection of benchmark and real-world datasets. The results of the experiments show that the proposed techniques can be used to obtain cautious learning classifiers that are competitive with, and often out-perform, state-of-the-art approaches. Further, through a qualitative medical case study we discuss the usefulness of cautious learning in the development of robust Machine Learning.
© 2021 Elsevier Inc. All rights reserved.

1. Introduction

In this article, we study the problem of *Cautious Learning* [7]. This latter is a generalization of supervised learning in which the Machine Learning (ML) models are allowed to express set-valued predictions. The set-valued predictions allow the ML models to highlight a possible state of uncertainty, that should require further intervention from a human decision maker [5].

Recently, such techniques have been advocated as a promising approach [17] to develop reliable ML-based decision support in so-called *decision-critical domains*, e.g. medicine, social policing. Indeed, in all these settings, errors induced by ML models could have high-impact consequences. Therefore the decision makers could accept less precise, but more reliable predictions. Set-valued predictions could then be used by the decision-maker either to take a decision, if the risk of doing so is not assumed to be too high; or to prompt the need to collect more information, so as to foster human-in-the-loop decision-making [14,23].

Cautious learning methods clearly entail a trade-off between different quality dimensions, that should be properly evaluated so as take into account different desirable properties. These may include:

- *Cost-sentitiveness* [10]: that is, whether the model properly takes into account information about the utilities and costs of the different alternative decisions;

* Corresponding author.
E-mail address: a.campagner@campus.unimib.it (A. Campagner).

https://doi.org/10.1016/j.ins.2021.08.009
0020-0255/© 2021 Elsevier Inc. All rights reserved.

- *Validity* [35]: that is, whether the performance of the model can be reliably bounded, usually through a theoretical analysis;
- *Efficiency* [36]: that is, whether the set-valued predictions provided by the model are as *informative* as possible.

In recent years, many different cautious learning techniques have been proposed to strike a balance among these properties. These include models based on imprecise probabilities [41], or belief functions [27]; selective classification [12]; three-way decision [43] (TWD); and conformal prediction [35] (CP).

While all the mentioned models have been successfully employed in empirical settings, their theoretical characterization largely remains an open problem. First, there is a lack of works attempting to characterize the validity of cautious learning methods (with the exception of CP [35]); second, the relationships and similarities among different approaches have not yet been investigated.

In this work, we address these gaps by focusing on two popular approaches, namely *three-way decision* (TWD) and *conformal prediction* (CP):

- TWD, inspired by Rough Set theory and human decision making [43], is a generalization of decision-theory to the setting of set-valued predictions. Intuitively, given a new instance and a loss function, a TWD-based classifier would assign the instance to the set-valued prediction associated with minimal loss;
- CP, by contrast, is a technique to obtain calibrated confidence predictors. For each new given instance, a conformal predictor would return a nested collection of set-valued predictions, each with an associated error probability lower bound [35]. A cautious learning algorithm can then be defined from a conformal predictor by selecting a specific probability threshold.

These differences notwithstanding, the two methods also share some similarities [6]. Indeed, both methods can be applied as a post-processing step to any standard (i.e., non-cautious) learning method [2,5]; both methods are distribution-free; and both methods make relatively weak assumptions. See Fig. 1 for a graphical representation of TWD-based classifiers and CP, in comparison with a standard (i.e., single-valued) ML model.

The aim of this paper, then, is to study the relationships among these two models, and to characterize when these two different approaches can be considered equivalent. To this purpose, we first define techniques to transform a TWD-based ML model into a CP one, and vice versa. Harnessing this relationship, we investigate two main theoretical questions:

- Under which conditions a TWD-based model is guaranteed to be valid, and with which error bounds? We answer this question through Theorems 2 and 3, by which it is shown that, under very general assumptions, TWD classifiers are valid;
- Under which conditions TWD and CP methods are equivalent? We answer this question through Theorems 4 and 5, by which conditions for the equivalence between TWD and CP methods are provided.

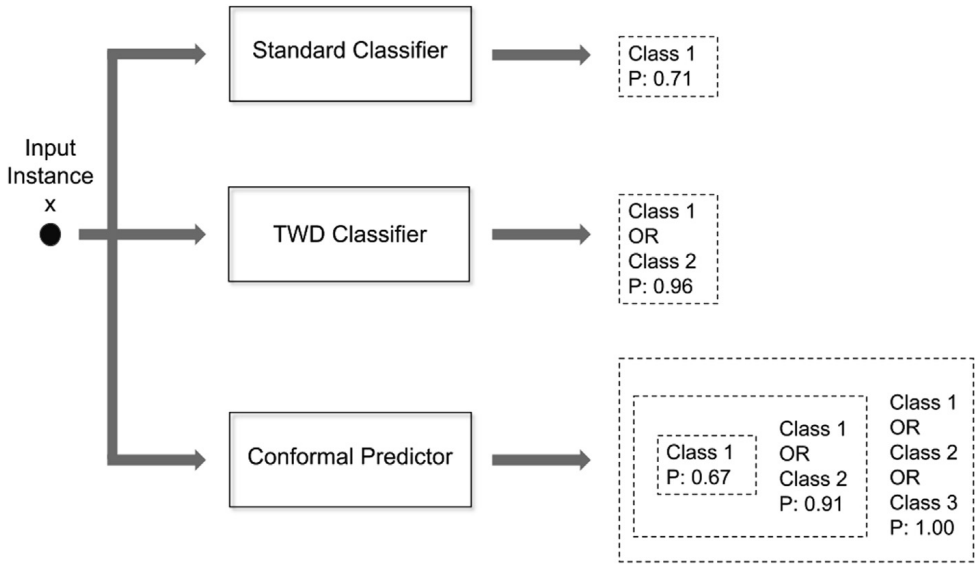


Fig. 1. A graphical representation of standard (classification) ML models, TWD models and CP models. Given an input instance x, a standard classifier would provide as output either a single class (in the example, Class 1) together with a confidence score, or a confidence score distribution. By contrast, a TWD classifier would provide as output the set of labels (in the example, Class 1 or Class 2) which is optimal w.r.t. a specific loss function, possibly together with an aggregated confidence score. Finally, a conformal predictor provides as output a nested collection of sets of labels, each with an associated (lower) probability bound.

Moreover, by means of a set of quantitative experiments, we show that, when the above mentioned assumptions do not hold, the proposed techniques can be used to improve the validity of TWD classifiers.

The rest of this article is structured as follows. In Section 2, the necessary technical background on Machine Learning, TWD in the ML domain, and CP is provided. In Section 3, we study the relationship between TWD and CP. Specifically, in Section 3.1, TWD is used as a basis to define a CP. Through this construction, the validity of TWD-based ML models is formally established. Conversely, in Section 3.2 we discuss how TWD can be used to define cost-sensitive cautious learning methods based on CP algorithms. Through these constructions conditions for the equivalence between TWD and CP are established. In Section 4, the empirical performance of the proposed constructions is investigated, through a set of experiments on real-world datasets. The results of these experiments are then discussed in Sections 4.2 and 4.3; while in Section 5 a short medical case study is discussed. Finally, in Section 6, the obtained results are summarized, and possible future lines of research are outlined.

2. Background

In this section, we recall the necessary background about ML, TWD and CP.

2.1. Supervised machine learning

Let X be the input space, i.e., a set of objects described as vectors of feature values. Let Y be the target space, i.e., the set of classes. Then, a classification algorithm, w.r.t. a sample space Z and a hypothesis space $\mathcal{H}$, is a function $\mathcal{A}: 2^Z \mapsto \mathcal{H}$. When $Z = X \times Y, \mathcal{A}$ is denoted as a *supervised* classification algorithm; by contrast, when $Z = X \times 2^Y$, $\mathcal{A}$ is a *weakly-supervised* [15] classification algorithm.

Let $S \subseteq Z$ be a sample drawn i.i.d. from an unknown distribution $\mathcal{D}$; $\mathcal{H}$ be an hypothesis space; and $l: \mathcal{H} \times Z \mapsto \mathbb{R}^+$ be a loss function. Then, the goal of the *machine learning problem* is to find a hypothesis $h \in \mathcal{H}$ with minimal (or small) *true risk*:

$$Risk_{\mathcal{D}}(h, l) = \int_{z \in Z} l(h, z) d\mathcal{D}(z). \tag{1}$$

Since $\mathcal{D}$ is unknown, the true risk cannot be computed. Hence, the aim is to minimize a proxy of the true risk, such as the *empirical risk*, based on the finite sample S:

$$Risk_S(h, l) = \frac{1}{|S|} \sum_{z \in S} l(h, z). \tag{2}$$

Empirical Risk Minimization (ERM) is the algorithm that, given $\mathcal{H}$ and a training set S, selects one of the $h \in H$ s.t. $Risk_S(h, l) = \min_{h' \in \mathcal{H}} Risk(h', l)$. We denote any such h as h_S. The ERM learning paradigm has been generalized to the setting of weakly supervised learning [15,16] by means of *generalized loss functions*. These latter are usually expressed in the form $l^S(h, \langle x, Y_x \rangle) = A(\{l(h, \langle x, y \rangle) : y \in Y_x\})$; where $A \in \{\min, \max, \text{mean}\}$.

In the following, we assume that $\mathcal{H}$ is a class of scoring classifiers. A scoring classifier is a function $h: X \mapsto Y$ s.t. $h = dec \circ s$, where $s: X \mapsto \mathbb{R}^{|Y|}$ is a scoring function (mapping an instance $x \in X$ to a distribution of scores); and $dec: \mathbb{R}^{|Y|} \mapsto Y$ is a decision function (mapping a distribution $s(x)$ to a single label). The decision function dec is usually defined as $dec(s(x)) = argmax_{y \in Y} s(x)_y$, where $s(x)_y$ denotes the score assigned to label y.

A *cautious* classifier [11] is a function $h: X \mapsto 2^Y$. Thus, a cautious classifier maps instances to *sets* of labels. The semantics attached to set prediction $h(x) \subseteq Y$ is that the correct label $\hat{y}$ is *likely* to be in $h(x)$.

2.2. Three-way decision

Three-way decision (TWD) [43] is a framework for information and uncertainty management, inspired by human decision-making and rough set theory [42], that generalizes standard decision theory.

In the binary setting, one considers three regions: a positive, or acceptance, region; a negative, or rejection, region; and a boundary, or non-commitment, region. This latter region, in particular, represents lack of knowledge, or (temporary) abstention, in regards to the status of the objects it contains.

With respect to the ML setting [42], according to TWD, every instance can be classified as either belonging to a given class (and thus not belonging to all others); not belonging to a given class; or being in the *boundary*, that is a region that represents lack of knowledge with respect to class assignment. This latter property makes TWD useful for the development of cautious classifiers, by means of a theoretically sound and cost-sensitive approach [6].

Indeed, this approach has been successfully applied in the ML literature for many tasks. Li et al. [22] proposed a cautious classification model for binary classification based on modeling uncertain boundaries; while Xu et al. [39] proposed a generalization of TWD-based cautious classification to multi-class problems, using sequential TWD. Liu et al. [24] proposed a TWD method based on the combination of logistic regression and decision-theoretic rough sets; Zhang et al. [49] proposed an approach for cost-sensitive cautious classification based on TWD and ensemble learning; similarly, Yue et al. [47] and

Savchenko [32] proposed computationally efficient techniques for cautious classification based on TWD and deep learning. Min et al. [29] proposed an approach for cautious classification of weakly-supervised data using TWD and active learning; Campagner et al. [5] proposed an approach for weakly supervised learning and multi-class cautious classification based on TWD and statistical learning methods; Gu et al. [13] studied approaches to TWD in group-decision making based on imprecise probabilistic linguistic assessments; Zhou et al. [50] studied and compared different approaches for TWD based on coarse and fuzzy data. More recently, Liu et al. [23] also discussed the interpretability and usefulness of cautious classification methods based on TWD. For a more general discussion about TWD in ML, we refer the reader to the recent surveys by Campagner et al. [6] and Liu et al. [25]. Furthermore, approaches for cautious classification based on TWD have recently been investigated also from a theoretical and conceptual perspective: Liu et al. [26] studied an alternative model for TWD based on optimization; Yao [46] studied the connections between TWD and set-based approaches; Yao [45] explored the foundations of TWD based on geometrical and numerical concepts; while Xu [38] studied the connections between TWD-based classification and the theory of confusion matrices. We refer the reader to the reviews by Yang et al. [40] and Yao [44] for further details.

In TWD, the loss function is generalized as a set-valued function $l : 2^Y \times Y \mapsto \mathbb{R}$, so as to model the loss w.r.t. a set-valued prediction. In this article, we consider the multi-class formulation of TWD classification [5]. In this latter approach, the loss function l can be decomposed in two parts, namely an *error cost function* and an *abstention cost function*. Formally, let

- $err : 2^Y \times Y \mapsto \mathbb{R}$ be an *error cost function*. Intuitively, $err(S, y)$ represents the cost of predicting S, when $y \notin S$ is the correct label;
- $\alpha : \mathbb{N}^+ \mapsto \mathbb{R}$ be an *abstention cost function*. We assume that

$$\forall i > 1, \alpha(1) = 0 < \alpha(i) \leqslant \alpha(i+1) \leqslant \min_{A \in 2^Y, y \in Y} err(A, y). \tag{3}$$

  Inuitively, $\alpha(|A|)$ represents the cost of making a set-valued prediction A that contains the correct label y.

Let h be a scoring classifier, its generalized loss is defined as:

$$Loss_{TWD}(A) = \sum_{y \notin A} h(x)_y \cdot err(A, y) + \alpha(|A|) \sum_{y \in A} h(x)_y. \tag{4}$$

Then, the TWD classifier $\mathcal{W}_h$ is defined, for each $x \in X$, as:

$$\begin{aligned} &\mathcal{W}_h(x) = \arg\min_{A \in 2^Y} \{|A| : \\ &A \in \arg\min_{B \in 2^Y} Loss_{TWD}(B)\}. \end{aligned} \tag{5}$$

Hence, for each x, the result of $\mathcal{W}_h(x)$ is (one of) the smallest sets having minimal generalized loss. In Example 1 we briefly describe the calculations involved in the definition of a simple TWD classifier.

Example 1. Let *err* be the constant 1 function, and $\alpha(|A|) = \frac{|A|-1}{|Y|}$, with $Y = \{1, 2, 3, 4, 5\}$.

Let h be a scoring classifier, and x an instance such that

$$h(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle.$$

Since the error cost function *err* is uniform, the optimization problem in Eq. (5) can be solved using a greedy algorithm [5]. Thus, the following holds:

$$\begin{aligned} Loss_{TWD}(\{2\}) &= 0.7 \\ Loss_{TWD}(\{2,5\}) &= 0.56 \\ Loss_{TWD}(\{1,2,5\}) &= 0.55 \\ LossTWD(\{1,2,3,5\}) &= 0.64 \\ Loss_{TWD}(Y) &= 0.8 \end{aligned}$$

Therefore, $\mathcal{W}_h(x) = \{1, 2, 5\}$.

By definition, the TWD classifier $\mathcal{W}_h(x)$ is the cautious classifier with minimal risk, *under the assumption* that the probability scores returned by h approximate the probability of error (i.e. h is calibrated). However, the calibration of h is, in general, only a sufficient condition for the correctness of the set-valued predictions issued by the TWD classifier $\mathcal{W}_h(x)$. In Section 3, based on the relationship between TWD and CP, we study some conditions under which TWD classifiers are guaranteed to be valid.

2.3. Conformal prediction

Conformal Prediction [35] (CP) is a cautious learning approach that allows to define calibrated classifiers. Since its introduction, the CP framework has been adapted to different settings, including clustering [30], anomaly detection [21], active

learning [28], semi-supervised learning [1]. Furthermore, CP has been successfully applied in many empirical settings, including cancer detection [48], cybersecurity [37], drug discovery [4]. See [2] for a recent review on CP.

Conventionally, CP is applied in the *transductive* learning setting [34]. In this latter setting, the instances are assumed to be sampled sequentially. Nonetheless, conformal predictors can be applied also to the standard *inductive* learning paradigm, by using a separate *validation* (or, *calibration*) set [35]. For simplicity of presentation, here and in Section 3, we focus on the transductive setting.

A *non-conformity measure* is a permutation-invariant function $M : 2^{X\times Y} \times (X \times Y) \mapsto \mathbb{R}$, i.e., given $S = (\langle x_1, y_1\rangle \ldots, \langle x_n, y_n\rangle)$, it holds that $M(S, \langle x, y\rangle) = M(\pi(S), \langle x, y\rangle)$ for every permutation π. Intuitively, a non-conformity measure quantifies how much a new instance $\langle x, y\rangle$ differs from past examples in S. More formally, the value of a non-conformity measure, for a given instance $\langle x, y\rangle$, represents a *statistic* for a non-parametric testing procedure [35].

Let $S_{x_i,x}$ be the result of exchanging $\langle x_i, y_i\rangle$ with $\langle x, y\rangle$ in S. Then, the *conformal predictor* determined by M is a function $\Gamma_M : 2^{X\times Y} \times X \times [0, 1] \mapsto 2^Y$, defined as:

$$\Gamma^{\epsilon}_M(S, x) = \{y | p^{x,y} > \epsilon\}, \tag{6}$$

where $\epsilon \in [0, 1]$ and $p^{x,y}$ is defined as:

$$p^{x,y} = \frac{|\{i = [1, n] : M(S, \langle x, y\rangle) \leqslant M(S_{x_i,x}, \langle x_i, y_i\rangle)\}| + 1}{n + 1}. \tag{7}$$

Intuitively speaking, relying on the above mentioned interpretation of the non-conformity measure as a testing statistic, the value $p^{x,y}$ is the *p-value* for the null hypothesis that the instance $\langle x, y\rangle$ comes from the same distribution as S [2]. Therefore, the labels in $\Gamma^{\epsilon}_M(S, x)$ are those for which the previously mentioned null hypothesis cannot be rejected (at a threshold confidence value of ϵ).

We denote with $im(\Gamma_M)$ the image of Γ_M, that is:

$$im(\Gamma_M) = \{A \subseteq Y : \exists \epsilon \in [0, 1] \text{s.t. } \Gamma^{\epsilon}_M(x) = A\}. \tag{8}$$

Thus, $im(\Gamma_M)$ is a nested collection of sets $A_1 = \varnothing \subseteq \ldots A_i \subseteq A_n = Y$. Each set $A_i \in im(\Gamma_M)$ has an associated ϵ_i s.t. $\epsilon_1 = 1, \epsilon_n = 0$. The map $p(i) = \epsilon_i$ represents the p-value function [2] of the statistical procedure defined by Γ_M.

Notably, a cautious classifier Γ^{ϵ}_M can be constructed from a conformal predictor Γ_M, by selecting an appropriate ϵ.

In Example 2, we illustrate the computations involved in the definition of a conformal predictor, by using an approach based on 1-nearest neighbor [35].

Example 2. In this example, the non-conformity measure will be defined as:

$$M_{1NN}(S, \langle x, y\rangle) = \frac{\min\limits_{x' \in S : y_{x'} = y} d(x, x')}{\min\limits_{x' \in S : y_{x'} \neq y} d(x, x')}, \tag{9}$$

where d is a metric. Thus, the similarity of a new example $\langle x, y\rangle$ w.r.t. the training set S is high when x is more similar to the instances in S associated with the same label, than to instances associated with a different label.

Consider the following single-feature training set

$$S = \{i_1 = \langle 0.75, 0\rangle, i_2 = \langle 0.90, 0\rangle, i_3 = \langle 0.48, 1\rangle\},$$

Let $x = 0.615$ be a new instance to be classified. Then $M_{1NN}(S, \langle x, 0\rangle) = M_{1NN}(S, \langle x, 1\rangle) = 1$ and, similarly:

$$\begin{aligned}
&M_{1NN}(S_{i_1,\langle x,0\rangle}, i_1) = 0.5\\
&M_{1NN}(S_{i_3,\langle x,1\rangle}, i_3) = 0.5\\
&M_{1NN}(S_{i_1,\langle x,1\rangle}, i_1) = 1.15\\
&M_{1NN}(S_{i_2,\langle x,0\rangle}, i_2) = 0.36\\
&M_{1NN}(S_{i_2,\langle x,1\rangle}, i_3) = 0.53.
\end{aligned}$$

By contrast, $M_{1NN}(S_{i_3,\langle x,0\rangle}, i_3)$ is undefined, as there is no instance with label 1 in the associated training set. Thus, we set $M_{1NN}(S_{i_3,\langle x,0\rangle}, i_3) = +\infty$. Therefore, $p^{x,0} = p^{x,1} = \frac{1}{2}$ and the corresponding conformal predictor is defined as:

$$\Gamma^{\epsilon}_{M_{1NN}} = \begin{cases} \varnothing & \epsilon > \frac{1}{2} \\ \{0, 1\} & otherwise \end{cases}$$

As previously mentioned, the main advantage of CP, compared to other cautious classification approaches, is that every conformal predictor is *valid*, i.e. the following result holds:

Theorem 1 (*Vovk et al. [35]*). *Let S, x be sampled i.i.d. from the same distribution $\mathcal{D}$, y be the true (but unknown) label associated with x. Let M be a non-conformity measure and $\epsilon \in [0, 1]$. Then, taken Γ_M the conformal predictor based on M, it holds that Γ_M is conservatively valid, that is:*

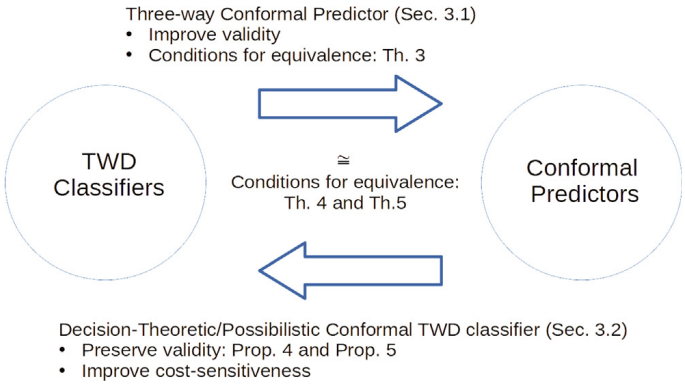


Fig. 2. A graphical illustration of the main results in Section 3.

$$Pr[y \notin \Gamma^{\epsilon}_{M}(S,x)] \leqslant \epsilon. \tag{10}$$

Thus, the probability of error of $\Gamma^{\epsilon}_{M}(S,x)$ is no greater than ϵ. Numerous approaches have been proposed in the literature to define conformal predictors, both based on algorithm-specific approaches [33]; and general-purpose ones [18]. One of the most popular general-purpose methods [18] is based on a score-based classifier h (see Section 2.1). In this case, a non-conformity measure based on h can be defined as:

$$M_h(S, \langle x, y\rangle) = \max_{y' \in Y}\{s(x)_{y'}\} - s(x)_y. \tag{11}$$

3. Methods

In this section, we study the relationships between TWD and CP. The main contents of this section, as well as the results of our study, are summarised in Fig. 2. As previously mentioned, in the following we focus on the transductive setting. As highlighted in Section 2.3, note, however, that the properties of conformal predictors we study in this section hold also in the inductive setting.

3.1. From three-way decision to conformal prediction

In this section, we address the first of the research questions mentioned in Section 1. Namely, we study whether, and under which conditions, TWD classifiers are valid. To this purpose, we first show that TWD can be used to design conformal predictors. Then, we provide sufficient and necessary conditions for the validity of TWD classifiers. The connection between TWD and CP is then generalized to the setting of weakly supervised learning.

Let us first consider the standard supervised setting (i.e. $Z = X \times Y$). Let $S = (\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle)$ be the training set. The *three-way non-conformity measure*, based on a given TWD classifier $\mathcal{W}$, can be defined as:

$$M_{\mathcal{W}}(S, \langle x, y\rangle) := l(\mathcal{W}_S, \langle x, y\rangle) = \begin{cases} \alpha(|\mathcal{W}_S(x)|) & y \in \mathcal{W}_S(x) \\ err(\mathcal{W}_S(x), y) + \alpha(|\mathcal{W}_S(x)|) & otherwise \end{cases} \tag{12}$$

where $l(\mathcal{W}_S, \langle x, y\rangle)$ denotes the loss of the prediction $\mathcal{W}_S(x)$, given $y \in Y$.

Thus, the *three-way non-conformity measure* assigns, to each instance $\langle x, y\rangle$, the loss incurred by using $\mathcal{W}_S$ to predict the label of x. It is easy to observe that, for any $\mathcal{W}, M_{\mathcal{W}}$ is indeed a non-conformity measure:

Proposition 1. *Let $\mathcal{W}$ be a three-way classifier, then $M_{\mathcal{W}}$, defined as in Eq. (12), is a non-conformity measure.*

Proof. Let S be a sample, and $\mathcal{W}_S$ the three-way classifier defined by S. Then, by definition, for any permutation S_1 of S, it holds that $\mathcal{W}_S = \mathcal{W}_{S_1}$. Therefore, the training algorithm is permutation-invariant. □

The construction described in Section 2.3, applied to the three-way non-conformity measure $M_{\mathcal{W}}$, allows to define the *three-way conformal predictor* (TWCP) as:

$$\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = \{y | p^{x,y} > \epsilon\}, \tag{13}$$

$$p^{x,y} = \frac{|\{i = 1, \ldots, n : \text{Pred is verified}\}| + 1}{n+1}, \tag{14}$$

$$Pred := l(\mathcal{W}_S, \langle x, y\rangle) \leqslant l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y_I\rangle). \tag{15}$$

The calculations involved in the definition of the TWCP are briefly illustrated in Example 3.

Example 3. Let $Y = \{0, 1, 2\}$, and let S be a training set s.t. $S = \{i_1 = \langle x_1, 0\rangle, i_2 = \langle x_2, 1\rangle, i_3 = \langle x_3, 1\rangle, i_4 = \langle x_4, 2\rangle\}$. Let $\mathcal{W}$ be a TW classifier s.t. $\mathcal{W}_S(x_1) = \{0, 2\}, \mathcal{W}_S(x_2) = \{0\}, \mathcal{W}_S(x_3) = \{1\}$ and $\mathcal{W}_S(x_4) = \{1, 2\}$.

Let x be a new instance. Assume, for simplicity, that $\forall i_j, \mathcal{W}_S = \mathcal{W}_{S_{i_j,x}}$ and that $\mathcal{W}_S(x) = \{0, 1\}$. Let $err = 1$ and $\alpha(|A|) = \frac{|A|-1}{|Y|-1}$.

Then $M_{\mathcal{W}}(S, \langle x, 1\rangle) = 0.5, M_{\mathcal{W}}(S, \langle x, 0\rangle) = 0.5, M_{\mathcal{W}}(S, \langle x, 2\rangle) = 1.5$, while

$$\begin{aligned} M_{\mathcal{W}}(S_{i_1,x}, i_1) &= 0.5 \\ M_{\mathcal{W}}(S_{i_4,x}, i_4) &= 0.5 \\ M_{\mathcal{W}}(S_{i_2,x}, i_2) &= 1 \\ M_{\mathcal{W}}(S_{i_3,x}, i_3) &= 0. \end{aligned}$$

Therefore $p^{x,0} = p^{x,1} = \frac{4}{5}, p^{x,2} = \frac{1}{5}$ and the TWCP $\Gamma_{\mathcal{W}}$ is defined as:

$$\Gamma^{\epsilon}_{\mathcal{W}} = \begin{cases} \varnothing & \epsilon > \frac{4}{5} \\ \{0, 1\} & \frac{1}{5} < \epsilon \leqslant \frac{4}{5} \\ Y & otherwise \end{cases}$$

Since $M_{\mathcal{W}}$ is a non-conformity measure, as a consequence of Theorem 1, it holds that the TWCP $\Gamma_{\mathcal{W}}$ is conservatively valid:

Corollary 1. *Let S, x be sampled i.i.d. from the distribution, and let $\hat{y}$ be the correct label associated with x. Then, for any $\epsilon, Pr[\hat{y} \notin \Gamma^{\epsilon}_{\mathcal{W}}(S, x)] \leqslant \epsilon$, that is $\Gamma_{\mathcal{W}}$ is conservatively valid.*

Proof. The result follows directly from Theorem 1 and the observation (see Prop. 1) that $M_{\mathcal{W}}$ is a non-conformity measure. □

The previous result holds for any CP algorithm, thus, in particular, for the TWCP. Nonetheless, the previous result does not provide any information about the validity of the original TWD classifier. Then, we ask two main questions: can the validity of a TWCP be used to obtain performance bounds for the corresponding TWD classifier? Under which conditions it holds that a TWD classifier and the corresponding TWCP are equivalent?.

In regard to the first question, note that the transformation from a TWD classifier $\mathcal{W}$ to the corresponding TWCP $\Gamma_{\mathcal{W}}$ provides a bound on the probability of error of $\mathcal{W}$. Indeed, if $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S, x))$, then, the following bound follows from Corollary 1:

$$Pr[y \notin \mathcal{W}_S(x)] \leqslant \arg\min_{\epsilon \in [0,1]} \{\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = \mathcal{W}_S(x)\}. \tag{16}$$

Consequently, a sufficient condition for the validity of the TWD classifier $\mathcal{W}$ would be that $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S, x))$. Then, the next result provides a characterization of this property:

Theorem 2. *The following two conditions are equivalent:*

1. $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S, x))$;
2. $\exists \langle x_i, y_i\rangle \in S$ *such that*

$$l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y_i\rangle) < \min_{y \notin \mathcal{W}_S} err(\mathcal{W}_S, y).$$

Proof. First, we prove that 1 implies 2. Note that $\forall y \in \mathcal{W}_S(x) = A$, then either $l(A, y) = 0$ (when $A = \{y\}$) or $l(A, y) = \alpha(|A|)$. Furthermore, by definition of α and err, it holds that $\forall A, \alpha(|A|) \leqslant \min_{B \subseteq Y, y \notin B} err(B, y)$. Thus, if 2 does not hold, then it exists $y \notin \mathcal{W}_S(x)$ s.t. $p^{x,y} = 1$. Consequently, the smallest A_i in $im(\Gamma_{\mathcal{W}}(S, x))$ is s.t. $\mathcal{W}_S \cup \{y\} \subseteq A_i$. The proof for the converse implication is analogous. □

Thus, as a consequence of Theorem 2, every non-trivial TWD classifier[1] is valid, and can be associated with an error upper bound. This latter error bound quantifies the probability that the correct label is not contained in the set-valued prediction issued by the TWD classifier.

Furthermore, this latter error bound is dependent on the predictive performance of the TWD classifier. This dependency is formalized through the following Theorem, which provides a characterization of the nested set structure for any TWCP:

Theorem 3. *Let $\epsilon \in [0, 1]$ and let $\mathcal{W}_S(x) = A$. Then $A = \Gamma^{\epsilon}_{\mathcal{W}}(S, x)$ iff both:*

1. *$\mathcal{W}$ makes at least $\lfloor \epsilon \cdot (n + 1) \rfloor$ predictions on S with risk greater than $\alpha(|A|)$;*
2. *$\mathcal{W}$ makes at most $\lceil \epsilon \cdot (n + 1) \rceil$ predictions on S with risk greater than $\min_{y \notin A} err(A, y)$.*

[1] Here, non-trivial refers to any TWD classifier that does not err on all of its predictions.

Proof. First, note that $\forall y \in A, l(A,y) = \alpha(|A|)$. Thus, if $y \in A$ is in $\Gamma^{\epsilon}_{\mathcal{W}}(S,x)$, then the same holds for all $y' \in A$. Thus, a sufficient (and necessary) condition for $y \in A$ to be included in $\Gamma^{\epsilon}_{\mathcal{W}}(S,x)$ is the existence of at least $\lfloor \epsilon \cdot (n+1) \rfloor$ instances $x' \in S$ s.t. $l(\mathcal{W}_{S_{x',x}}(x'), y') \geqslant \alpha(|A|)$. Otherwise $\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = \varnothing$.

As for the second condition, note that for any $y \notin A$

$$l(\mathcal{W}_S, \langle x, y \rangle) \geqslant \min_{y' \notin A} l(\mathcal{W}_S, \langle x, y' \rangle) > \alpha(|A|).$$

Thus a sufficient and necessary condition for excluding any $y \notin A$ from $\Gamma^{\epsilon}_{\mathcal{W}}(S,x)$ is that for at most $\lfloor \epsilon \cdot (n+1) \rfloor$ instances $\langle x', y' \rangle \in S$, it holds that

$$l(\mathcal{W}_{S_{x',x}}, \langle x', y' \rangle) \geqslant \min_{y \notin A} l(\mathcal{W}_S, \langle x, y \rangle).$$

Thus, the theorem follows. □

Finally, with respect to our second question, we note that in the uniform-cost classification setting, a finer version of Theorem 3 can be derived. This result shows that any TWD classifier and its corresponding TWCP are equivalent (see also Example 4 for a brief illustration of the following Theorem):

Corollary 2. *Let $\epsilon \in [0,1]$, then in the uniform-cost classification setting it holds that:*

- *If $|\mathcal{W}_S(x)| = 1$, then $\mathcal{W}_S(x) = \Gamma^{\epsilon}_{\mathcal{W}}(S,x)$ iff $\mathcal{W}$ makes at most $\lfloor \epsilon \cdot (n+1) \rfloor$ errors on S (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = Y$);*
- *Otherwise, $\mathcal{W}_S(x) = \Gamma^{\epsilon}_{\mathcal{W}}(S,x)$ iff $\mathcal{W}$ makes at most $\lceil (1-\epsilon) \cdot (n+1) \rceil$ predictions on S with risk lower than $\alpha(|\mathcal{W}_S(x)|)$ (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = \varnothing$) and at most $\lfloor \epsilon \cdot (n+1) \rfloor$ errors (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = Y$).*

Thus, $\Gamma_{\mathcal{W}}(S,x)$ is completely determined by two thresholds $0 \leqslant \epsilon_1 < \epsilon_2 \leqslant 1$ s.t.

$$\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = \begin{cases} \varnothing & \epsilon_2 < \epsilon \\ \mathcal{W}_S(x) & \epsilon_1 < \epsilon \leqslant \epsilon_2 \\ Y & 0 \leqslant \epsilon \leqslant \epsilon_1 \end{cases} \tag{17}$$

Proof. The result follows directly from Theorem 3, applying the result to the case of uniform-cost classification. □

Example 4. Consider the TWCP introduced in Example 3. Then, $|\mathcal{W}_S(x)| > 1$, and $\mathcal{W}$ makes exactly one prediction with risk lower than $\alpha(\mathcal{W}_S(x)) = 0.5$. Hence, by Theorem 2 it holds that $\epsilon_2 = \frac{4}{5}$. Similarly, $\mathcal{W}$ makes exactly 1 error, hence by Theorem 2 it holds that $\epsilon_1 = \frac{1}{5}$.

We now focus on the more general weakly supervised learning setting (i.e. $Z = X \times 2^Y$). Let $S = (\langle x_1, Y_1 \rangle, \ldots, \langle x_n, Y_n \rangle)$ be the current training set. The three-way nonconformity measure can be generalized as follows:

$$M^{\min}_{\mathcal{W}}(S, \langle x, Y_x \rangle) = \min_{y \in Y_x} l(\mathcal{W}_S, \langle x, y \rangle). \tag{18}$$

where $\mathcal{W}$ is a three-way in/three-way out classifier [5]. Thus, the *superset* TWCP is defined as:

$$\Gamma^{\epsilon}_{\mathcal{W},\min}(S,x) = \{y | p^{x,y} > \epsilon\}, \tag{19}$$

$$p^{x,y} = \frac{|\{i = 1, \ldots, n : \text{Pred is verified}\}| + 1}{n+1}, \tag{20}$$

$$\textit{Pred} = l(\mathcal{W}_S, \langle x, y \rangle) \leqslant \min_{y' \in Y_i} l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y' \rangle). \tag{21}$$

The nonconformity measure $M^{\min}_{\mathcal{W}}$ is defined in terms of the *minimum* operator. Thus, it is similar to the optimistic loss minimization [15] approach for weakly supervised learning. Remarkably, however, the role of the minimum operator in the two formulations is different. In the generalized loss minimization framework, the minimum operator selects the instantiation of the set labels that minimizes the empirical loss, over all possible instantiations. On the other hand, in Eq. (19), the minimum operator acts as a conservative bound for the similarity between x and the training set S. Indeed, given $\langle x_i, Y_i \rangle$, the corresponding nonconformity score is

$$\min_{y \in Y_i} M_{\mathcal{W}}(S, \langle x_i, y \rangle) \leqslant M \leqslant \max_{y \in Y_i} M_{\mathcal{W}}(S, \langle x_i, y \rangle).$$

Thus, the nonconformity score of x is compared against the most conservative threshold, among those that are considered possible. In this sense, Eq. (19) is more similar to the principle underlying pessimistic loss minimization [16].

As a second remark, we study the efficiency of the superset TWCP. It is not hard to see that changing min, in Eq. (19), with max or *mean* would equally result in a non-conformity measure. However, it is easy to observe that the approach based on the minimum operator is more efficient than those based on, either, the maximum or mean operators. Indeed, denote these latter non-conformity measures as, resp., $M^{\max}_{\mathcal{W}}, M^{mean}_{\mathcal{W}}$. Similarly, denote the corresponding conformal predictors as, resp., $\Gamma_{\mathcal{W},\max}, \Gamma_{\mathcal{W},mean}$. Then, the following result holds:

Proposition 2. *Let $\mathcal{W}$ be a TWD classifier, S a training set and x a new instance. Then, for any $\epsilon \in [0,1]$, $\Gamma^{\epsilon}_{\mathcal{W},\min}(x) \subseteq \Gamma^{\epsilon}_{\mathcal{W},mean}(x) \subseteq \Gamma^{\epsilon}_{\mathcal{W},\max}(x)$.*

Proof. Note that, for any set of positive numbers $\{n_1, \ldots, n_m\}$, it holds that $\arg\min_i\{n_i\} \leqslant \frac{1}{m}\sum n_i \leqslant \arg\max_i\{n_i\}$. Then, the result easily follows. □

3.2. From conformal prediction to three-way decision

In this section we address the second of the research questions mentioned in Section 1. Namely, we study conditions under which TWD and CP methods are equivalent. To this aim, we first outline two approaches to define a cost-sensitive cautious classifier from any conformal predictor. These latter approaches can be used to transform any conformal predictor into a TWD classifier. We then study the equivalence between TWD and CP methods, by applying the above mentioned approaches to the case in which the conformal predictor is defined as in Section 3.1.

The first approach to obtain a TWD classifier, starting from a CP algorithm Γ_M, relies on the observation that Γ_M is defined as a collection of nested sets, associated with corresponding (lower) probabilities.

Let M be any non-conformity measure, and let Γ_M be the corresponding conformal predictor. Let $A \subseteq Y$ be s.t. $A \in im(\Gamma_M(x))$, i.e. $\exists \epsilon \in [0,1]$ s.t. $\Gamma^{\epsilon}_M(x) = A$. Denote with ϵ^A, the (unique) solution of the following equality:

$$\epsilon^A := \arg\min_{\epsilon \in [0,1]} \{\Gamma^{\epsilon}_M(x) = A\}. \tag{22}$$

Eq. (22) implies that, given $A \in im(\Gamma_M(x))$, it is known that $Pr_{\langle x,y\rangle\sim\mathcal{D}}[y \notin A] \leqslant \epsilon_A$. Therefore, the loss $Loss_{\Gamma_M}(x)$ w.r.t. A can be bounded as follows:

Proposition 3. *Let $A \subseteq Y$ be in the image of $\Gamma_M(x)$, and let ϵ^A be the corresponding solution of Eq. (22). Then:*

$$\alpha(|A|) \leqslant Loss_{\Gamma_M(x)}(A) \leqslant \alpha(|A|) \cdot (1-\epsilon^A) + \epsilon^A |Y \setminus A| \cdot \max_{y \notin A}\{err(A,y)\}. \tag{23}$$

Proof. Let $y \in Y$ be the real label attached to x. Then, by Theorem 1, $Pr[y \notin A] \leqslant \epsilon^A$. Further, by definition of *err* and α, it holds that

$$\alpha(|A|) \leqslant \min_{y \notin A} err(A,y) \leqslant \max_{y \notin A}\{err(A,y)\}.$$

Note, also, that the rightmost summand in Eq. (23) is monotonically increasing w.r.t. $\epsilon \in [0, \epsilon^A]$, and the left and right side of the inequality chain coincide when $\epsilon = 0$. Then, the result easily follows. □

Denote the right-most side of Eq. (23) as $Loss^*_{\Gamma_M(x)}(A)$. Then, given a conformal predictor Γ^{ϵ}_M, the *decision-theoretic conformal TWD* (DCTWD) classifier $\mathcal{W}_\Gamma$ is defined as follows:

$$\mathcal{W}^{dec}_\Gamma(x) = \arg\min_{A \in im(\Gamma_M(x))} Loss^*_{\Gamma_M(x)}(A). \tag{24}$$

On the other hand, the second approach to transform a conformal predictor into a TWD classifier relies on the observation that a conformal predictor Γ_M defines a *possibility distribution* over Y. Indeed, given $A \in im(\Gamma_M)$, it holds that $1 - \epsilon_A$ is a lower bound on the probability that the correct label y is in A.

Denote as $\varnothing \subset A_1 \subseteq \ldots \subseteq A_k$ the nested sets in $im(\Gamma_M)$. Given any $y \in Y$, let $j(y) = \max\{i : y \notin A_i\}$. Then, a possibility distribution π_Γ can be defined as follows [8]:

$$\pi_\Gamma(y) = \begin{cases} 1 & A_1 = \{y\} \\ \epsilon^{A_{j(y)}} & otherwise \end{cases} \tag{25}$$

The possibility distribution π_Γ can then be used to define a TWD classifier by transforming π_Γ into a probability distribution, so that the Loss function in Eq. (5) is well-defined. This transformation is performed by means of the possibility-probability transformation [9]:

$$Pr_{\pi_\Gamma}(y) = \sum_{i=1}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} \mathbf{1}_{y \in B_i}, \tag{26}$$

where $\hat{\pi}$ is the ordering of π_Γ in terms of decreasing possibility value; B_i is the $\hat{\pi}_i$ α-cut (i.e., $B_i = \{y \in Y : \pi_\Gamma(y) \geqslant \hat{\pi}_i\}$). Then, the *possibilistic conformal three-way*(PCTWD) classifier is defined as:

$$\mathcal{W}^{poss}_{\Gamma}(x) = Loss^{poss}_{\Gamma_M(x)}(A) = \arg\min_{A \in 2^Y} \sum_{y \notin A} Pr_{\pi_\Gamma}(y) \cdot err(A, y) + \alpha(|A|) \sum_{y \in A} Pr_{\pi_\Gamma}(y). \tag{27}$$

Example 5 below provides an illustration of the calculations involved in the construction of a DCTWD and a PCTWD.

Example 5. Consider the TWCP $\Gamma_{\mathcal{W}}$ and loss function defined in Example 3.

Then, considering the instance x, it holds that $Loss^{*}_{\Gamma_{\mathcal{W}}(x)}(\{0,1\}) = \frac{4}{5} * \frac{1}{3} + \frac{1}{5} = 0.47$; while $Loss^{*}_{\Gamma_{\mathcal{W}}(x)}(Y) = 0.67$. Hence $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}(x) = \{0,1\}$.

By contrast, the corresponding PCTWD can be defined by noting that $\pi_\Gamma(0) = \pi_\Gamma(1) = 1$ and $\pi_\Gamma(2) = 0.25$, therefore $Pr_{\pi_\Gamma} = \langle 0 : 0.46, 1 : 0.46, 2 : 0.08 \rangle$.

Hence, $Loss^{poss}_{\Gamma}(0) = 0.54, Loss^{poss}_{\Gamma}(1) = 0.54$, while $Loss^{poss}_{\Gamma}(\{0,1\}) = 0.39$ and $Loss^{poss}_{\Gamma}(Y) = 0.67$. Therefore $\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}}}(x) = \{0,1\}$.

The two above mentioned constructions allow to transform any conformal predictor into a cost-sensitive TWD classifier. Furthermore, it is easy to see that these constructions preserve validity. Indeed, for the case of a DCTWD $\mathcal{W}^{dec}_{\Gamma}$, the following result holds:

Proposition 4. *$Pr[y_x \notin \mathcal{W}^{dec}_{\Gamma}(x)] \leqslant \epsilon^A$, where ϵ^A is defined as in Eq. (22).*

Proof. By construction, it holds that $\mathcal{W}^{dec}_{\Gamma}(x) = A \in im(\Gamma(x))$. Then, the result follows by Theorem 1, and the definition of ϵ^A. □

By contrast, for the case of a PCTWD $\mathcal{W}^{poss}_{\Gamma}$, there is no guarantee that $\mathcal{W}^{poss}_{\Gamma}(x) \in im(\Gamma(x))$. Nonetheless, a weaker bound can be obtained through the following result:

Proposition 5. *$Pr[y_x \notin \mathcal{W}^{poss}_{\Gamma}(x)] \leqslant \epsilon^{B^*}$, where*

$$B^* = \arg\max_{B \in im(\Gamma)(x) : B \subseteq \mathcal{W}^{poss}_{\Gamma}(x)} |B|. \tag{28}$$

Proof. The result directly follows from the definition of Γ and $\mathcal{W}^{poss}_{\Gamma}(x)$. In particular, for all $B \subseteq B^*$ it holds that $Pr[y_x \notin \mathcal{W}^{poss}_{\Gamma}(x)] \leqslant \epsilon^B$. □

We now consider our main research question: namely, we ask under which conditions a given TWD classifier, and the corresponding DCTWD (resp. PCTWD) classifier, are equivalent. Such conditions would then establish an isomorphism between the class of TWD classifiers and (three-way) conformal predictors.

To this aim, let $\mathcal{W}$ be a TWD classifier, let $\Gamma_{\mathcal{W}}$ be the TWCP obtained from $\mathcal{W}$ and, finally, let $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}$ (resp. $\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}}}$) be the corresponding DCTWD (resp. PCTWD). The following result provides sufficient and necessary conditions for the equivalence between the TWD classifier $\mathcal{W}$ and the DCTWD classifier $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}$.

Theorem 4. *Let $\mathcal{W}(x) = A$, then $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}(x) = A$ holds iff the following two conditions are satisfied:*

1. *$\exists \epsilon \in [0,1]$ s.t. $\mathcal{W}$ makes at least $\lfloor \epsilon \cdot (n+1) \rfloor$ predictions on S with risk greater than $\alpha(|A|)$ and makes at most $\lfloor \epsilon \cdot (n+1) \rfloor$ predictions on S with risk greater than $\min_{y \notin A} R(y, A)$;*
2. *$\epsilon^A \leqslant \min_{B \in im(\Gamma^{dec}_{\mathcal{W}}(x))} \frac{Loss^{*}_{\Gamma_{\mathcal{W}}}(B) - \alpha(|A|)}{\max_{y - in A} err(A,y) - \alpha(|A|)}$.*

Proof. The first condition, by Theorem 3, ensures that $\mathcal{W}(x) \in im(\Gamma_{\mathcal{W}}(x))$. The second condition, on the other hand, ensures that the transformation preserves the minimal element w.r.t. the ordering of 2^Y in terms of the *Loss* value. Thus, if both conditions hold, then $\mathcal{W}(x)$ is the unique solution to Eq. (24), and the result follows. □

The following corollary shows that, in the uniform-cost setting, the conditions required by Theorem 4 can be relaxed:

Corollary 3. *Let S be the current training set with $|S| = n, \mathcal{W}_S(x) = A$, then $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}(x) = A$ iff*

$$m \leqslant \frac{\alpha(|Y|) - \alpha(|A|)}{err - \alpha(|A|)} \cdot n, \tag{29}$$

where m is the number of errors made by $\mathcal{W}_S$.

Proof. The result directly follows from Theorems 2 and 4. □

We now discuss the case of the PCTWD classifier. First of all, irrespective of the non-conformity measure used, $\forall A \in im(\Gamma)$, the following proposition holds:

Proposition 6. $Pr_{\pi_\Gamma}(A) \geqslant 1 - \epsilon^A$.

Proof. Let j be s.t. $B_j = A$ (i.e. A is the i^{th} α-cut). Then:

$$\begin{aligned} Pr_{\pi_\Gamma}(A) &= (1 - \epsilon^{B_1}) + (\epsilon^{B_1} - \epsilon^{B_2}) + \ldots + (\epsilon^{B_{j-1}} - \epsilon^{A}) + |A|\sum_{i=j}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} = (1 - \epsilon^A) + |A|\sum_{i=j}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} \\ &\geqslant (1 - \epsilon^A) + \frac{|A|}{|Y|}\epsilon^A > 1 - \epsilon^A \end{aligned} \tag{30}$$

□

This result implies, in particular, that $Loss^{poss}_{\Gamma}(A) \leqslant Loss^{*}_{\Gamma}(A)$.

Furthermore, note that if $\Gamma = \Gamma_{\mathcal{W}}$ (i.e. Γ is a TWCP) and the cost function is uniform, then the penultimate inequality in Eq. (30) holds with equality (as a consequence of Theorem 2). Then, the following result provides sufficient and necessary conditions for the equivalence between a TWD classifier and the corresponding PCTWD classifier in the uniform-cost setting:

Theorem 5. *Let S be the current training set. Let $\mathcal{W}_S(x) = A$. Then, $\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}}}(x) = A$ iff all the following conditions hold:*

$$\frac{\alpha(|A|+1)}{\alpha(|A|)} > f(|A|), \tag{31}$$

$$\forall k < |A|,\ 1 - \epsilon^A > g(|A|, k), \tag{32}$$

$$\forall k > |A|,\ \epsilon^A \leqslant g(|A|, k), \tag{33}$$

where

$$\begin{aligned} f(|A|) &= \frac{1 - \frac{|A|}{|Y|}}{(|A|+1)(\frac{1}{|A|} - \frac{1}{|Y|})}, \\ g(|A|, k) &= \frac{\frac{k}{|A|}\alpha(k) - \alpha(|A|)}{D(|A|, k)}, \\ D(|A|, k) &= \frac{k}{|A|}\alpha(k) + \frac{k}{|Y|} + \frac{|A|}{|Y|}\alpha(|A|) \\ &\quad -\alpha(|A|) - \frac{k}{|Y|}\alpha(k) - \frac{|A|}{|Y|}. \end{aligned}$$

and ϵ^A is equal to ϵ_1 in Theorem 2.

Proof. The result directly follows from standard algebraic manipulations and the observation that, in the uniform-cost setting, the penultimate inequality in (30) holds with equality. □

The generalization of Theorem 5 to general, non-uniform, loss functions is left as an open problem.

In regard to the significance of Theorems 4 and 5, we discussed in Section 2.2 that, while TWD is optimal w.r.t. cost-sensitiveness, its results may in general be not valid. In particular, the latter may happen when the underlying classifier is not calibrated. Therefore, the transformation from a TWD classifier to a CP one (by means of TWCP and then, either, a DCTWD, or a PCTWD, classifier), can be seen as an approach to correct this lack of validity. In particular, then, Theorems 4 and 5 show that calibration is not a necessary condition for a TWD classifier to be valid, and provide conditions for validity.

Indeed, the two Theorems show that, under the condition that the TWD classifier $\mathcal{W}$ is sufficiently accurate, the correction implemented by means of CP has no effect. In this latter case, the set-valued predictions obtained before and after the validity correction are identical. Consequently, Theorems 4 and 5 establish an isomorphism between the class of (non-trivial) TWD classifiers and (three-way) conformal predictor. An illustration of these latter observations is shown in Example 6 and in Fig. 2.

Example 6. Let us refer to the TWCP $\Gamma_{\mathcal{W}}$ defined in Example 3 and the corresponding DCTWD and PCTWD classifiers defined in Example 5. In Example 5, it was shown that the predictions provided by the three TWD classifiers were equivalent, hence Theorems 4 and 5 should hold, as they provide sufficient and necessary conditions for such equivalences.

Table 1
List of used datasets.

Dataset	Instances	Features	Classes
Digits	1797	64	10
Breast Cancer	569	30	2
Wine	178	13	3
Covertype	581012	54	7
20Newsgroups	18846	130107	20
Diabetes	786	8	2
Epileptic Seizure	11500	179	2
Diabetic Retinopathy	1151	20	4
Hepatitis C virus	1385	29	4
Chronic Kidney Disease	400	25	2
Abalone	4177	8	27
Arrhythmia	452	279	16

Indeed, as regards the DCTWD, we note that $\mathcal{W}$ made exactly $1 < \frac{2/3-1/3}{1-1/3} \cdot |S| = 2$ error and thus the conditions in Theorem 4 are satisfied.

Similarly, with respect to the PCTWD, we note that Eq. (31) reduces to $2 > \frac{1-2/3}{3(1/2-1/3)} = 0.67$, Eq. (32) reduces to $\frac{4}{5} > 0.5$ and Eq. (33) reduces to $\frac{1}{5} < \frac{2}{5}$ which are all obviously satisfied.

4. Results

4.1. Experimental design

The theoretical study of the previous sections shows some important connections between TWD and CP. In particular, it provides conditions for the equivalence among these two cautious classification methods. Based on these results, in this section, we describe a set of experiments to investigate the relationship between TWD and CP also from an empirical point of view.

More in detail, we address three research questions:

1. In Sections 3.1 and 3.2, we studied conditions for the equivalence between a TWD classifier $\mathcal{W}$ and the corresponding DCTWD (resp., PCTWD) classifier. In particular, we showed that these latter two classes of classifiers are equivalent, provided that the original TWD classifier is sufficiently accurate. Do these conditions hold in real-world datasets?
2. In Section 3.2 we proposed the DCTWD and PCTWD classifiers as techniques to obtain cost-sensitive cautious classifiers, starting from any conformal predictor. Nonetheless, we did not study any difference, in terms of validity or efficiency, between the DCTWD and PCTWD construction. Are there any empirical differences among these two latter methods, in terms of either classification accuracy or efficiency?
3. The proposed constructions can be seen as techniques to both improve the predictive performance of a TWD, as well as objective[2] approaches to obtain a cautious classifier from any CP method. Do these constructions result in an increase in predictive performance compared with other state-of-the-art TWD and CP algorithms?

To this end, we considered a set of experiments, based on 12 datasets from the UCI repository. These datasets are listed in Table 1.

We considered two different classes of scoring classifiers, namely Random Forest and k-Nearest Neighbors. For each of these latter two classes, we compared the results of 6 different methods:

- The (standard, single-valued prediction) classifiers h_{RF}, h_{KNN};
- The TWD classifiers $\mathcal{W}_{RF}, \mathcal{W}_{KNN}$;
- The TWCP-based DCTWD and PCTWD classifiers $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}, \mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{KNN}}}$ based on $\mathcal{W}_{RF}, \mathcal{W}_{KNN}$;
- The DTCWD and PCTWD classifier $\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}, \mathcal{W}^{dec}_{\Gamma_{h_{KNN}}}$, directly based on h_{RF}, h_{KNN} (see Section 2.3).

The loss function (used to determine the TWD classifiers and to evaluate the performance of the models) was defined through the following abstention cost function:

$$\alpha(n) = \frac{n-1}{|Y|}, \tag{34}$$

[2] Here, by objective it is meant that there is no a priori selection of a probability threshold.

while the *err* function was

- Uniform, with all costs equal to 1, for the Abalone, Digits, Wine, Covertype and 20Newsgroups datasets;
- Equal to 1 when the true class label was associated to healthy status, and equal to 2 otherwise, for the medical datasets.

The CP algorithms were implemented using the inductive approach, i.e., by relying on a validation set. The size of the validation set was set to 20% of the training set. We decided to use the inductive approach, rather than the sequential one, in order to reduce the computational cost of re-training the classification algorithms.

All algorithms were evaluated in terms of the complement of the above mentioned loss function,[3] henceforth *accuracy*, as well as in terms of coverage, as a measure of efficiency. This latter measure, in particular, was defined, for a set-valued prediction S, as:

$$coverage(S) = 1 - \frac{|S| - 1}{|Y| - 1}. \tag{35}$$

All performances were computed using 5-fold cross-validation. Thus, we report the results in terms of both the average performance and the corresponding 95% confidence interval. In order to assess the presence of statistically significant differences, if any, we performed the Friedman rank test. Namely, for each cautious learning approach and each model class (Random Forest, kNN), we computed the ranks with respect to each of the considered datasets; for each cautious learning approach and each dataset we then averaged the Random Forest and kNN ranks.

4.2. Experimental results

The results of the Experiments, in terms of the average accuracy, are reported in Tables 2, 3 and in Fig. 3. The average coverage values are reported in Tables 4, 5 and Fig. 4.

As regards the observed accuracies, the average ranks are reported in Table 6: the observed test statistic was $Q = 36.11$, which was significant at the 95% confidence level (p-value < 0.00001). Thus, we also performed a post hoc pairwise comparison using the Nemenyi test procedure. The critical value of the test (with 12 datasets and 6 compared methods), at the 95% confidence level, is 2.176. The pairwise comparisons are reported in Table 7.

As regards the observed coverage values, the average ranks are reported in Table 8. The observed test statistic was $Q = 26.33$ which was significant at the 95% confidence level (p-value $= 0.00003$). Thus, we performed the Nemenyi post hoc pairwise test. The critical value of the test (with 12 dataset 5 compared methods), at the 95% confidence level, is 1.761. The pairwise comparison are reported in Table 9.

4.3. Discussion

Commenting the results reported in Section 4.2, we can see that the TWD classifiers (i.e., $\mathcal{W}, \Gamma^{dec}_{\mathcal{W}}$ and $\Gamma^{poss}_{\mathcal{W}}$) outperformed the corresponding single-valued classifiers in terms of accuracy. This finding should not be surprising. Indeed, the considered cautious classifiers are, by construction, *cost-sensitive*. Hence, they always return the set-valued prediction that maximizes the accuracy. Nonetheless, it shows that both TWD classifiers and the corresponding CP-based corrections can be useful to obtain significantly improved predictive performance (if set-valued predictions are allowed).

More interestingly, we can observe that the standard CP-based classifiers (i.e., Γ^{dec}_{h} and Γ^{poss}_{h}) were not significantly different from the single-valued classifiers in terms of accuracy. In particular, the PCTWD classifier was significantly outperformed by all TWD-based classifiers. Thus, we can provide a positive answer to our third experimental research question. Indeed, the proposed methods out-performed the state-of-the-art CP methods, in terms of predictive accuracy, with comparable or even better efficiency.

In regard to our first research question, we can see that there were no significant differences among the three TWD-based cautious classifiers (i.e., $\mathcal{W}, \Gamma^{dec}_{\mathcal{W}}$ and $\Gamma^{poss}_{\mathcal{W}}$). This finding lends empirical support to the results proven in Section 3. Indeed, in Section 3, we proved that a TWD classifier is equivalent to the corresponding CP-based model, provided the original TWD classifier is sufficiently accurate. The experimental results, then, show that the conditions of Theorems 3, 4, 4 are usually satisfied in real-world setting. Hence, TWD methods can usually be expected to have the same level of validity as the corresponding CP-based correction.

More in detail, in the case of Random Forest, the results for the TWD-based cautious classifiers were almost always identical. By contrast, in the case of kNN, there were 4 datasets on which the DCTWD and PCTWD classifiers achieved increased performance. This observation can be explained by noting that kNN classifiers usually have lower accuracy and generalization capability than Random Forest ones. As a consequence of the results in Section 3, this observation implies that TWD classifiers based on kNN are expected to satisfy the conditions of Theorems 3–5 less often than classifiers based on Random Forest. Therefore, the proposed TWCP, DCTWD and PCTWD constructions would result in less efficient (but more accurate)

[3] Note that when the *err* function was uniform, then the complement of the loss function is equivalent to a penalized accuracy, in which the penalization depends on the size of the set-valued prediction.

Table 2
Average loss value and 95% confidence intervals for the Random Forest-based classifiers, on all 12 datasets.

Dataset	h_{RF}	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma^{\mathcal{W}_{RF}}}$	$\mathcal{W}^{dec}_{\Gamma^{h_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma^{h_{RF}}}$
Abalone	0.79 ± 0.02	0.91 ± 0.01	$\mathbf{0.92 \pm 0.01}$	0.91 ± 0.01	0.90 ± 0.01	0.85 ± 0.03
Arrhythmia	0.80 ± 0.04	0.81 ± 0.02	0.81 ± 0.02	$\mathbf{0.84 \pm 0.03}$	0.80 ± 0.04	0.81 ± 0.04
Breast Cancer	0.86 ± 0.03	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	0.95 ± 0.01	0.96 ± 0.01
Chronic Kidney	0.85 ± 0.03	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.87 \pm 0.02}$	0.85 ± 0.03	0.85 ± 0.03
Covertype	0.82 ± 0.02	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	0.85 ± 0.01	0.88 ± 0.03
Diabetes	0.73 ± 0.05	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.77 \pm 0.02}$	0.74 ± 0.02	0.75 ± 0.01	0.73 ± 0.05
Diabetic Retinopathy	0.78 ± 0.04	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.84 \pm 0.02}$	0.81 ± 0.02	0.82 ± 0.02	0.78 ± 0.04
Digits	0.94 ± 0.02	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$	0.91 ± 0.01	0.90 ± 0.01
Epileptic Seizure	0.73 ± 0.01	$\mathbf{0.88 \pm 0.00}$	$\mathbf{0.88 \pm 0.00}$	0.77 ± 0.05	0.86 ± 0.00	0.70 ± 0.01
Hepatitis C	0.56 ± 0.03	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.86 \pm 0.04}$	0.79 ± 0.03	0.77 ± 0.02
Wine	0.81 ± 0.05	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.96 \pm 0.02}$	0.92 ± 0.02	0.91 ± 0.03
20Newsgroups	0.85 ± 0.01	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.01}$	0.81 ± 0.01

Table 3
Average loss value and 95% confidence intervals for the kNN-based classifiers, on all 12 datasets.

Dataset	h_{KNN}	$\mathcal{W}_{KNN}$	$\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{KNN}}}$	$\mathcal{W}^{poss}_{\Gamma^{\mathcal{W}_{KNN}}}$	$\mathcal{W}^{dec}_{\Gamma^{h_{KNN}}}$	$\mathcal{W}^{poss}_{\Gamma^{h_{KNN}}}$
Abalone	0.67 ± 0.02	0.74 ± 0.02	$\mathbf{0.91 \pm 0.00}$	0.81 ± 0.04	0.88 ± 0.02	0.73 ± 0.04
Arrhythmia	0.69 ± 0.08	0.75 ± 0.07	0.75 ± 0.07	0.75 ± 0.07	$\mathbf{0.77 \pm 0.05}$	0.69 ± 0.08
Breast Cancer	0.93 ± 0.02	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	0.73 ± 0.06	0.93 ± 0.01
Chronic Kidney	0.74 ± 0.04	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	0.74 ± 0.04
Covertype	0.71 ± 0.02	$\mathbf{0.87 \pm 0.06}$	$\mathbf{0.87 \pm 0.06}$	0.78 ± 0.01	0.50 ± 0.00	0.71 ± 0.02
Diabetes	0.68 ± 0.03	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.72 \pm 0.03}$	0.70 ± 0.03	0.70 ± 0.03	0.68 ± 0.03
Diabetic Retinopathy	0.63 ± 0.03	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.79 \pm 0.02}$	0.78 ± 0.02	0.71 ± 0.02	0.73 ± 0.03
Digits	0.96 ± 0.01	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	0.96 ± 0.01	0.90 ± 0.00	0.96 ± 0.01
Epileptic Seizure	0.59 ± 0.01	0.74 ± 0.01	$\mathbf{0.83 \pm 0.00}$	0.78 ± 0.01	0.75 ± 0.00	0.66 ± 0.01
Hepatitis C	0.56 ± 0.03	0.79 ± 0.01	$\mathbf{0.83 \pm 0.01}$	0.79 ± 0.04	$\mathbf{0.83 \pm 0.01}$	0.70 ± 0.02
Wine	0.71 ± 0.04	$\mathbf{0.85 \pm 0.02}$	$\mathbf{0.85 \pm 0.02}$	0.78 ± 0.06	0.78 ± 0.03	0.71 ± 0.04
20Newsgroups	0.67 ± 0.01	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	0.50 ± 0.00	0.76 ± 0.01

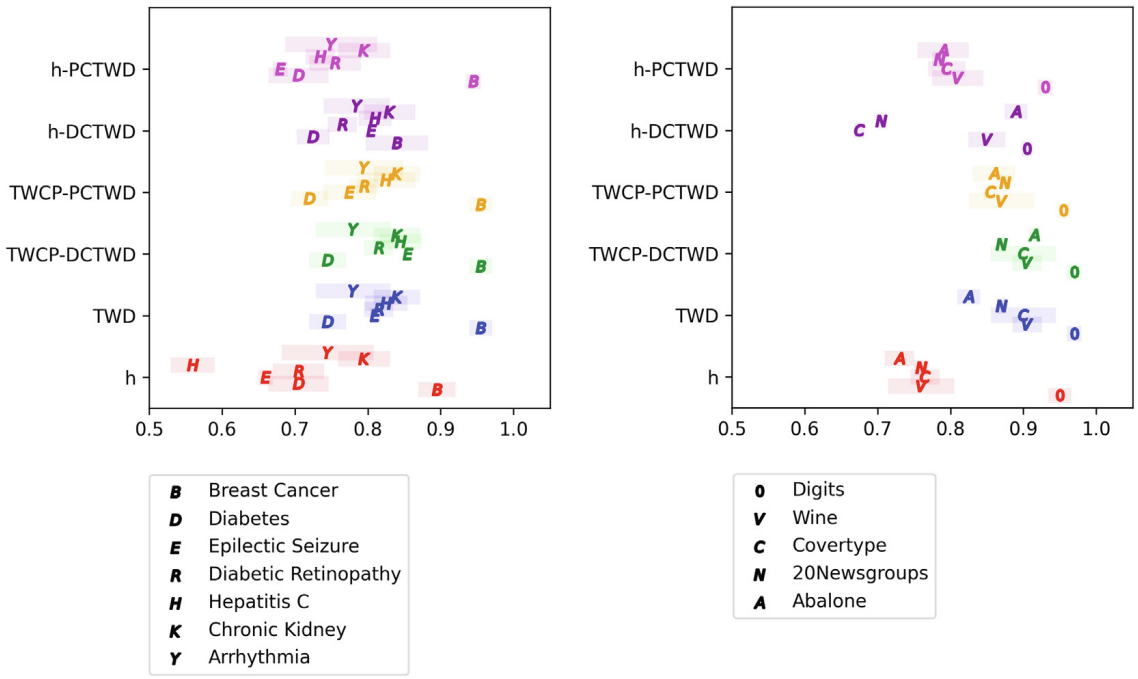


Fig. 3. Average accuracy and 95% confidence intervals for each of the evaluated classifiers, on the medical (left) and non-medical (right) datasets. Each marker refers to the average of the corresponding Random Forest-based and kNN-based classifiers. In the legend, h denotes the standard classifiers (h_{RF}, h_{KNN}); TWD the three-way decision classifiers ($\mathcal{W}_{RF}, \mathcal{W}_{KNN}$); TWCP-DCTWD (resp., TWCP-PCTWD) the DCTWD (resp., PCTWD) classifier based on the TWCP ($\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{RF}}}, \mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{RF}}}$); while h-DCTWD (resp., h-PCTWD) the DCTWD (resp. PCTWD) classifier based on the standard conformal predictors ($\mathcal{W}^{dec}_{\Gamma^{h_{RF}}}, \mathcal{W}^{dec}_{\Gamma^{h_{RF}}}$).

Table 4
Average coverage value and 95% confidence intervals for the Random Forest-based classifiers, on all 12 datasets.

Dataset	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}}_{RF}}$	$\mathcal{W}^{poss}_{\Gamma^{\mathcal{W}}_{RF}}$	$\mathcal{W}^{dec}_{\Gamma^{h}_{RF}}$	$\mathcal{W}^{poss}_{\Gamma^{h}_{RF}}$
Abalone	$\mathbf{0.74 \pm 0.01}$	0.64 ± 0.05	0.69 ± 0.01	0.60 ± 0.02	0.69 ± 0.05
Arrhythmia	0.86 ± 0.02	0.86 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.82 ± 0.09	$\mathbf{1.00 \pm 0.00}$
Breast Cancer	0.97 ± 0.02	0.97 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.54 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Chronic Kidney	0.83 ± 0.05	0.83 ± 0.05	$\mathbf{1.00 \pm 0.00}$	0.75 ± 0.06	$\mathbf{1.00 \pm 0.00}$
Covertype	$\mathbf{0.76 \pm 0.03}$	$\mathbf{0.76 \pm 0.02}$	$\mathbf{0.76 \pm 0.03}$	0.30 ± 0.05	0.62 ± 0.16
Diabetes	0.94 ± 0.04	0.94 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.81 ± 0.05	$\mathbf{1.00 \pm 0.00}$
Diabetic Retinopathy	0.73 ± 0.02	0.73 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.74 ± 0.01	$\mathbf{1.00 \pm 0.00}$
Digits	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.93 \pm 0.03}$	0.60 ± 0.02	0.65 ± 0.05
Epileptic Seizure	0.58 ± 0.01	0.58 ± 0.00	0.77 ± 0.09	0.47 ± 0.01	$\mathbf{0.81 \pm 0.01}$
Hepatitis C	0.44 ± 0.01	0.44 ± 0.01	0.44 ± 0.01	0.25 ± 0.06	$\mathbf{0.64 \pm 0.04}$
Wine	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.87 \pm 0.05}$	0.44 ± 0.16	$\mathbf{0.87 \pm 0.05}$
20Newsgroups	$\mathbf{0.64 \pm 0.01}$	$\mathbf{0.64 \pm 0.01}$	$\mathbf{0.64 \pm 0.01}$	0.57 ± 0.00	0.59 ± 0.01

Table 5
Average coverage value and 95% confidence intervals for the kNN-based classifiers, on all 12 datasets.

Dataset	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}}_{RF}}$	$\mathcal{W}^{poss}_{\Gamma^{\mathcal{W}}_{RF}}$	$\mathcal{W}^{dec}_{\Gamma^{h}_{RF}}$	$\mathcal{W}^{poss}_{\Gamma^{h}_{RF}}$
Abalone	$\mathbf{0.77 \pm 0.00}$	0.60 ± 0.03	0.69 ± 0.10	0.57 ± 0.03	$\mathbf{0.77 \pm 0.07}$
Arrhythmia	0.79 ± 0.04	0.79 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.55 ± 0.17	$\mathbf{1.00 \pm 0.00}$
Breast Cancer	0.96 ± 0.02	0.96 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.50 ± 0.12	$\mathbf{1.00 \pm 0.00}$
Chronic Kidney	0.79 ± 0.06	0.79 ± 0.06	$\mathbf{1.00 \pm 0.00}$	0.62 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Covertype	0.94 ± 0.02	0.94 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$
Diabetes	0.82 ± 0.04	0.82 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.72 ± 0.07	$\mathbf{1.00 \pm 0.00}$
Diabetic Retinopathy	0.75 ± 0.04	0.75 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.59 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Digits	0.98 ± 0.01	0.98 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.55 ± 0.00	$\mathbf{1.00 \pm 0.00}$
Epileptic Seizure	0.71 ± 0.00	0.50 ± 0.00	0.62 ± 0.01	0.68 ± 0.01	$\mathbf{0.78 \pm 0.01}$
Hepatitis C	0.32 ± 0.01	0.23 ± 0.04	0.39 ± 0.07	0.37 ± 0.04	$\mathbf{0.79 \pm 0.01}$
Wine	0.76 ± 0.04	0.71 ± 0.09	0.83 ± 0.11	0.62 ± 0.09	$\mathbf{0.97 \pm 0.01}$
20Newsgroups	0.83 ± 0.01	0.83 ± 0.01	$\mathbf{0.87 \pm 0.01}$	0.00 ± 0.00	0.82 ± 0.01

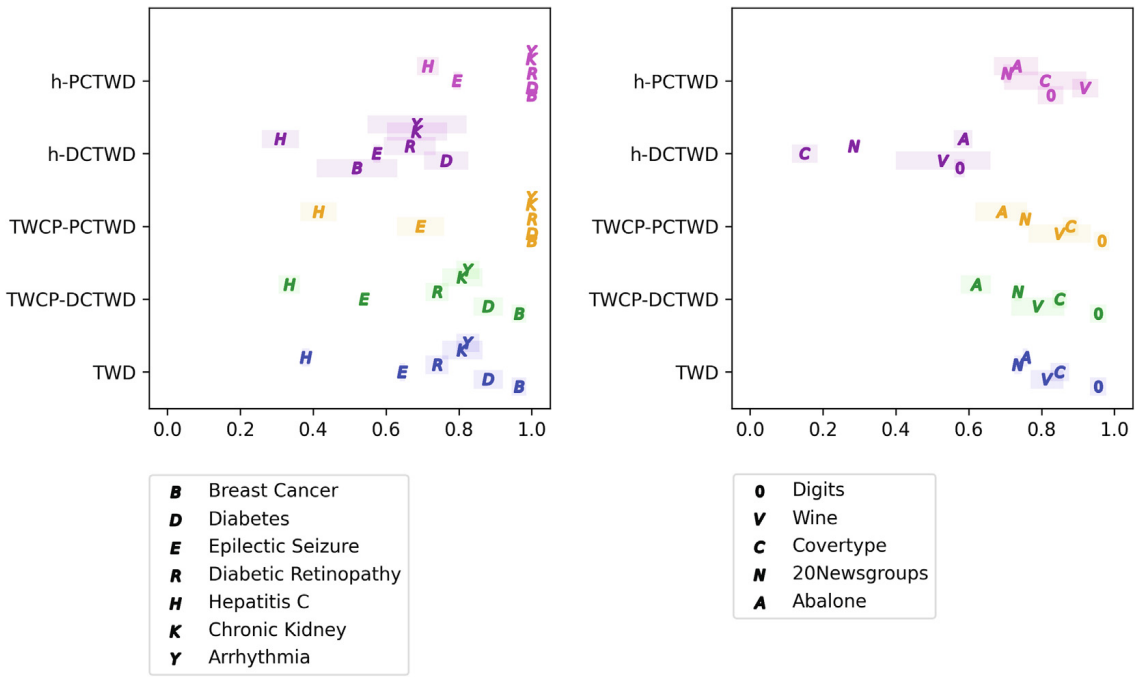


Fig. 4. Average accuracy and 95% confidence intervals for each of the evaluated classifiers, on the medical (left) and non-medical (right) datasets. Each marker refers to the average of the corresponding Random Forest-based and kNN-based classifiers. In the legend, h denotes the standard classifiers (h_{RF}, h_{KNN}); TWD the three-way decision classifiers ($\mathcal{W}_{RF}, \mathcal{W}_{KNN}$); TWCP-DCTWD (resp., TWCP-PCTWD) the DCTWD (resp., PCTWD) classifier based on the TWCP ($\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}}_{RF}}, \mathcal{W}^{dec}_{\Gamma^{\mathcal{W}}_{RF}}$); while h-DCTWD (resp., h-PCTWD) the DCTWD (resp. PCTWD) classifier based on the standard conformal predictors ($\mathcal{W}^{dec}_{\Gamma^{h}_{RF}}, \mathcal{W}^{dec}_{\Gamma^{h}_{RF}}$).

Table 6
Average ranks of the compared learning algorithms, in terms of observed loss value, according to the Friedman test procedure.

	h	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
Average rank	5.29	2.23	1.83	2.67	4.01	4.91

Table 7
Pairwise differences in ranks, in terms of observed loss values, among the compared learning algorithms. Statistically significant differences (according to the Nemenyi test) are denoted in bold and with an asterisk.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
h	3.06*	3.46*	2.62*	1.28	0.38
$\mathcal{W}$	–	0.40	0.44	1.78	2.68*
$\Gamma^{dec}_{\mathcal{W}}$	–	–	0.84	2.18*	3.08*
$\Gamma^{poss}_{\mathcal{W}}$	–	–	–	1.34	2.24*
Γ^{dec}_{h}	–	–	–	–	0.9

Table 8
Average ranks of the compared learning algorithms, in terms of observed coverage, according to the Friedman test procedure.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
Average rank	3.02	3.46	2	4.75	1.90

Table 9
Pairwise differences in ranks, in terms of observed coverage values, among the compared learning algorithms. Statistically significant differences (according to the Nemenyi test) are denoted in bold and with an asterisk.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
$\mathcal{W}$	–	0.44	1.02	1.73	1.12
$\Gamma^{dec}_{\mathcal{W}}$	–	–	1.46	1.29	1.36
$\Gamma^{poss}_{\mathcal{W}}$	–	–	–	**2.75***	0.10
Γ^{dec}_{h}	–	–	–	–	**2.85***

predictions, as observed in Tables 3, 5. More generally, the proposed construction can be applied to improve the predictive accuracy of any TWD classifier whose underlying single-valued ML models may be prone to either under- or over-fitting, as in the case of kNN.

In regard to our second research question, we did not find significant differences among the DCTWD classifiers and the PCTWD classifiers in terms of predictive accuracy, though the PCTWD classifiers were on average less accurate than the DCTWD ones. On the other hand, in terms of efficiency, the PCTWD classifiers reported a larger coverage than the DCTWD ones. In particular, the difference between Γ^{dec}_{h} and Γ^{poss}_{h} was statistically significant.

Thus, the DCTWD and PCTWD classifiers offer a trade-off between greater accuracy (for the DCTWD classifier) and greater efficiency (for the PCTWD classifier). The selection among the two methods should then be made by the decision-maker, based on the quality dimension which is deemed most important for the specific decision-making task at hand.

As a general final remark, we focus on the TWD-based classifiers, i.e., the TWD classifier $\mathcal{W}$, the DCTWD classifier $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}}}$ and the PCTWD classifier $\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}}}$. Compared with $\mathcal{W}$, the DCTWD classifier reported, on average, improved predictive accuracy but slightly lower efficiency. By contrast, the PCTWD reported, on average, improved efficiency with comparable but slightly reduced accuracy. Therefore, the application of the proposed CP-based corrections could be useful not only for classifiers whose predictions are insufficiently accurate, or for classifiers that are known to be prone to over-fitting, but also for more general TWD classifiers. Indeed, in the worst case situation, the set-valued predictions provided by TWD and the corresponding DCTWD and PCTWD would be equivalent, as a consequence of the results in Section 3. In all other cases, however, the proposed constructions would allow to achieve either more accurate (using the DCTWD classifier) or more specific (using the PCTWD classifier) predictions, compared with a standard TWD classifier.

5. A medical case study

Up to now, we have discussed the relationship between TWD and CP. Through this relationship we studied some formal property of TWD, by introducing the TWCP, DCTWD and PCTWD classifiers as a means to both study validity bounds for TWD

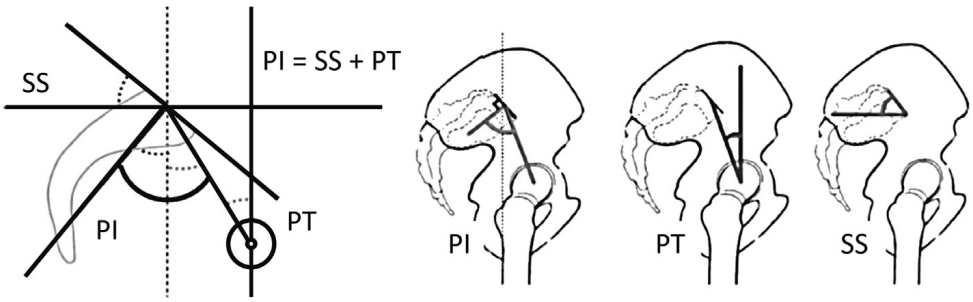


Fig. 5. The main angles considered in the sagittal imbalance classification.

and to improve the validity of standard TWD classifiers. In this section, we address the potential of the proposed approach for human decision making in classification tasks and, therefore, for its integration into Decision Support Systems.

As we argued in the Introduction, cautious learning approaches could be useful to develop valid and reliable decision support in human decision making. Nonetheless, to the knowledge of the authors, no previous study evaluated the usefulness of such set-valued advice compared to standard support. Indeed, even though the recent study by Liu et al. [23] assessed the effectiveness of TWD from the perspective of interpretability, the authors did not specifically evaluate the usefulness of set-valued advice.

In order to understand whether set-valued advice could be supportive in naturalistic decision making [19], we tested this approach in the case of the assessment of sagittal misalignment. This latter is a kind of spine deformity regarding an imbalance along the front-to-back direction of the outward curve of the middle spine called kyphosis.

We chose this case for three main reasons. First, there is a lack of standard criteria to classify imbalance [20], as this is characterized in terms of a number of angles, among which the main ones are called pelvic tilt (PT), sacral slope (SS) and pelvic incidence (PI, which can be defined as the sum of PT and SS - see Fig. 5). Second, it fits well a set-valued output. Indeed, real cases form a continuous range, where specific instances of pathological shape of the spine might be borderline, sharing characteristics of two "adjacent" patterns. On the other hand, existing classification schema provide discrete and mutually exclusive categories by which to characterize spine misalignment. Lastly, and more importantly, the diagnosis of this kind of spine deformity is strongly related to treatment. That is, recognizing a kyphosis type, and therefore classifying sagittal misalignments into a specific pattern, provides spine surgeons with a range of treatment guidelines to restore a physiological profile and reduce the odds of adverse events or of poor outcome [3].

To this aim, we considered a dataset of 120 patients (26 male subjects), whose imaging and sets of 14 spine angles were analyzed and annotated by two senior expert spine surgeons. The two surgeons annotated each case with one out of 7 mutually exclusive labels, namely: normal and 6 different types of kyphosis. The normal cases (N) were 14% of the sample; of the abnormal cases, 36% were affected by lumbar kyphosis (L), 23% suffered from thoracic kyphosis (T), 21% from global kyphosis (G), 17% from thoracolumbar kyphosis (TL), while the other disorders (Lower Lumbar (LL), Cervical (C)) accounted for the remaining 9%.

As proof that the classification task was not a trivial one, although the two expert surgeons shared a taxonomic framework that they had jointly published [20], they could only agree on slightly more than two thirds of the cases (68.3%) and only exhibited a moderated agreement (Cohen's Kappa and Krippendorff's Alpha both equal to 0.62). Thus, the considered dataset was a natural example of the weakly supervised learning setting, discussed in Section 2. So, for model training, we applied the techniques proposed in Section 3.1, using as base TWD classifier the state-of-the-art TW Random Forest method [5].

One of the authors, an expert spine surgeon, reviewed 15 predictions provided by a classical (weakly supervised) predictive model h (which also gave the probability score associated with the diagnostic advice), with moderate accuracy (approximately 70%), and compared them with the set-valued predictions provided by a corresponding DCTWD classifier (defined on the basis of a TWD $\mathcal{W}_h$ classifier grounding on h), together with the related probability bound as described in Section 3.2. See Table 10 for a brief summary of the annotated cases. The spine surgeon evaluated the usefulness of the advice and, then, discussed about the rationale for the potential adoption of these approaches in clinical decision support.

From the quantitative point of view, the output of the DCTWD classifier was found to be more useful (or informative), with a mean score of 4.13 (SD = 1.21) (vs 3.80, SD = 0.92), but not significantly so (Mann Whitney test, $U = 97.5$. p-value = 0.55).

On a more qualitative level, the traditional approach was deemed preferable whenever the classifier would be able to provide the decision makers with highly-confident advice. In the case at hand, which we recall was a 7-class diagnostic task, a probability score for a single class was considered high if it was at least 3 times higher than those from a uniform probability distribution (i.e., $1/6$). Nonetheless, also the conformal approach proposed in this paper was deemed valuable in these cases, for its capability to point out the most plausible alternatives, that the decision maker could further consider to definitely rule them out in favor of the diagnostic class singled-out by the traditional approach.

Table 10
Summary of the information regarding the medical cases reviewed by the domain expert. For each case, we report both the single-valued prediction and the set-valued prediction provided by the DCTWD (in parentheses, the probability scores of the two methods), the target labels, and the perceived usefulness of the two types of predictions, measured in an ordinal scale ranging from 1 (very low) to 5 (very high). We also report, for each case, the pelvic tilt (PT) and sacral slope (SS) angles.

Case ID	PT	SS	Target	h (prob.)	DCTWD (prob.)	Usefulness (h)	Usefulness (DCTWD)
83	29	8	L	L (0.68)	L, LL (0.86)	4	4
8	23	24	G	L (0.64)	L, G (0.86)	5	5
87	43	24	L	L (0.76)	L (0.86)	4	5
100	11	22	TL, L	L (0.32)	N, TL, L (0.71)	3	5
115	21	35	TL, LL	T (0.62)	T, L, LL (0.86)	5	3
71	27	12	G, L	G (0.29)	T, L, LL, G (0.94)	5	2
3	39	4	L	L (0.33)	L, G (0.71)	3	5
24	16	39	LL	L (0.71)	L, LL (0.94)	3	4
101	13	33	TL	T (0.64)	T, TL (0.86)	2	4
4	32	26	L	T (0.74)	T, L (0.94)	2	4
124	57	16	L	L (0.75)	L, LL (0.94)	5	3
109	16	28	N	N (0.67)	N, T (0.86)	4	5
104	22	24	N, LL	N (0.32)	N, L, LL (0.71)	5	4
2	20	19	L	N (0.33)	N, T, L (0.86)	2	5
61	15	44	N	N (0.89)	N, L (0.86)	5	4

Conversely, when the traditional approach gives predictions associated with low confidence scores, the set-valued output (provided by the DCTWD) was found to be more useful. This was mainly the case because the set-valued prediction provides an enumeration of the main alternatives, and thus indirectly suggests what further evidence or elements should be considered by the decision maker to rule out some options and keep those that are more compatible with the case at hand.

In the 15 cases considered in this evaluation, the set-valued predictions provided by the DCTWD of alternatives were deemed to be always close to the set of natural candidates for the correct diagnosis that an expert surgeon would have considered if unaided (only for case 115 the DCTWD did not include the label TL in the set-valued predictions, although it included both T and L).

Generalizing, we can assert that whenever the traditional approach provides confidence scores close to the uniform probability distribution, and the probability bounds of the DCTWD are sufficiently high, then presenting both these pieces of advice would be the best option.

The medical expert also provided some comments on two noteworthy cases that we report in what follows, to highlight the kind of reasoning that cautious learning can facilitate in diagnostic tasks:

- Case 1 was described as an odd one (see Fig. 6a). The involved surgeon said that it was probably a normal subject (because the pelvic tilt and the SVA were normal and the combined normality of both parameters leaves little room for a pathological case to be confirmed), who nevertheless exhibited a value of lumbar lordosis that was too low. He agreed upon the fact that other, equally expert, colleagues could have defined the unusual shape of the spine exhibited by case 100 (presenting small pelvic incidence, relatively small lumbar lordosis and small thoracic kyphosis) as unharmonious and weird, irrespective of its occurrence in asymptomatic subjects. Interestingly, the DCTWD was capable to capture this "oddness" and it provided a set-valued prediction that encompassed both normality, lumbar lordosis and thoracic kyphosis as plausible labels.
- Case 2 was deemed extremely interesting. In [20], subjects like case 8, who present values of lumbar lordosis lower than the normative values, and thoracic kyphosis above the normative values, are considered clear instances of global kyphosis. Nonetheless, insufficient lumbar lordosis, combined with decreased thoracic kyphosis, indicates cases of manifest lumbar kyphosis. This puts patients like case 8 in an area of uncertainty, and no current spine deformity classification can associate these patients with a clear-cut category, without the risk of misdiagnosis. Interestingly, the DCTWD method recognizes and reflects this intrinsic uncertainty, by not imposing any specific diagnosis over the others.

The described qualitative evaluation, and the brief discussion of the two cases mentioned above, are just exemplifications. Nonetheless, they allow us to hint at how computational tools, like those integrating some form of machine learning, can support human reasoning, and how decision makers and these tools should interact in naturalistic settings and real-world scenarios.

This latter aspect also relates to *how* plausible classes should be presented, that is *how many* and whether in terms of confidence or probability. Likewise, the usefulness of set-valued predictions was appreciated in almost all the decision settings, as long as the interval did not encompass more than 3 or 4 alternative candidates, irrespective of the number of potential disjoint options.

In our short, but indicative, use case, we showed how human decision makers can collect observations in a medical scenario; combine this information with knowledge on spinal bio-mechanics, developed in either direct or indirect clinical experience (e.g., historical trial and errors, case reports, clinical comparisons); and formulate hypotheses on the basis of what

(a) A sagittal x-ray for the case no. 100. This depicts a thoracolumbar kyphosis that is so light that the alignment is almost normal.

(b) A sagittal x-ray that depicts a manifest, easy-to-detect thoracolumbar kyphosis.

Fig. 6. (a) A sagittal X-ray for the case No. 100. This depicts a thoracolumbar kyphosis that is so light that the alignment is almost normal. (b) A sagittal X-ray that depicts a manifest, easy-to-detect thoracolumbar kyphosis.

a computational decision support gives them. In regard to set-valued output, we saw how this type of support can reflect compatible patterns of spine deformation and compensation, and hence be a useful aid to choose appropriate treatments even if a single option is not highlighted. In fact, the predictions provided by the DCTWD classifier were found to be useful even for the cases for which traditional systems could suggest a single diagnosis with high accuracy, because they acted as triggers for double check and review of less-than-obvious options.

In light of our study, we then make the point that decision support in real-world settings should always leverage some form of cautious prediction; either in conjunction with more traditional approaches, or in isolation. Their usefulness especially emerges in those cases where real life comes in shades of grey, and even well-trained and long-experienced experts cannot classify specific cases with total certainty. Cautious learning approaches can better reflect this intrinsic uncertainty, compared with traditional approaches, and thus they could provide more useful and interpretable decision support [14,23] for decision makers in critical settings, or for under-specified tasks.

6. Conclusion

In this article, we studied the relationship between TWD and CP, two popular cautious learning approaches. To this aim, we introduced the *three-way non-conformity measure*, as well as the *three-way conformal predictor* (TWCP), and discussed two classes of conformal TWD classifiers (i.e., the DCTWD and PCTWD classifiers) by which a conformal predictor can be transformed into a TWD classifier. Through this relationship, the validity of TWD-based ML models is proven for the first time (to our knowledge): this allows to establish reliable learning-theoretic guarantees and error bounds for TWD classifiers.

Furthermore, the definition of optimal cost-sensitive cautious classification algorithms is addressed, along with a characterization of the conditions under which CP and TWD would provide identical results.

From an empirical point of view, we illustrated how the proposed constructions can be used to obtain TWD classifiers that were shown to outperform state-of-the-art TWD, and CP, methods.

Finally, we highlighted the positive potential of the proposed approaches – and cautious learning methods more in general – in the development of reliable decision support, through an illustrative use case, involving a subject-matter expert in a complex medical classification problem.

In conclusion, we believe that our theoretical analysis and the promising results from the empirical study represent a first step, as well as a foundation, for further investigations aimed at characterizing the theoretical aspects of TWD-based ML, and of cautious-learning approaches more in general. For this reason, we believe that the following open problems should be further investigated:

- In Section 3.2, a characterization of the conditions for the equivalence between a TWD classifier and the corresponding PCTWD classifier was proved, under the assumption of a uniform-error loss function. It would be interesting to generalize this characterization to general-loss functions;
- In this paper, we focused on the most basic notion of *validity* (i.e. conservative validity). It would thus be interesting to study also the *probabilistic validity* of TWD classifiers, or their validity in non-i.i.d. settings [2];
- The three-way non-conformity measure was introduced to define CP algorithms based on TWD classifiers. Though this approach allowed a natural comparison among the two studied approaches, it is not optimal in terms of efficiency. It would thus be interesting to study appropriate generalizations of other, efficient [31], CP approaches to the TWD setting;
- The proven validity bounds are instance-wise and can be applied in both online and inductive settings (using a validation set). Nonetheless, it could be interesting to study validation-independent finite-sample bounds. This would require, in turn, to generalize the framework of *PAC learning* theory to TWD-based ML and, more in general, to cautious learning [12];
- Finally, through a simple but indicative case study, in Section 5, we discussed the usefulness of the proposed approaches to develop more reliable and supportive decision support tools. We deem that further assessing the perceived usefulness of TWD, CP and other cautious learning approaches as support tools for human decision makers could be of great interest towards the development of truly reliable Decision Support Systems.

CRediT authorship contribution statement

Andrea Campagner: Conceptualization, Software, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Federico Cabitza:** Conceptualization, Validation, Investigation, Writing - original draft, Writing - review & editing, Supervision. **Pedro Berjano:** Investigation, Writing - review & editing. **Davide Ciucci:** Validation, Writing - original draft, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Adamskiy, I. Nouretdinov, A. Gammerman, Conformal predictors in semisupervised case, in: Statistical Learning and Data Science, 2011, p. 43.
[2] V. Balasubramanian, S.-S. Ho, V. Vovk, Conformal prediction for reliable machine learning: theory, adaptations and applications, Newnes, 2014.
[3] S. Bhagat, V. Vozar, L. Lutchman, R. Crawford, A. Rai, Morbidity and mortality in adult spinal deformity surgery: Norwich spinal unit experience, Eur. Spine J. 22 (2013) 42–46.
[4] N. Bosc, F. Atkinson, E. Felix, A. Gaulton, A. Hersey, A.R. Leach, Large scale comparison of qsar and conformal prediction methods and their applications in drug discovery, J. Cheminf. 11 (2019) 4.
[5] A. Campagner, F. Cabitza, D. Ciucci, The three-way-in and three-way-out framework to treat and exploit ambiguity in data, Int. J. Approximate Reasoning 119 (2020) 292–312.
[6] A. Campagner, F. Cabitza, D. Ciucci, Three-way decision for handling uncertainty in machine learning: a narrative review. In Proceedings of International Joint Conference on Rough Sets 2020 (pp. 137–152). Springer volume 12179 of LNCS.
[7] J.J. Del Coz, J. Díez, A. Bahamonde, Learning nondeterministic classifiers, J. Mach. Learn. Res. 10 (2009).
[8] D. Dubois, H. Prade, Practical methods for constructing possibility distributions, Int. J. Intell. Syst. 31 (2016) 215–239.
[9] D. Dubois, H. Prade, S. Sandri, On possibility/probability transformations, in: Fuzzy logic, Springer, 1993, pp. 103–112.
[10] Elkan, C. (2001). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (pp. 973–978). Lawrence Erlbaum Associates Ltd volume 17.
[11] C. Ferri, J. Hernández-Orallo, Cautious classifiers, ROCAI 4 (2004) 27–36.
[12] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in neural information processing systems (pp. 4878–4887).
[13] P. Gu, J. Liu, X. Zhou, Approaches to three-way decisions based on the evaluation of probabilistic linguistic terms sets, Symmetry 13 (2021) 764.
[14] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, Brain Inf 3 (2016) 119–131.
[15] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization, Int. J. Approximate Reasoning 55 (2014) 1519–1534.
[16] E. Hüllermeier, S. Destercke, I. Couso, Learning from imprecise data: adjustments of optimistic and pessimistic variants, in: International Conference on Scalable Uncertainty Management, Springer, 2019, pp. 266–279.
[17] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach. Learn. 110 (2021) 457–506.
[18] U. Johansson, H. Boström, T. Löfström, Conformal prediction using decision trees, in: 2013 IEEE 13th international conference on data mining, IEEE, 2013, pp. 330–339.
[19] G. Klein, Naturalistic decision making, Hum. Factors 50 (2008) 456–460.
[20] C. Lamartina, P. Berjano, Classification of sagittal imbalance based on spinal alignment and compensatory mechanisms, Eur. Spine J. 23 (2014) 1177–1189.
[21] R. Laxhammar, G. Falkman, Conformal prediction for distribution-independent anomaly detection in streaming vessel data, in: Proceedings of the first international workshop on novel data stream pattern mining techniques, 2010, pp. 47–55.

[22] Y. Li, L. Zhang, Y. Xu, Y. Yao, R.Y.K. Lau, Y. Wu, Enhancing binary classification by modeling uncertain boundary in three-way decisions, IEEE Trans. Knowl. Data Eng. 29 (2017) 1438–1451.
[23] D. Liu, The effectiveness of three-way classification with interpretable perspective, Inf. Sci. 567 (2021) 237–255.
[24] D. Liu, T. Li, D. Liang, Incorporating logistic regression to decision-theoretic rough sets for classifications, Int. J. Approximate Reasoning 55 (2014) 197–210.
[25] D. Liu, X. Yang, T. Li, Three-way decisions: beyond rough sets and granular computing, Int. J. Mach. Learn. Cybern. 11 (2020) 989–1002.
[26] J. Liu, H. Li, X. Zhou, B. Huang, T. Wang, An optimization-based formulation for three-way decisions, Inf. Sci. 495 (2019) 185–214.
[27] Z.-G. Liu, Q. Pan, J. Dezert, G. Mercier, Credal classification rule for uncertain data based on belief functions, Pattern Recogn. 47 (2014) 2532–2541.
[28] L.E. Makili, J.A.V. Sánchez, S. Dormido-Canto, Active learning using conformal predictors: application to image classification, Fusion Sci. Technol. 62 (2012) 347–355.
[29] F. Min, F.-L. Liu, L.-Y. Wen, et al, Tri-partition cost-sensitive active learning through knn, Soft. Comput. 23 (2019) 1557–1572.
[30] I. Nouretdinov, J. Gammerman, M. Fontana, D. Rehal, Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection, Neurocomputing 397 (2020) 279–291.
[31] M. Sadinle, J. Lei, L. Wasserman, Least ambiguous set-valued classifiers with bounded error levels, J. Am. Stat. Assoc. 114 (2019) 223–234.
[32] A.V. Savchenko, Fast inference in convolutional neural networks based on sequential three-way decisions, Inf. Sci. 560 (2021) 370–385.
[33] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (2008) 371–421.
[34] Vovk, V. (2002). On-line confidence machines are well-calibrated. In The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings. (pp. 187–196). IEEE.
[35] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, Springer Science & Business Media, 2005.
[36] V. Vovk, I. Nouretdinov, V. Fedorova, I. Petej, A. Gammerman, Criteria of efficiency for set-valued classification, Ann. Math. Artif. Intell. 81 (2017) 21–46.
[37] H. Wechsler et al, Cyberspace security using adversarial learning and conformal prediction, Intel. Inf. Manage. 7 (2015) 195.
[38] J. Xu, Y. Zhang, D. Miao, Three-way confusion matrix for classification: A measure driven view, Inf. Sci. 507 (2020) 772–794.
[39] Y. Xu, J. Tang, X. Wang, Three sequential multi-class three-way decision models, Inf. Sci. 537 (2020) 62–90.
[40] B. Yang, J. Li, Complex network analysis of three-way decision researches, Int. J. Mach. Learn. Cybern. (2020) 1–15.
[41] G. Yang, S. Destercke, M.-H. Masson, Cautious classification with nested dichotomies and imprecise probabilities, Soft. Comput. 21 (2017) 7447–7462.
[42] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: International Conference on Rough Sets and Knowledge Technology, Springer, 2009, pp. 642–649.
[43] Y. Yao, An outline of a theory of three-way decisions, in: International Conference on Rough Sets and Current Trends in Computing, Springer, 2012, pp. 1–17.
[44] Y. Yao, Tri-level thinking: models of three-way decision, Int. J. Mach. Learn. Cybern. (2019) 1–13.
[45] Y. Yao, The geometry of three-way decision. Applied Intelligence, 2021a, pp. 1–28.
[46] Y. Yao, Set-theoretic models of three-way decision, Gran. Comput. 6 (2021) 133–148.
[47] X. Yue, Y. Chen, B. Yuan, Y. Lv, Three-way image classification with evidential deep convolutional neural networks, Cognit. Comput. (2021) 1–13.
[48] X. Zhan, Z. Wang, M. Yang, Z. Luo, Y. Wang, G. Li, An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction, Measurement (2020) 107588.
[49] Y. Zhang, D. Miao, J. Wang, Z. Zhang, A cost-sensitive three-way combination technique for ensemble learning in sentiment classification, Int. J. Approximate Reasoning 105 (2019) 85–97.
[50] J. Zhou, W. Pedrycz, C. Gao, Z. Lai, X. Yue, Principles for constructing three-way approximations of fuzzy sets: A comparative evaluation based on unsupervised learning, Fuzzy Sets Syst. 413 (2021) 74–98.

Chapter 6

Ensembling of Cautious Predictors

As discussed in the previous chapters, one of the most relevant problems in the cautious inference literature regards the so-called *validity-efficiency trade-off* [174, 251], that is the trade-off between less precise but more accurate predictions, or vice-versa. Even though in the recent years several theoretical advancements have been made in regard to this question, most relevantly through the proposal in [100, 148, 203] of cautious inference methods based on conformal prediction having optimal asymptotic efficiency, the application of such results in practical settings has been limited. As a possible, conceptually simpler, alternative to such statistical methods, in recent years the application of ensemble techniques for the combination of cautious predictors has attracted some interest [14, 18, 236], also due to the popularity and efficacy of ensemble learning in the standard supervised setting [116, 204]. Ensemble learning refer to techniques that allow to obtain, from a collection of models $h_1, \ldots, h_n \in \mathcal{H}$, a new aggregated model $\hat{h}$ (not necessarily in $\mathcal{H}$) that improves on each of the h_i in regard to some quality dimension of interest, e.g. by improving the generalization, or by reducing the variance. Within the setting of standard supervised learning, ensemble methods have been proposed as a flexible approach to control the bias-variance trade-off [124, 140, 191], and have since become one of the most popular learning paradigms. In the cautious inference setting, on the other hand, the issue of how to ensemble cautious predictors, and the properties of such combination techniques, has been investigated in two different strands of research.

On the one hand, as mentioned above, the ensembling of cautious predictors has been investigated, mainly within the framework of conformal prediction, with the aims of improving the data efficiency of such methods while at the same time improving their efficiency and preserve their validity [57, 138, 250] or allowing the effective application of cautious inference techniques in multiple data-source or information fusion settings [14, 223]. On the other hand, ensemble of cautious predictors have been considered as an alternative to state-of-the-art ensemble methods, to reduce base models' overfitting [48] and improve generalization [55, 167], to improve interpretability [18] or to improve robustness to noise in the data [29, 71, 143].

Within the first strand of research, several ensemble techniques have been proposed to address the above mentioned aims, starting from seminal work on cross-conformal, bootstrap-conformal and out-of-bag conformal prediction [154, 250] as well as aggregated conformal prediction [57, 155, 153]. These latter are techniques that are directly based on the ensemble learning scenario and have been proposed as a way to mitigate the need for a separate calibration set in inductive conformal prediction [183], by training different conformal predictors $\Gamma_1, \ldots, \Gamma_n$ on different datasets $S_1, \ldots, S_n$ obtained from a single dataset S either by partitioning or resampling. The base conformal predictors are then ensembled by averaging their p-value functions $\hat{p}_x(y) = \frac{1}{n} \sum_i p^i_x(y)$, thus allowing full utilization of the training set S. At the same time, while it has been shown that such techniques tend to have a stabilizing effect on the p-value functions, thus improving the efficiency, it has been shown in practical settings [155, 153] that the resulting ensemble models often fail to preserve the validity properties of the base cautious predictors that are combined: indeed, existing results on the validity of the above mentioned methods [57] rely on strong assumptions on the resampling procedure.

More recently, the above mentioned ensemble learning-based combination techniques have been extended also to ensembles of conformal predictors possibly trained on different and unrelated datasets, as a way to allow application of such techniques in information fusion, multi-source or multi-modal data, as mentioned above. Among such techniques, aside from direct generalizations of the aggregated conformal pre-

dictor, the two most popular approaches have been inspired either by research in multiple hypothesis testing [163] and meta-analysis [119], using approaches such as quantile and order-statistic methods [14, 236, 237], or by voting theory [65]. In the first case, an approach based on Fisher's method for combining p-values has gained popularity due to its good empirical performance [14, 236]. According to this procedure, the p-values given by different conformal predictors $\Gamma_1, \ldots, \Gamma_n$ are combined by applying the rule $\hat{p}_x(y) = k \sum_{i=0}^{n-1} \frac{(-\log k)^i}{i!}$, where $k = \Pi_i p_x^i(y)$: the intuition behind this approach relies on the observation that, if the p-values are independent, the values $-log_2 p_x^i$ are distributed as a χ^2 variable, from which the above formula can be derived. In the second case, on the other hand, combination is performed at the level of confidence sets, rather than p-values [65]: given confidence sets $T_1^\epsilon(x), \ldots, T_n^\epsilon(x)$ at a specific threshold level ϵ, these are combined either by selecting the set-valued prediction $\hat{T}(x)$ containing all classes for which $|\{i : y \in T_i^\epsilon(x)\}| \geq \frac{n}{2}$. The intuition for this approach stems from voting theory, in which Condorcet jury theorem [31] provides an asymptotic bound on the probability of correctness for the aggregation of votes according to majority voting, which directly translates into a weak form of validity for the above mentioned combination rule. Despite the empirical effectiveness of such combination rules, and similarly to the above mentioned ensemble learning-inspired combination methods, all of the above mentioned approaches have been shown to lose the validity properties enjoyed by the combined cautious predictors, and their theoretical properties have not been studied [14].

Within the second strand of research mentioned above, focused on the application of cautious predictors as base models for standard ensemble learning techniques, most existing approaches have been grounded on the application of approval voting schemes [37], by which multiple set-valued predictions $T_1, \ldots, T_n$, no matter by which cautious predictor have they been obtained, are combined into a single, precise prediction $\hat{y} = \arg\max_{y \in Y} |\{i : y \in T_i(x)\}|$. Similarly to the first setting mentioned above, however the theoretical properties of such ensemble techniques have not yet been clarified in general settings, and research has mostly focused on the empirical evaluation of their effectiveness in comparison with state-of-the-art ensemble

methods [18, 48]. Also from this latter perspective, however, the assessment of such approaches has been mostly based on limited benchmarks, with a lack of study focusing on showing whether ensembles of cautious predictors really provide an effective improvement as compared to state-of-the-art ensemble methods.

The aim of this chapter, then, will be to study the ensembling of cautious predictors so as to address research problem **P2.2**, within the two setting described above, both from an empirical as well as a theoretical point of view.

First, in Section 6.1, the main aim will be to study the empirical performance of ensembles of cautious predictors as a way to improve the generalization and robustness of state-of-the-art ensemble methods. Such an aim will be addressed by means of an extensive experimental analysis of ensemble techniques based on a large set of benchmark datasets. The focus will be devoted to ensembles of cautious predictors obtained by the application of three-way decisions, as described in Section 5, to the ensemble base classifiers [48] which are then aggregated by means of approval voting. This class of cautious inference-based ensemble models will be compared against both state-of-the-art standard ensemble methods, as well as ensemble techniques inspired by other uncertainty quantification approaches [133]. The main contribution of Section 6.1 will then be to show that such ensembles of cautious predictors, and more generally ensembles of models building on uncertainty quantification schemes [133], can provide significantly better performance, as well as improved robustness to uncertainty in the data and the curse of dimensionality, as compared to commonly adopted state-of-the-art techniques.

On the other hand, in Section 6.2, the main aim will be to study the theoretical properties, in terms of validity and efficiency, of ensembles of cautious predictors, focusing on techniques based on conformal prediction. To this aim, the two main theoretical contributions of this section will be the proposal of a general framework for the definition and analysis of conformal prediction combination rules, based on the correspondence between conformal prediction and possibility theory [62] and the application of copula theory [178], as well as the theoretical study of several such combination rules in a general, information fusion-inspired setting [14] which relaxes

the main assumptions introduced in the previous proposals in this sense [57, 65]. These theoretical contributions will be complemented by an empirical contribution by which the effectiveness of the above mentioned combination rules will be evaluated in the setting of multi-variate time series classification (MTSC). This task was selected due to its practical relevance as well as due to the certain characteristics that make MTSC particularly relevant for the evaluation of the above mentioned models In particular, one of the main characteristics of the MTSC setting relates to the distinction between bespoke methods, which employ all the multivariate information at the same time and thus make full use of the available information, and univariate methods, which combine classifiers trained on each univariate slice of the time series of interest, and hence may be preferable in terms of computational efficiency or applicability in resource-constrained settings. To this aim, the proposed combination rules will be evaluated, on a standard benchmark collection of MTSC datasets, as way to implement robust univariate MTSC methods, in comparison with both standard state-of-the-art bespoke classifiers proposed in this setting, other cautious inference ensemble methods proposed in the literature, as well as state-of-the-art cautious predictors. The main empirical contribution in this sense, then, will be to show that combination of conformal predictors, each of which based only on a univariate slice of the time series to be classified, can perform as well as, or better than, state-of-the-art bespoke methods, both based on multi-variate standard classifiers as well as cautious inference methods.

Information Fusion 90 (2023) 241–252

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Aggregation models in ensemble learning: A large-scale comparison

Andrea Campagner [a,*], Davide Ciucci [a], Federico Cabitza [a,b]

[a] Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano–Bicocca, viale Sarca 336, 20126, Milano, Italy
[b] IRCCS Istituto Ortopedico Galeazzi, Milano, Italy

ARTICLE INFO

Dataset link: https://github.com/AndreaCampagner/Aggregation-Models-in-Ensemble-Learning

Keywords:
Aggregation methods
Ensemble learning
Information fusion
Uncertainty management
Social choice theory
Collective intelligence

ABSTRACT

In this work we present a large-scale comparison of 21 learning and aggregation methods proposed in the ensemble learning, social choice theory (SCT), information fusion and uncertainty management (IF-UM) and collective intelligence (CI) fields, based on a large collection of 40 benchmark datasets. The results of this comparison show that Bagging-based approaches reported performances comparable with XGBoost, and significantly outperformed other Boosting methods. In particular, ExtraTree-based approaches were as accurate as both XGBoost and Decision Tree-based ones while also being more computationally efficient. We also show how standard Bagging-based and IF-UM-inspired approaches outperformed the approaches based on CI and SCT. IF-UM-inspired approaches, in particular, reported the best performance (together with standard ExtraTrees), as well as the strongest resistance to label noise (together with XGBoost). Based on our results, we provide useful indications on the practical effectiveness of different state-of-the-art ensemble and aggregation methods in general settings.

1. Introduction

The term ensemble learning [1] (EL) refers to algorithms that rely on the combination and fusion of multiple *base models* to obtain a more accurate averaged model. This learning paradigm is currently one of the most popular ones thanks to the outstanding empirical performance [2] of its two main variants, namely *bagging* [3] and *boosting* [4], on a variety of applications including survival analysis [5], rule mining [6], outlier detection [7].

Interestingly, the main theoretical notion underlying EL, that is the information fusion procedure by which high-quality information is extracted from large ensembles of weak classifiers, is not unique to Machine Learning (ML) and has long been explored also in other research fields including *social choice theory* (SCT) [8]; *collective intelligence* (CI) [9]; as well as *information fusion and uncertainty management* (IF-UM) [10–13].

Nonetheless, research in these communities has been devoted to largely different concerns. Research in EL has largely focused on the development of more effective *learning strategies* [14–16] and their theoretical analysis [17–19]. Conversely, research in SCT, CI and IF-UM has largely focused on the development and analysis of novel *aggregation rules* [8,12,20], focusing in particular on the validity and efficiency properties of such aggregation rules [13,21]. These latter approaches have shown promising performances compared with traditional majority-based aggregation [22], usually employed also in EL. Moreover, research at the intersection of these fields has been lagging behind. Notably, the different aggregation mechanisms developed in SCT and CI have rarely been evaluated in the EL literature, with few exceptions [23–25].

This work aims to bridge this gap in the literature, by means of a large-scale empirical comparison of different aggregation procedures in EL. The main goal of this paper, then, is to provide useful guidance about the pros and cons of these different procedures in real-world applications. To this aim, we evaluated 21 different aggregation methods on a large collection of 40 benchmark datasets, selected so as to encompass different application domains, different number of instances ($150-58509$), different number of classes ($2-20$), and different number of features ($4-1000$).

2. Related work

In the recent years, there has been an increasing interest toward the relationships among EL and other related research fields, including SCT, CI and IF-UM. One of the first works in this line of research is [26], which studied the properties of different EL methods in the framework of SCT. Since then, most research works have focused on the development of EL methods inspired by aggregation methods proposed in SCT, CI and IF-UM. Ruta et al. [27] evaluated the use of plurality voting for classifier selection; Gandhi et al. [28] developed an hybrid ensemble method based on the plurality voting rule; Cornelio et al. [24] described a general approach for developing EL algorithms based on SCT-inspired

* Corresponding author.
E-mail address: a.campagner@campus.unimib.it (A. Campagner).

https://doi.org/10.1016/j.inffus.2022.09.015
Received 2 December 2021; Received in revised form 17 September 2022; Accepted 18 September 2022
Available online 23 September 2022
1566-2535/© 2022 Elsevier B.V. All rights reserved.

voting mechanisms; Luo et al. [29] and Campagner et al. [23] described an EL approach based on the Surprisingly Popular Algorithm [20]; while Abellan et al. [10], Balasubramanian [21], Campagner et al. [23] and Toccaceli et al. [13] considered the application of approaches inspired by IF-UM.

Nonetheless, and despite claims of alternative methods outperforming traditional ensemble-based ones, few evaluation studies have aimed at comparing different ensemble aggregation methods. Shipp et al. [30] performed a comparison of 10 different aggregation methods, and studied the relationship between these latter and ensemble diversity measures. Narassiguin et al. [31] performed an extensive comparison of traditional EL methods in the binary classification setting; Leon et al. [25] compared the Bagging-based EL approach with 3 SCT-inspired aggregation methods on 3 toy datasets. Bagging-based methods and SCT methods were also compared in [24,32], through a more extensive evaluation on more than 20 benchmark datasets. Though no statistical analysis was performed, these latter research works highlighted how SCT-based methods can be competitive with traditional EL methods. However, these findings were based on relatively small and low-dimensional datasets. By contrast, Campagner et al. [23] compared Random Forest and Boosting with methods from CI, SCT and IF-UM on a collection of 10 small datasets and found SCT-based methods to be out-performed by all other methods, while CI and IF-UM methods were found to out-perform traditional EL approaches.

3. Methods

In this section, we recall the basic notions of EL, as well as of the aggregation rules that we compared in our experiments. Finally, we describe the adopted experimental setting. In the rest of the article, we denote by X the set of instances and by Y the set of class labels. As we focus on classification, we assume that $Y = \{0, \dots, |Y|\}$. Further, we focus on scoring classifiers, that is functions $h : X \mapsto [0,1]^{|Y|}$, where $h^y(x)$ denotes the probability score that h assigns label y to instance x, i.e. the empirical estimate of $Pr(y|x)$ given by the classifier h. Given a scoring classifier h, we denote with $\hat{h} = [h^{(1)}, \dots, h^{(|Y|)}]$ the vector obtained by sorting the probability scores (and corresponding class labels) in decreasing probability order, i.e. such that $h^{(y)} \geq h^{(y+1)}$.

3.1. Ensemble learning methods

EL refers to a ML paradigm based on the aggregation of multiple ML models into a single model (i.e., an *ensemble*) [33], with the aim of flexibly controlling the bias–variance trade-off, i.e., the trade-off between model capacity and generalization. Formally, given a set of *base* predictors $h_1, \dots, h_n$, the goal is to obtain a combined predictor h_{ens} such that its predictive performance is better than the performance of each h_i. Since its original proposal in the context of statistical learning theory, many EL approaches have been proposed. The main methods include boosting [34], bagging [3], stacking [35], Bayesian model averaging [36] and post-aggregation [37], each of which encompasses different variations such as AdaBoost [38], Random Forest [39], Rotation Forest [40], Random Patches [41], Gradient Boosting [42], and others [43]. In this article, we focus on the two most popular strategies, namely bagging and boosting [1].

Bagging is an EL method used to improve the performance and generalization of unstable estimators [3,44], such as Decision Trees. Given a training set T, B bootstrap samples are generated from T, each of which is used to train a base classifier. These latter ones are then aggregated by some form of majority voting, either simple or weighted. We evaluated two state-of-the-art approaches related to Bagging, i.e., *Random Forest* [39],[1] and *Extra-Trees* [45].[2]

Boosting was first introduced in [4,46], and denotes a family of learning algorithms that iteratively train and aggregate models: as new models are added to the ensemble, increasingly greater relevance is associated to those instances that were mis-classified by the previous models. While this relevance assignment procedure depends on the specific Boosting algorithm [34], most methods are based on the evaluation of a loss function that drives the construction of the ensemble. In our experiments we considered AdaBoost [38], Gradient Boosting [42], and XGBoost [47], which are among the most commonly used and best-performing methods in real-world practice [2].

3.2. Aggregation techniques

The main source of variation among EL approaches introduced in the previous section lies in the selection of techniques for model training: by contrast, model combination is usually performed by majority voting, either simple (i.e., *Plurality* voting, in SCT) or weighted. Nonetheless, in SCT, CI and IF-UM many other aggregation algorithms have been proposed. In what follows, we describe the combination approaches that we evaluated in this paper.

Plurality Voting is the aggregation rule most frequently considered in SCT, CI and EL [48]. In this approach the most voted class, among the base models, becomes the outcome of the EL model. A common variation is *weighted plurality*, often employed in the EL setting: for example, the aggregation method adopted in AdaBoost can be understood as a variation of weighted plurality in which every base classifier is assigned a weight based on its accuracy on the training set; similarly, a common variation on the Bagging approach assigns to each base classifier h_i, for each instance x, a weight which depends on the probability scores that the base classifiers assign to the classes.

Borda Count is an aggregation rule, proposed in SCT [8], that requires the models to rank the classes in a sorted list. This ranking is obtained based on the ordering of the class labels by the models' probability scores. Based on the assigned ranks, a numerical score is assigned to each option [49]. More precisely, let y be a class label and $h_1, \dots, h_n$ be the base models. Then, the score of y for base model h_i is computed as $\frac{1}{|\{j \in Y : h_i^j(x) \geq h_i^y(x)\}|}$, i.e., the reciprocal of the number of class labels whose probability score is greater than or equal to that of the given label y. Intuitively, the higher y's probability score is, the higher its rank and the closer to 1 its score. After the scores are computed for all models and classes, the scores for each class are then summed across the models, and the class with larger total score is the output of the ensemble. The full algorithm of the Borda Count aggregation method is reported in Algorithm 1. We note that the time complexity of Algorithm 1, for each instance x, is $O(n \cdot |Y| log |Y|)$, where n is the number of base models in the ensemble.

Algorithm 1 Borda aggregation method

```
procedure BORDA-PREDICT({h_i}_{i=1}^n, x)
    scores ← int[|Y|] filled with 0
    for i = 1 to n do
        for y = 1 to |Y| do
            scores[y] += 1 / |{j ∈ Y : h_i^j(x) ≥ h_i^y(x)}|
        end for
    end for
    return argmax_{y∈Y} scores[y]
end procedure
```

Copeland Rule is an aggregation method studied in SCT [50] satisfying many desirable rationality properties [8]. According to *Copeland* rule, for each pair of classes y_1, y_2, a net score is defined as $Net(y_1, y_2) =$

[1] Compared to standard Bagging, in Random Forest, also the features used by the base models are randomly selected.

[2] Formally, Extra-Trees is not a Bagging approach, as it is not based on bootstrapping. Nonetheless, as most implementations of Extra-Trees allow bootstrapping as a further way to reduce variance, we include the algorithm in the Bagging category.

$|\{h_i : h_i^{y_1} > h_i^{y_2}\}| - |\{h_i : h_i^{y_2} > h_i^{y_1}\}|$. The output of the ensemble is the class with the largest total net score, i.e., $argmax_{y \in Y} \sum_j Net(y, j)$. The algorithm for the Copeland Rule aggregation method is reported in Algorithm 2. The time complexity of Algorithm 2 is $O(n \cdot |Y|^2)$, where n is the number of base models.

Algorithm 2 Copeland aggregation method

```
procedure COPELAND-PREDICT({h_i}_{i=1}^n, x)
    scores ← int[|Y|, |Y|] filled with 0
    for y_r = 1 to |Y| do
        for y_c = 1 to |Y| do
            p ← |{h_i : h_i^{y_r}(x) > h_i^{y_c}(x)}|
            n ← |{h_i : h_i^{y_r}(x) < h_i^{y_c}(x)}|
            scores[y_r, y_c] ← p − n
        end for
    end for
    return arg max_{y∈|Y|} Σ_j scores[y, j]
end procedure
```

Approval Voting refers to an aggregation rule in which each voter is allowed to report a set of *approved alternatives* [51]. In terms of EL, this means that each base classifier is transformed into a cautious classifier [52] that provides as output a set of classes (rather than a single-valued prediction). A cautious classifier can then be represented as a function $h : X \mapsto 2^Y$. The class which was reported most frequently is then the output of the ensemble model. This IF-UM-based technique [23] was proposed as a form of regularization, inspired by label smoothing [53,54] and cautious inference [55], to avoid the over-fitting of the base classifiers and mitigate label noise. It was also considered in the conformal prediction literature [56] as a method to improve the efficiency of a set of cautious classifiers to be combined, while preserving their validity. We considered two different approaches for transforming the base learners into set-valued classifiers, namely either by selecting all classes associated with a score larger than a fixed threshold $\tau = \frac{1}{nClasses}$, or through the three-way reduction [23,57]. In this latter case, if as said above we denote with $\hat{h}_i(x) = [h_i^{(1)}(x), \ldots, h_i^{(|Y|)}(x)]$ the ranking of the class labels in decreasing probability order for model h_i on input x, then $three-way(i)(x) = \{y_1, \ldots, y_m\} \subseteq Y$; where $m = \min\{k \in Y : \sum_{j=1}^{k} h_i^{(j)}(x) \geq \tau\}$, with τ a numeric threshold. Intuitively, the three-way reduction maps a scoring classifier into a set-valued one that outputs the first a_j classes, ranked by their probability scores, s.t. their sum is greater than τ. In the experiments, we set $\tau = 0.75$ for both multi-class and binary problems. The rationale for this threshold is that, due to the equivalence between three-way decision and conformal prediction [55], under weak assumptions on the calibration of the base classifiers the correct class label is guaranteed to be contained in the set-valued prediction with probability greater than 75%. The pseudocode for the Approval and Three-way aggregation methods is reported in Algorithms 3 and 4. It is easy to observe that the computational complexity of Algorithms 3 and 4 is given by, respectively, $O(n \cdot |Y|)$ and $O(n \cdot |Y| log |Y|)$.

Algorithm 3 Approval aggregation method

```
procedure APPROVAL-PREDICT({h_i}_{i=1}^n, x, τ)
    scores ← int[|Y|] filled with 0
    for i = 1 to n do
        for y ∈ Y do
            if h_i^y ≥ τ then
                scores[y] += 1
            end if
        end for
    end for
    return argmax_{y∈|Y|} scores[y]
end procedure
```

Algorithm 4 Three-way aggregation method

```
procedure THREE-WAY-PREDICT({h_i}_{i=1}^n, x, τ)
    scores ← int[|Y|] filled with 0
    for i = 1 to n do
        Let ĥ_i = [h_i^(1)(x), ..., h_i^(|Y|)(x)] be s.t. h_i^(j)(x) ≥ h_i^(j+1)(x)
        acc ← 0
        j = 1
        s = set()
        while acc < τ do
            v, y ← h_i^(j)(x) and corresponding label
            acc += v
            s.add(y)
            j+ = 1
        end while
        for y ∈ s do
            scores[y] += 1/|s|
        end for
    end for
    return argmax_{y∈|Y|} scores[y]
end procedure
```

Possibilistic Reduction is an IF-UM-based aggregation method, proposed in [23] and inspired by possibility theory [58] and label smoothing [53,54], as well as by the *averaging aggregation* in the conformal prediction literature [21,59]. Thus, the possibilistic reduction is an approach to regularize base models in EL methods. For each base classifier h_i in the ensemble model, the probability scores, for a given instance x, are transformed to a possibility distribution as $poss(i)(x) = \left[\frac{h_i^{(1)}(x)}{h_i^{(1)}(x)}, \frac{h_i^{(2)}(x)}{h_i^{(1)}(x)}, \ldots, \frac{h_i^{(|Y|)}(x)}{h_i^{(1)}(x)}\right]$; where, as before, $h_i^{(1)}(x), \ldots, h_i^{(|Y|)}(x)$ denotes the ranking of the class labels in decreasing probability order. In the experiments, we considered two aggregation rules: either by summing the possibility scores, or by multiplying them. Intuitively, the sum-based aggregation is similar to weighted plurality, in which, however, an even larger weight is assigned to the top-ranked class [23]. On the other hand, the product-based aggregation corresponds to (unnormalized) Dempster's combination rule on possibility distributions [10,60], and strongly penalizes the classes that have been assigned low possibility scores by at least one of the base models.

Surprisingly Popular Algorithm (SPA) is a CI-inspired aggregation method proposed in [20]. The main intuition for this voting rule is that, in real-life situations, raters do not just express their own preference, but also make assumptions on how others could have voted. In EL, the *SPA* approach has been applied by [23,29], showing excellent predictive performance. In SPA, the base models in the ensemble not only make their prediction $y = \arg\max_{c \in Y} h_i^c(x)$, but are also tasked with predicting the final output of the ensemble (that is, the most popular class within the ensemble). We denote these latter probability scores, for model i and class label $y \in Y$, as $s_i^y(x)$. The final output of the ensemble is the class c that maximizes the difference $SP(x) = |\{i : argmax_{c \in Y} h_i^c(x) = y\}| - |\{i : argmax_{c \in Y} s_i^c(x) = y\}|$. In this article, we considered the Bagging-based implementation of *SPA* introduced in [23,29] and reported in Algorithm 5. If we denote with T the time complexity of training a base model, then it is easy to observe that the computational complexity of Algorithm 5 is given by $O(2nT + 2n \cdot |Y|)$.

3.3. Experimental design

In our experiments we considered 21 different aggregation and ensemble learning methods. First, we considered standard ensemble learning methods, including three boosting methods, namely AdaBoost, Gradient Boosting and XGBoost (AB, GB, XGB), as well as two bagging methods, namely Random Forest and ExtraTrees, using both the

Algorithm 5 Surprisingly Popular Algorithm

procedure SURPRISINGLY POPULAR ENSEMBLING($D = [X, y]$: $dataset, n$: num of models)
  Sample $D_1 = \langle X_1, y_1 \rangle, ..., D_n = \langle X_n, y_n \rangle$ with replacement from D
  Train base model h_i on each D_i
  E = Bagging-Ensemble($h_1, ..., h_n$)
  $\hat{y}$ = E.predict(X)
  for $i = 1$ to n **do**
    Train base model s_i on $(X_i, \hat{y})$
  end for
  return$(h_1, ..., h_n), (s_1, ..., s_n)$
end procedure
procedure SURPRISINGLY POPULAR PREDICT($((h_1, ..., h_n), (P_s, ..., s_n))$: ensemble, x : instance)
  $c = (0, ..., 0)$ s.t $|c| = |Y|$
  $p = (0, ..., 0)$ s.t $|p| = |Y|$
  for $i = 1$ to n **do**
    for $y \in Y$ **do**
      $c[y] = c[y] + h_i^y(x)$
      $p[y] = p[y] + s_i^y(x)$
    end for
  end for
  $\hat{y} = \arg\max_{y \in Y} \{c[y] - p[y]\}$
return $\hat{y}$
end procedure

weighted majority (RandomForest, ExtraTrees) and simple Plurality (Plurality_DT, Plurality_ET) aggregation rules.

Then, we considered Bagging-based implementations of all the above mentioned aggregation rules: in all cases, the ensemble learning algorithm was implemented by training the base models and then using each of the aggregation rules to combine the predictions of the base models. In detail, we considered two Bagging-based implementations of SPA: these two implementations were realized by the pseudo-code reported in Algorithm 5 with two different base classifiers, either Decision Tree (SPA_DT) or ExtraTree (SPA_ET). Similarly, we also considered two Bagging-based implementations for each of Borda Count and Copeland Rule (Borda_DT and Copeland_DT using Decision Tree as base classifier, Borda_ET and Copeland_ET using ExtraTree as base classifier) based respectively on Algorithms 1 and 2; four Bagging-based implementations of Approval voting, two of which using the threshold-based method described in Algorithm 3 and two of which using the three-way method described in Algorithm 4 (Approval_DT and Threeway_DT using Decision Tree as base classifier, Approval_ET and Threeway_ET using ExtraTree as base classifier); four Bagging-based implementations of the Possibilistic reduction, two of which using the sum-based aggregation and two of which using the product-based aggregation (PossSum_DT and PossProd_DT using Decision Tree as base classifier, PossSum_ET and PossProd_ET using ExtraTree as base classifier). In particular, the selected ensemble learning algorithms (AB, GB, XGB, RF, ET) are the most commonly adopted ensembling approaches, both in the literature and in practice, while the selected aggregation rules (weighted majority, plurality, Borda count, Copeland rule, Approval voting, Three-way reduction, Possibilistic reduction, SPA) can be considered a representative selection of some commonly adopted combination methods in SCT, CI and IF-UM. We decided to consider two alternative implementations of each ensemble learning method, i.e., using either Decision Tree or ExtraTree as base classifiers, to evaluate the interplay between aggregation methods and base classifiers. Nonetheless, we also performed a non-differentiated analysis by which we did not distinguish methods by base classifier but only by aggregation method employed. A summary of all considered ensemble and aggregation methods, along with the corresponding source of inspiration, is reported in Table 1.

Table 1
Summary of the considered ensemble and aggregation methods along with their respective sources of inspiration.

Method	Inspiration
AdaBoost	EL (Boosting)
Gradient Boosting	EL (Boosting)
XGBoost	EL (Boosting)
Random Forest	EL (Bagging)
ExtraTrees	EL (Bagging)
Plurality Voting	SCT, CI
Borda Count	SCT
Copeland Rule	SCT
Approval Voting (threshold-based)	SCT, IF-UM
Approval Voting (Three-way reduction)	SCT, IF-UM
Possibilistic reduction (sum-based)	IF-UM
Possibilistic reduction (product-based)	IF-UM
Surprisingly Popular Algorithm	CI

All algorithms were evaluated on 40 benchmark datasets, of which 35 obtained from the UCI repository [61], and 5 synthetic datasets (data0 to data50). The UCI datasets were selected to provide as much variation as possible in terms of number of features, number of instances, number of classes, and application field. The aim was to ensure that our results were generalizable across different domains and tasks. On the other hand, the synthetic datasets were randomly generated, to evaluate the susceptibility of the models to label noise (i.e. errors in the target variable). We set the number of instances to 10 000, the number of features to 1000, the number of classes to 10, and varied the amount of label noise (i.e. the probability of observing an incorrect label) in $[0\%, 5\%, 10\%, 25\%, 50\%]$. The full list of datasets is reported in Appendix A, in Table A.2.

In the experiments, for all ensemble models, we set the number of base classifiers to 100 and all other hyper-parameters (i.e., for all algorithms: maximum tree depth, maximum number of evaluated features, split criterion; additionally for boosting methods: learning rate, undersampling rate) were optimized through a grid-search procedure. The full range of hyper-parameters, for all evaluated models, is reported in Appendix A, in Table A.1: DecisionTree (resp. ExtraTree) was used as a base estimator for all algorithms whose name ended with _DT (resp. _ET). For all algorithms involving randomization the value of the seed was set to 0, so as to ensure reproducibility.

Training, hyper-parameter selection and testing were performed by means of 5–3 nested cross-validation (CV), that is 5 folds for the outer CV and 3 folds for the inner CV. Each dataset was first split in five equal-sized folds. For each iteration of the outer CV, 4 folds of the dataset (80% of the data) were used for training and hyper-parameter selection (*train/valid data*) and the remaining fold (20% of the data) for testing. For each iteration of the outer CV, a 3-fold CV was applied on *train/valid data*: at each iteration of the inner CV, 2 folds (66.67% of *train/valid data*) were used for training and the remaining fold (33.33% of *train/valid data*) for hyper-parameter selection.

To measure the performance of the models, we considered the average test performance across the 5 iterations of the outer CV. In regard to evaluation metrics, in order to account for label imbalance, we considered the balanced accuracy. We also measured the running time (in ms), so as to identify potential differences in terms of computational efficiency among the different aggregation methods. For both balanced accuracy and running time, comparison among the models was performed by considering the average ranks obtained on the collections of 40 datasets. Namely, for each dataset we ranked the ensemble methods from best to worst. For each ensemble method we then considered its average rank across the 40 datasets. The average value of the performance metrics on the 40 datasets is also reported. Statistically significant differences were assessed by means of the Friedman omnibus test [62] and Nemenyi post-hoc test [63], both with $\alpha = 0.05$. We decided to use the Friedman test since it is a non-parametric, rank-based alternative to the repeated measures ANOVA, which is thus more robust to violations of the assumptions of this latter test.

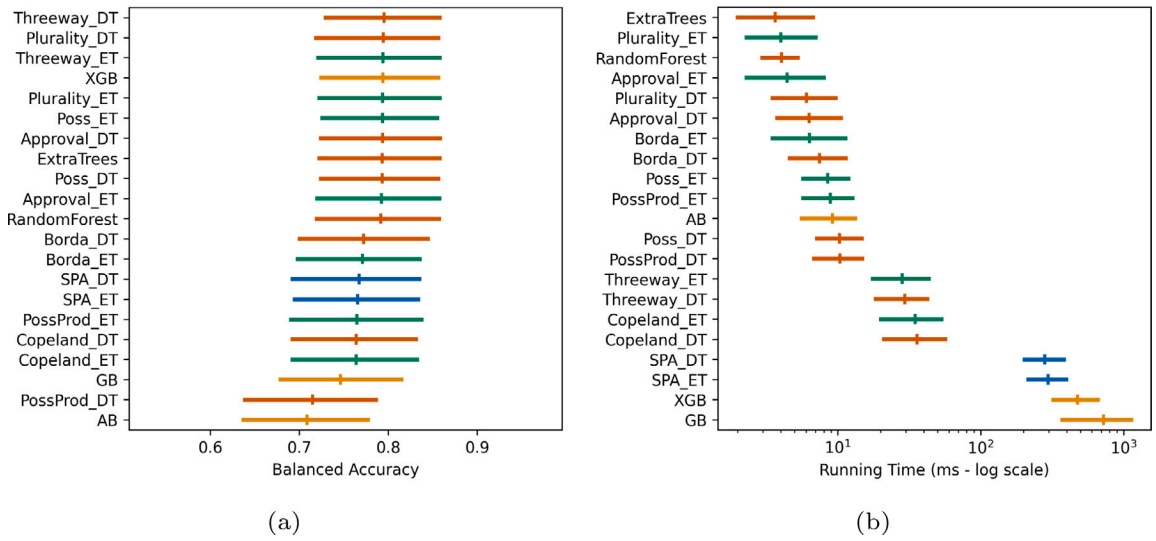


Fig. 1. Results of the experiments. Left: mean balanced accuracy scores of the models under study (higher is better), Error bars denote 95% C.I. Mean running times (ms) of the models under study (lower is better). Error bars denote 95% C.I. Legend, okra: Boosting-based, green: ExtraTree-based, orange: Decision Tree-based, blue: SPA-based.

In order to provide further information on the behavior of the ensemble models on different types of datasets we also considered three additional comparisons, by which we evaluated the ensemble models' behavior w.r.t. increasing number of classes, increasing number of features and increasing levels of label noise. All code was implemented in Python (ver. 3.9.5), using pandas (ver. 1.2.4), scikit-learn (ver. 0.24.2), xgboost (ver. 1.4.2), numpy (ver. 1.19.5), scipy (ver. 1.6.3) and scikit-posthocs (ver. 0.6.7), and is publicly available online (together with all results and employed datasets) on GitHub at https://github.com/AndreaCampagner/Aggregation-Models-in-Ensemble-Learning.

4. Results

The results of the experimental comparison, in terms of average balanced accuracy and average running time (with respective 95% confidence intervals), are reported in Figs. 1(a) and 1(b). The average ranks of the algorithms, across all datasets, are reported in Figs. 2(a) and 3(a) . The p-value for the omnibus Friedman test w.r.t. balanced accuracy and running time were both < 0.0001. Thus, since the result of the omnibus comparison was statistically significant, we performed a post-hoc comparison using the Friedman–Nemenyi procedure. The p-values for the post-hoc analysis are reported in Figs. 2(b) and 3(b). The performance of the models, differentiated by the number of classes, number of features, and level of label noise, are reported, respectively, in Figs. 4, 5 and 6. The results of the aggregation methods, measured without distinction by base classifier, are reported in Appendix B in Figs. B.1(a) and B.1(b), in terms of balanced accuracy and running time, while their rank comparison and statistical analysis is reported in Appendix B in Figs. B.2(a) and B.2(b), in terms of balanced accuracy, and Figs. B.3(a) and B.3(b), in terms of running time.

5. Discussion

We observed significant differences among the evaluated algorithmic families. More in detail, the *Boosting* algorithms, except for XGB, were out-performed by almost all other aggregation approaches. Indeed, in terms of the mean ranks (w.r.t. balanced accuracy) both AB and GB reported the worst classification performance, while in terms of the mean balanced accuracy only PossProd_DT obtained a performance comparable to that of AB and GB. Also, most of these observed differences were statistically significant. Indeed, as seen in Fig. 2(b) the p-values for almost all comparisons involving either AB and GB were lower than the adopted significance threshold. GB, in particular, was also the worst performing algorithm in terms of average running time and among the 5 worst algorithms in terms of mean ranks (w.r.t. running time), comparable only with XGB, SPA_DT, SPA_ET and Threeway_DT. A possible explanation for the observed poor performance of AB and GB can be observed in Figs. 4, 5 and 6. Indeed, both algorithms were the most impacted by both increasing data dimensionality and number of classes. In particular, while AB was the best performing algorithm (along with XGB) in binary classification tasks, its performance sharply decreased with more classes. This finding can be explained by the fact that boosting models [64] are known to be more affected by overfitting in the case of overlapping classes, whose occurrence is obviously more likely when the number of classes increases. AB and GB were also strongly impacted by increasing label noise, a well known and widely reported fact in the previous literature [65]. Our findings suggest that while AB and GB can be very effective in low-dimensional or binary classifications settings, they should not be used in problems that are either high-dimensional or likely to be affected by label noise. In particular, our findings suggest that AB is preferable to GB, since even though the two algorithms have similar performance, AB is significantly more computationally efficient (see Fig. 3(a)).

In contrast with the poor performance reported by AB and GB, XGB was among the best performing algorithms, both in terms of mean balanced accuracy (4th best) and mean ranks (6th best). Moreover, in terms of the comparison obtained by unifying the base classifiers (see Figs. B.1(a), B.2(a) and B.2(b) in Appendix B), XGB was similarly among the best models (3rd best in terms of mean balanced accuracy, 4th best in terms of mean ranks). In particular, XGB was significantly more accurate than the other boosting methods, of SPA algorithms, of the SCT-based Borda and Copeland aggregation algorithms, as well as of PossProd. This finding is in agreement with the widely reported effectiveness of XGB [2,66,67]. In particular, Figs. 4, 5 and 6 show that XGB was among the most robust methods in regard to both increasing data dimensionality and label noise. In regard to data dimensionality, even if the performance of XGB decreased with increasing number of features, the reported drops were not statistically significant except for the case of more than 100 features. In regard to label noise, even if the performance of XGB significantly dropped with increasing levels of label noise, the reported decrease was among the lowest ones among the considered models. This latter finding confirms the previous results reported in [68], which showed that XGB was the most noise-resistant algorithm among boosting methods. An explanation for these findings could be found in the use of curvature information

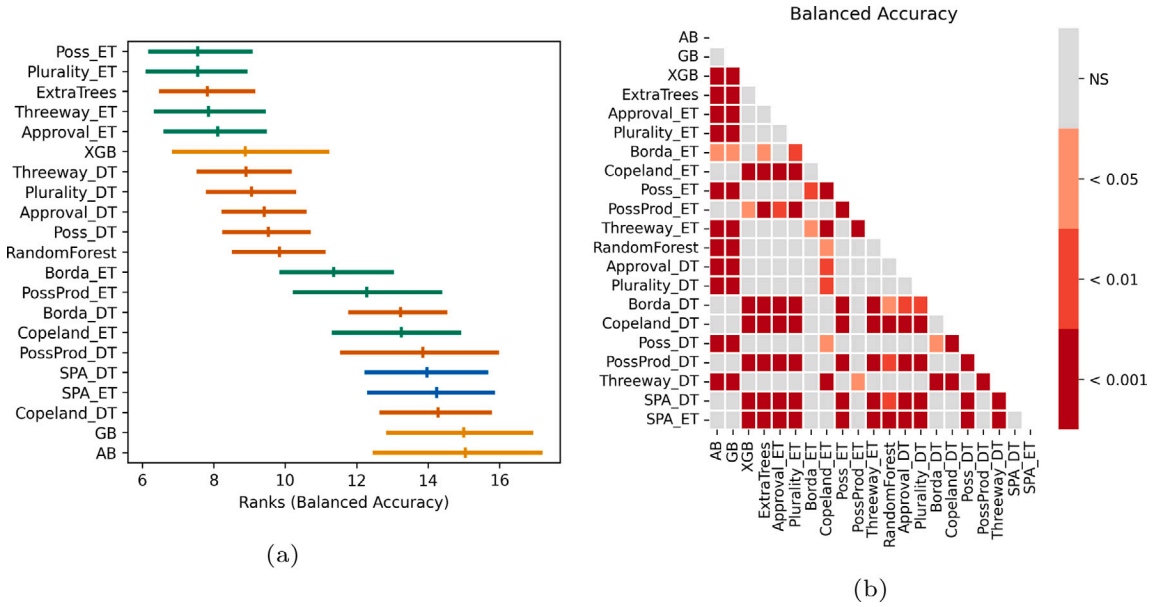


Fig. 2. Comparison of the models under study in terms of balanced accuracy. Left: pointplot of the mean ranks (lower is better), error bars denote 95% C.I. Right: heatmap of p-values obtained with the post-hoc Friedman–Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

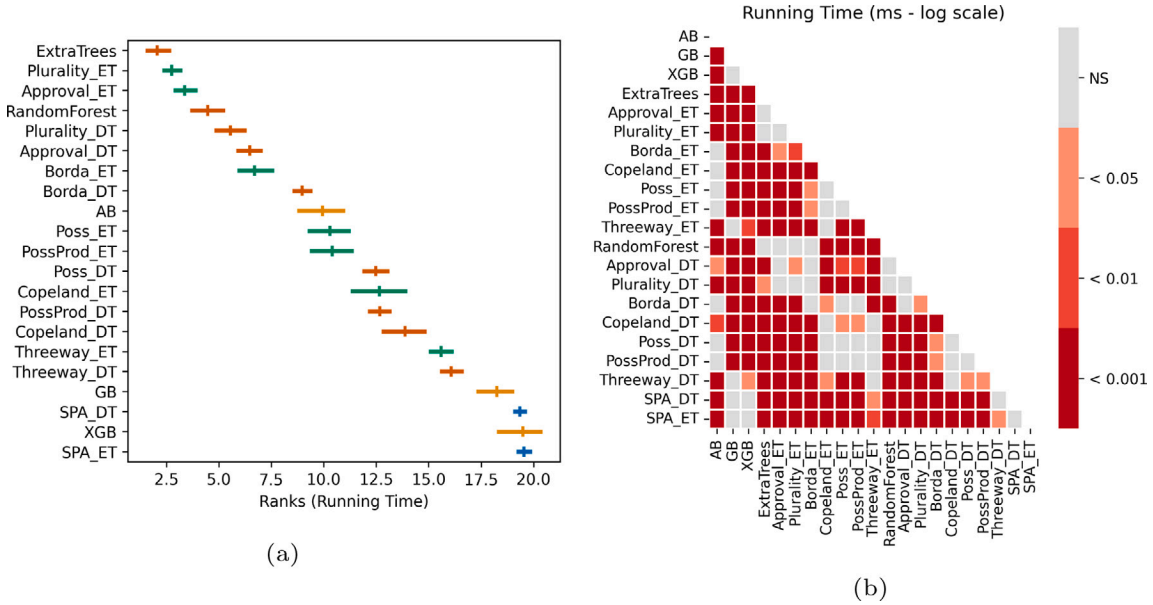


Fig. 3. Comparison of the models under study in terms of running time. Left: pointplot of the mean ranks (lower is better), error bars denote 95% C.I. Right: heatmap of p-values obtained with the post-hoc Friedman–Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(i.e., Newton boosting) and extensive regularization [69] in the XGB implementation: indeed, regularization could be helpful for reducing the effective data-dimensionality (by reducing the weight assigned to redundant or useless features) as well as for reducing overfitting and improving generalization. Nonetheless, despite its outstanding classification performance, XGB was among the worst algorithms in terms of running time, scoring 2nd worst both in terms of mean running time and in terms of mean ranks (w.r.t. running time). In particular, the difference in running time w.r.t. XGB was significant for all algorithms except SPA and GB. In this sense, it is easy to observe that the same characteristics that allow XGB to obtain better classification performance can have a large impact on its computational efficiency: indeed, the use of boosting makes XGB harder to parallelize than Bagging-based methods while the high number of hyper-parameters makes finding an optimal configuration in limited running-time harder than for simpler models.

In contrast with the previous results reported in the literature [23, 29], where SPA was reported as the best-performing ensemble approach, the SPA algorithm was among the worst performing models, both in terms of balanced accuracy and in terms of running time, and independently of the used base classifier. In particular, in terms of balanced accuracy, even though SPA was not significantly different from AB and GB, the differences w.r.t. all Bagging-based approaches,

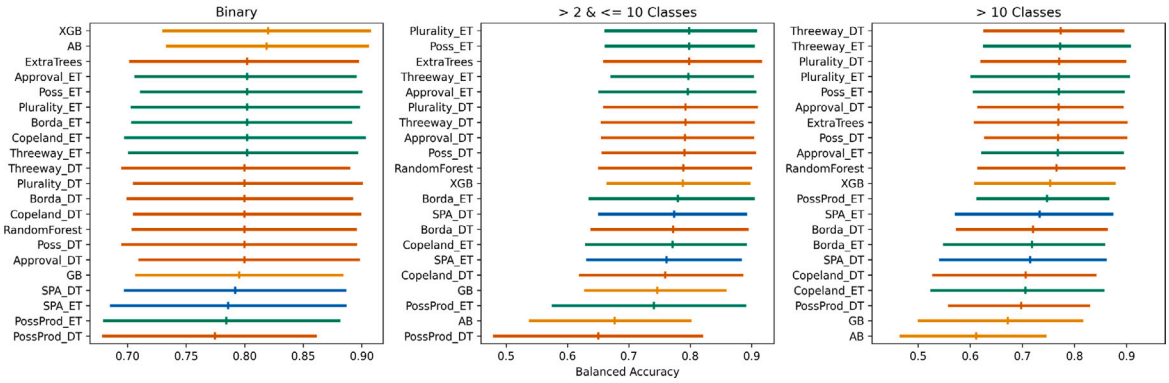


Fig. 4. Balanced accuracy scores of the tested aggregation methods using multiple subsets of possible datasets aggregated by number of classes. Error bars denote 95% C.I.

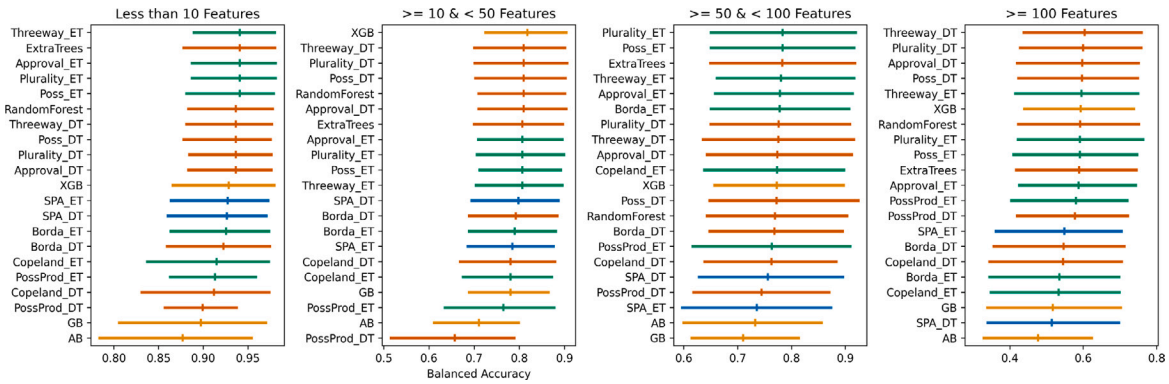


Fig. 5. Balanced accuracy scores of the tested aggregation methods using multiple subsets of possible datasets aggregated by number of features. Error bars denote 95% C.I.

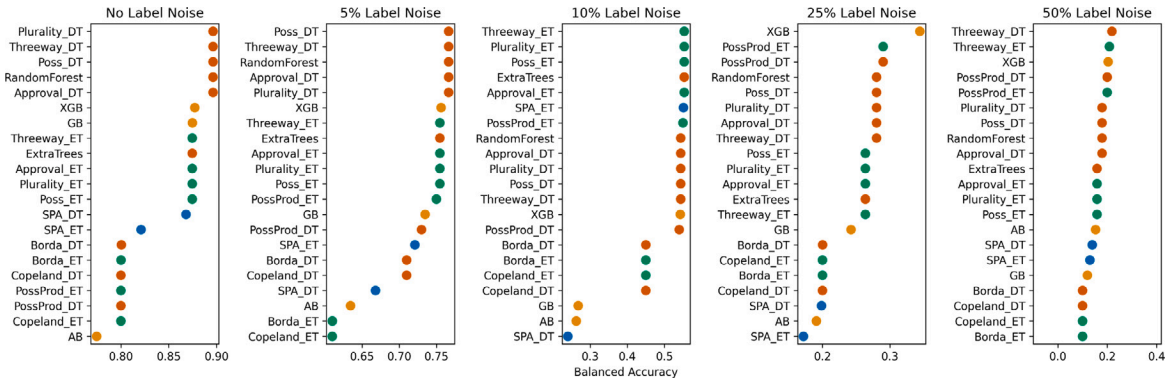


Fig. 6. Balanced accuracy scores of the tested aggregation methods at different levels of label noise, based on the synthetic datasets.

with the exception of the SCT-based ones and PossProd, were statistically significant (see Fig. 2(b)). The poor performance of the SPA algorithm, which in human ensembles often performs better than plurality voting [20], can be related to the overfitting of the models whose target is the ensemble's prediction: indeed, the base models whose target was the ensemble's prediction (i.e., the s_i models in Algorithm 5) did not generalize to the test set, thus failing to improve the overall performance of the algorithm. Moreover, while in human collectives SPA aims at identifying *real experts* by assuming that they can better conjecture the performance of the others, this association is much weaker in the case of ML models: the ability of a model to predict correctly the output of other classifiers in the ensemble does not necessarily entail a lower generalization error and may instead denote a similar overfitting pattern for the majority of the base classifiers in the ensemble. Similarly, also in terms of running time, SPA was significantly worse than all other algorithms excluded GB and XGB (see Fig. 3(b)). The increased time required by the SPA algorithms is due to the need to train twice the number of base models [23,29].

Similarly, also the SCT-based Borda and Coperland aggregation algorithms reported lower performance than other Bagging-based approaches, as highlighted in the rank comparison, see Figs. 2(a) and B.2(a) in Appendix B. In particular, the Copeland rule was significantly out-performed by all other Bagging-based approaches, except PossProd and SPA. Similarly, also the Borda_DT algorithm was significantly out-performed by most other Bagging-based approaches, with the exception of PossProd_DT and PossProd_ET. On the other hand, while also the Borda_ET reported worse performance than other Bagging approaches, most of these differences were not statistically significant: only the differences w.r.t. ExtraTrees, Plurality_ET, Poss_ET, Threeway_ET were significant. The poorer performance of the Borda Count and Copeland Rule approaches cannot be solely explained as being derived by the insufficient informativeness of the ranking information as conjectured

in [23]. Indeed, we note that the SCT-based Plurality (and, particularly so, the ExtraTree-based implementation Plurality_ET) was among the best models, despite being similarly based solely on ranking information. A possible explanation for this behavior is reported in Figs. 4, 5 and 6. Indeed, we easily note that the Borda Count and Copeland Rule approaches were the less stable among the Bagging-based approaches w.r.t. increasing number of features, classes or label noise. This observation also provides an indication about the observed discrepancy w.r.t. the previous results in [24,25,32], which were mostly focused on datasets with relatively low dimensionality and small number of classes. Then, the observed good performance of Plurality in comparison with Borda and Copeland could be explained by considering the different scoring functions adopted by the three methods. Indeed, while Plurality only scores the top-ranked class label, both Borda and Copeland use all ranking information. As shown in Fig. 4, the difference between the three methods becomes particularly significant in multi-class settings. As can be easily derived from basic results in probability and voting theory (see e.g. Condorcet jury theorem [70]), since Plurality only assigns a score to the top-ranked alternative, when multiple classifiers report the same top-ranked alternative, this is likely to be the correct class label. By contrast, Borda and Copeland use the full ranking information and the scoring functions they employ may tend to smooth-out [71] the score of the correct class when there are many classes. Furthermore, we also note that the Copeland Rule was among the worst-performing algorithms in terms of running time. This increased running time can be easily explained by observing that computing the Copeland score has time complexity $O(n \cdot |Y|^2)$ [8], where n is the number of base models.

All other approaches, which were all Bagging-based, reported similar performance. While in terms of average balanced accuracy the Decision Tree-based approaches reported higher performance than the ExtraTree-based ones (see Fig. 1(a)), this difference was not statistically significant and was not observed in the rank-based analysis (see Fig. 2(a)). Indeed, in terms of the rank-based analysis the five best performing algorithms were all ExtraTree-based. The improved performance of ExtraTree-based methods can be understood as stemming from the regularization effect due to the random selection of split thresholds in the training of the base classifiers, which may help in reducing overfitting as compared to the use of CART-based Decision Trees. This finding confirms and generalizes the previous results in [31], obtained in the setting of *binary* classification, by which ExtraTree base classifiers were found to be preferable to CART trees as base classifiers in EL, due to the similar classification performance for the two methods, with ExtraTree-based methods having however lower computational costs. Our results confirm and extend this analysis also to the multi-class setting: for all aggregation methods, the ExtraTree-based variant reported a performance similar to (or better than) the corresponding Decision Tree-based one, while being also more computationally efficient. Furthermore, the default ExtraTrees ensemble method was among the 5 best performing algorithms (2nd best when unifying the base classifiers, see Fig. B.2(a) in Appendix B) and the best in terms of running time.

Interestingly, both in terms of balanced accuracy and ranks, the best performing approaches were IF-UM-based: Threeway_DT and Poss_ET, respectively, with the Threeway aggregation method being the general best aggregation algorithm in the comparison obtained by unifying the base classifiers (see Figs. B.1(a) and B.2(a) in Appendix B). More in general, the IF-UM-based approaches ranked consistently among the models with the best performance: indeed, both in terms of balanced accuracy (Threeway_DT, Threeway_ET, Poss_ET) as well as ranks (Poss_ET, Threeway_ET, Approval_ET), 3 of the 5 best performing methods were IF-UM-based. This observation confirms the previous results in [23] and shows that label smoothing regularization [53] and cautious inference mechanisms could be useful in the development of effective EL methods. The effectiveness of label smoothing regularization can be observed more clearly in the additional analyses, as reported in Figs. 4, 5, and 6. Indeed, the IF-UM-based methods were among the most robust to both increasing dimensionality, number of classes and, most interestingly, label noise. This is particularly evident for PossProd, which while being among the worst methods in the general analysis, was surprisingly resistant to label noise, showing (in the worst case of 50% label noise) performance on par with three of the best performing methods (namely, XGB, Threeway_ET and Threeway_DT). Nonetheless, despite these strongly positive results in terms of classification performance, all IF-UM-based approaches but Approval_ET were significantly less computationally efficient than the most efficient algorithm (i.e., ExtraTrees).

Finally, as reported in Figs. 4 and 5, we can observe that the previous observations still hold also for the differentiated analysis. Moreover, as mentioned previously, this latter analysis sheds light on the poor performance of some of the ensemble models (i.e., the Boosting-based methods, SPA, Borda Count and Copeland Rule): If we observe Fig. 4, we note that on the binary datasets all classifiers (except Threeway_AB) reported very similar performances; by contrast, when we increase the number of classes, the ensemble methods which reported poorer performances in the aggregated analysis were affected by a much larger reduction in performance. In particular, the AB and GB Boosting-based approaches reported a decrease in balanced accuracy as large as 30%. By contrast, the Bagging-based approaches (and, particularly, so the IF-UM-based approaches) were significantly more stable w.r.t. increasing the number of classes. Similar observations also hold when we consider the reduction in performance with respect to increasing data dimensionality (see Fig. 5).

Table A.1
Range of hyper-parameters for the evaluated models.

Algorithm	Param.	Values
Adaboost	algorithm	SAMME, SAMME.R
	learning_rate	1.0, 0.5, 0.2, 0.1, 0.05, 0.01, 0.001
	base_estimator	DecisionTreeClassifier
	max_depth[a]	1, 2, 3, 5, 10, 20, 50, 100, *None*
GradientBoosting	learning_rate	1.0, 0.5, 0.2, 0.1, 0.05, 0.01, 0.001
	subsample	1.0, 0.9, 0.75, 0.5
	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
XGBoost	learning_rate	1.0, 0.5, 0.2, 0.1, 0.05, 0.01, 0.001
	subsample	1.0, 0.9, 0.75, 0.5
	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
ExtraTrees	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
	max_features	sqrt, log2
	criterion	gini,entropy
	bootstrap	True, False
	class_weight	balanced
RandomForest	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
	max_features	sqrt, log2
	criterion	gini,entropy
	class_weight	balanced
DecisionTree	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
	max_features	sqrt, log2
	criterion	gini,entropy
	class_weight	balanced
ExtraTree	max_depth	1, 2, 3, 5, 10, 20, 50, 100, *None*
	max_features	sqrt, log2
	criterion	gini,entropy
	class_weight	balanced

[a]This hyper-parameter is related to the DecisionTreeClassifier used as base_estimator.

6. Conclusion

In this article, we have compared different EL aggregation methods, inspired by SCT, CI and IF-UM, by means of a large-scale experimental evaluation based on a highly heterogeneous set of datasets and tasks. We believe that our results can provide useful indications for the application of EL in practical contexts. Thus, in conclusion, we provide

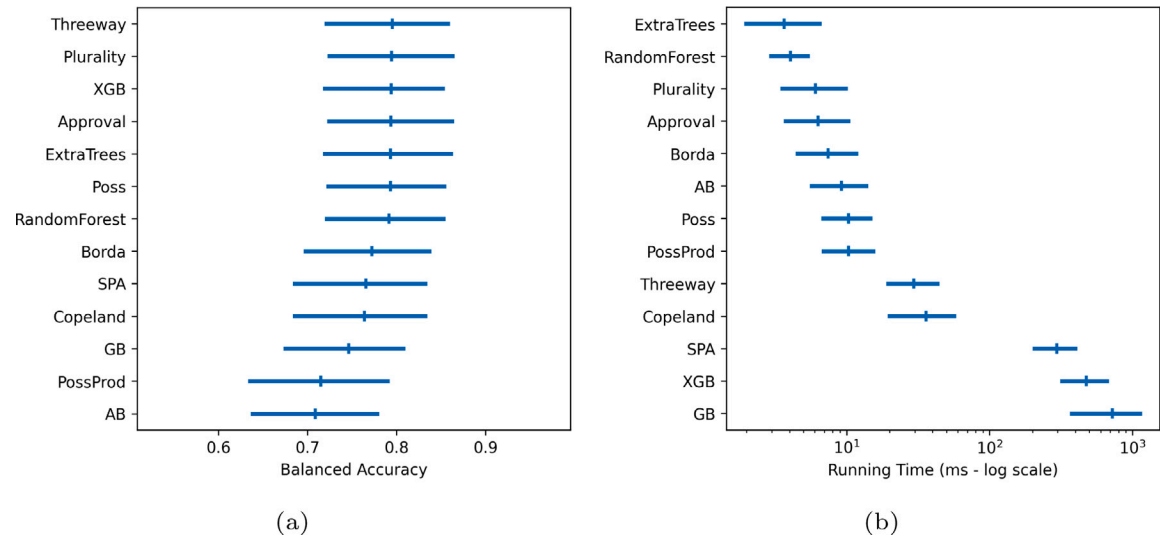


Fig. B.1. Results of the experiments. Left: mean balanced accuracy scores of the models under study (higher is better), Error bars denote 95% C.I. Mean running times (ms) of the models under study (lower is better). Error bars denote 95% C.I.

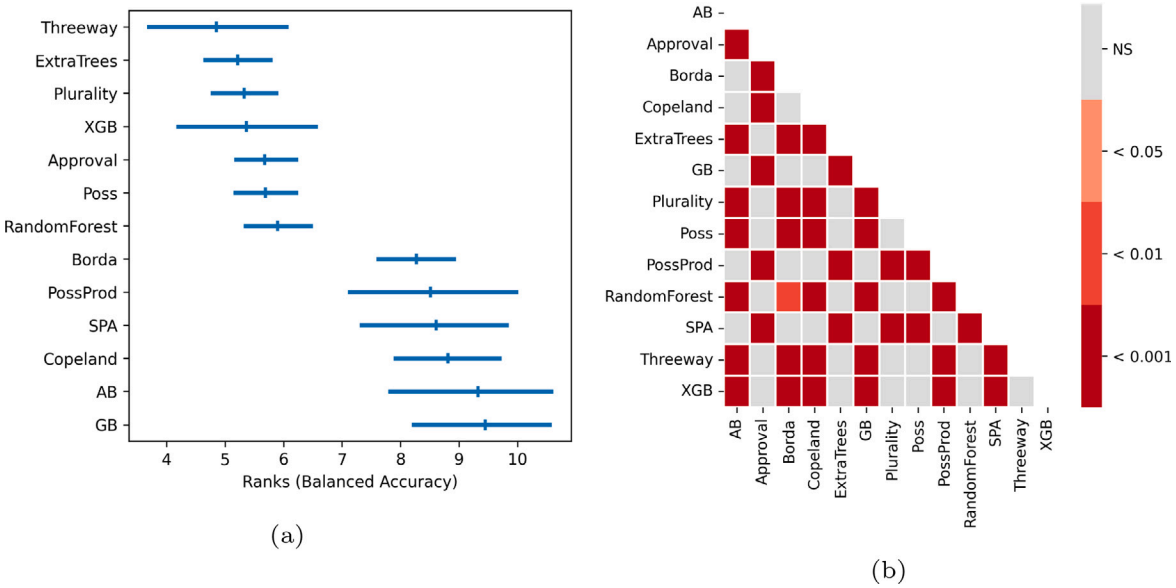


Fig. B.2. Comparison of the models under study in terms of balanced accuracy. Left: pointplot of the mean ranks (lower is better), error bars denote 95% C.I. Right: heatmap of p-values obtained with the post-hoc Friedman–Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank.

Table A.2
List of used datasets. For each dataset, we report the number of classes, features and instances.

Dataset	Classes/Feats./Insts.	Dataset	Classes/Feats./Insts.	Dataset	Classes/Feats./Insts.	Dataset	Classes/Feats./Insts.
20newsgroups	20/1000/11313	data25	10/100/10000	ionosphere	2/33/351	qualitywine	7/11/4898
avila	10/10/20768	data5	10/100/10000	iranian	2/45/7032	robot	4/24/5456
banknote	2/4/1372	data50	10/100/10000	iris	3/4/150	sensorless	11/48/20000
cancer	2/9/683	diabetes	2/8/76	mice	8/78/972	shill	2/9/6321
car	4/16/64	digits	10/62/5620	micromass	20/1300/571	sonar	2/60/208
cargo	3/11/3942	frog-family	4/22/7195	mushroom	6/99/5644	taiwan	2/94/6819
credit	2/61/1000	frog-genus	8/22/7195	myocardial	2/111/1700	thyroid	3/21/7200
crowd	6/28/10845	frog-species	10/22/7195	obesity	7/31/2111	vowel	11/9/990
data0	10/100/10000	hcv	4/12/582	occupancy	2/5/20560	wifi	4/7/2000
data10	10/100/10000	htru	2/8/17898	pen	10/16/10992	wine	3/13/178

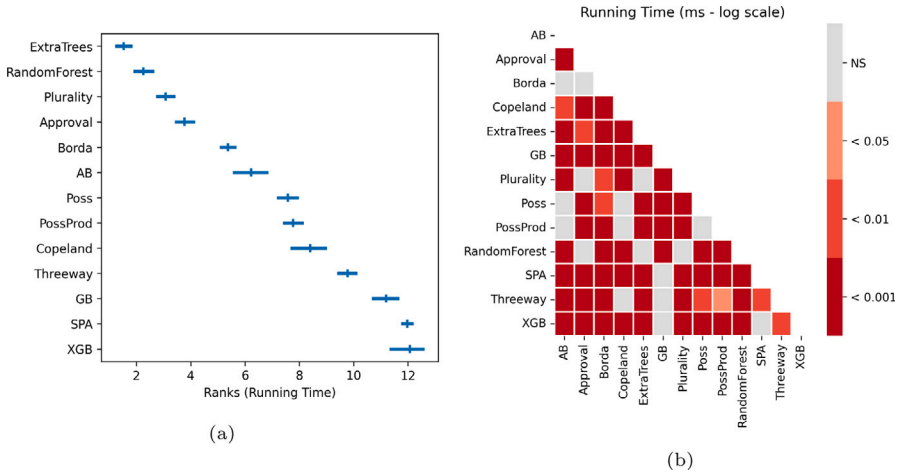


Fig. B.3. Comparison of the models under study in terms of running time. Left: pointplot of the mean ranks (lower is better), error bars denote 95% C.I. Right: heatmap of p-values obtained with the post-hoc Friedman–Nemenyi test, significance at different thresholds is denoted with shades of red. For each significant comparison in the right side, the best method in the corresponding pair of models can be assessed from the left side, by looking at which of the two models had a lower mean rank.

Table B.1
Numerical results of the experiments in terms of both average values and average ranks. Numbers in bold denote the best aggregation or ensemble method for each metric.

Algorithm	Balanced accuracy (rank)	Running time (rank)	Balanced accuracy	Running time
AB	15.04	9.95	0.71	9.21
GB	15.00	18.25	0.75	725.85
XGB	8.75	19.50	0.79	476.97
ExtraTrees	7.83	**2.05**	0.79	**3.66**
Approval_ET	8.11	3.38	0.79	4.43
Plurality_ET	**7.55**	2.75	0.79	4.00
Borda_ET	11.36	6.70	0.77	6.35
Copeland_ET	13.25	12.66	0.76	34.85
Poss_ET	**7.55**	10.30	0.79	8.52
PossProd_ET	12.29	10.40	0.76	8.87
Threeway_ET	7.83	15.60	0.79	28.25
RandomForest	9.84	4.47	0.79	4.06
Approval_DT	9.41	6.46	0.79	6.31
Plurality_DT	9.06	5.55	0.79	6.04
Borda_DT	13.22	8.97	0.77	7.45
Copeland_DT	14.28	13.88	0.76	36.02
Poss_DT	9.53	12.47	0.79	10.32
PossProd_DT	13.85	12.68	0.71	10.36
Threeway_DT	8.75	16.07	**0.80**	29.54
SPA_DT	13.96	19.35	0.77	280.96
SPA_ET	14.24	19.55	0.77	296.20

some guidelines for the selection of ensemble methods, based on the reported results:

- Boosting-based approaches (except XGBoost) were significantly outperformed by all other methods, and were also more sensitive to data dimensionality, number of classes and label noise. Thus, use of these algorithms should be limited to binary settings, in which case they are among the most effective methods;
- While XGBoost was as accurate and as resistant to label noise as the best Bagging-based methods, it was significantly less computationally efficient. Therefore, in general, Bagging-based approaches should be preferred, unless the learning problem is affected by label noise, or computation time is not a limiting factor. Even in these latter cases, however, standard Bagging-based, as well as IF-UM-based, approaches could be preferable;
- ExtraTree-based methods were found to be as accurate as XGBoost and Decision Tree-based ones. Since the former models are more computationally efficient and massively parallelizable, they could be preferable in general settings. In particular, standard ExtraTrees was among the most accurate models as well as the most efficient one, thus we suggest it could be used as an effective baseline ML model;
- IF-UM-inspired approaches reported the best performance among the evaluated models, and were the most resistant to label noise. While less computationally efficient than standard ExtraTrees, we believe that the *label smoothing* and *cautious inference*-inspired regularization of ensemble models should be further investigated as an effective way to improve the performance, especially in settings affected by uncertainty [72].

We believe that future works should evaluate the effectiveness of EL aggregation methods on different data types (e.g., images, text), in different learning tasks (e.g., regression, multi-label classification), and using different base classifiers (e.g. Deep Learning models [73]), as well as their resistance to dataset shifts [74,75] or adversarial perturbations [76].

CRediT authorship contribution statement

Andrea Campagner: Conceptualization, Formal analysis, Investigation, Writing, Visualization. **Davide Ciucci:** Conceptualization, Writing, Supervision. **Federico Cabitza:** Conceptualization, Writing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All used data is publicly available on GitHub at: https://github.com/AndreaCampagner/Aggregation-Models-in-Ensemble-Learning.

Appendix A. Datasets and model settings

See Tables A.1 and A.2.

Appendix B. Additional results

See Table B.1 and Figs. B.1–B.3.

References

[1] O. Sagi, L. Rokach, Ensemble learning: A survey, in: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, 2018, e1249.
[2] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data? 2022, arXiv preprint arXiv:2207.08815.
[3] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[4] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.
[5] A. Bender, D. Rügamer, F. Scheipl, B. Bischl, A general machine learning framework for survival analysis, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 158–173.
[6] M. Rapp, E.L. Mencía, J. Fürnkranz, V.L. Nguyen, E. Hüllermeier, Learning gradient boosted multi-label classification rules, in: Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2021, pp. 124–140.
[7] L. Cheng, Y. Wang, X. Liu, B. Li, Outlier detection ensemble with embedded feature selection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 3503–3512.
[8] F. Brandt, V. Conitzer, U. Endriss, J. Lang, A.D. Procaccia, Handbook of Computational Social Choice, Cambridge University Press, 2016.
[9] S. Suran, V. Pattanaik, D. Draheim, Frameworks for collective intelligence: a systematic literature review, ACM Comput. Surv. 53 (2020) 1–36.
[10] J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Inf. Fusion 14 (2013) 423–430.
[11] M. Barandas, D. Folgado, R. Santos, R. Simão, H. Gamboa, Uncertainty-based rejection in machine learning: Implications for model development and interpretability, Electronics 11 (396) (2022).
[12] A. Campagner, D. Ciucci, C.M. Svensson, M.T. Figge, F. Cabitza, Ground truthing from multi-rater labeling with three-way decision and possibility theory, Inform. Sci. 545 (2021b) 771–790.
[13] P. Toccaceli, A. Gammerman, Combination of conformal predictors for classification, in: Conformal and Probabilistic Prediction and Applications, PMLR, 2017, pp. 39–61.
[14] T. Duan, A. Anand, D.Y. Ding, K.K. Thai, S. Basu, A. Ng, A. Schuler, Ngboost: Natural gradient boosting for probabilistic prediction, in: International Conference on Machine Learning, PMLR, 2020, pp. 2690–2700.
[15] L. Gautheron, P. Germain, A. Habrard, G. Metzler, E. Morvant, M. Sebban, V. Zantedeschi, Landmark-based ensemble learning with random fourier features and gradient boosting, in: Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2021, pp. 141–157.
[16] V. Gómez-Rubio, R.S. Bivand, H. Rue, Bayesian model averaging with the integrated nested laplace approximation, Econometrics 8 (23) (2020).
[17] Y. Bian, H. Chen, When does diversity help generalization in classification ensembles? IEEE Trans. Cybern. (2021).
[18] Y. Wang, H. Zhang, H. Chen, D. Boning, C.J. Hsieh, On lp-norm robustness of ensemble decision stumps and trees, in: International Conference on Machine Learning, PMLR, 2020, pp. 10104–10114.
[19] A. Webb, C. Reynolds, W. Chen, H. Reeve, D. Iliescu, M. Luján, G. Brown, To ensemble or not ensemble: When does end-to-end training fail? in: Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 109–123.
[20] D. Prelec, H.S. Seung, J. McCoy, A solution to the single-question crowd wisdom problem, Nature 541 (2017) 532–535.
[21] V.N. Balasubramanian, S. Chakraborty, S. Panchanathan, Conformal predictions for information fusion, Ann. Math. Artif. Intell. 74 (2015) 45–65.
[22] J.F. Laslier, And the loser is… plurality voting, in: Electoral Systems, Springer, 2012, pp. 327–351.
[23] A. Campagner, D. Ciucci, F. Cabitza, Ensemble learning, social choice and collective intelligence, in: Modeling Decisions for Artificial Intelligence, Springer, 2020, pp. 53–65.
[24] C. Cornelio, M. Donini, A. Loreggia, M.S. Pini, F. Rossi, Voting with random classifiers (vorace): theoretical and experimental analysis, Auton. Agents Multi-Agent Syst. 35 (2021) 1–31.
[25] F. Leon, S.A. Floria, C. Bădică, Evaluating the effect of voting methods on ensemble-based classification, in: 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA, IEEE, 2017, pp. 1–6.
[26] D.M. Pennock, P. Maynard-Reid II, C.L. Giles, E. Horvitz, A normative examination of ensemble learning algorithms, in: International Conference on Machine Learning, 2000, pp 735–742.
[27] D. Ruta, B. Gabrys, Classifier selection for majority voting, Inf. Fusion 6 (2005) 63–81.
[28] I. Gandhi, M. Pandey, Hybrid ensemble of classifiers using voting, in: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), IEEE, 2015, pp. 399–404.
[29] T. Luo, Y. Liu, Machine truth serum, 2019, arXiv preprint arXiv:1909.13004.
[30] C.A. Shipp, L.I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, Inf. Fusion 3 (2002) 135–148.
[31] A. Narassiguin, M. Bibimoune, H. Elghazel, A. Aussem, An extensive empirical comparison of ensemble learning methods for binary classification, Pattern Anal. Appl. 19 (2016) 1093–1128.
[32] K.T. Leung, D.S. Parker, Empirical comparisons of various voting methods in bagging, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 595–600.
[33] Z.H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012.
[34] R.E. Schapire, Y. Freund, Boosting: Foundations and algorithms, Kybernetes (2013).
[35] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259.
[36] J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian model averaging: a tutorial, Statist. Sci. (1999) 382–401.
[37] A. Omari, A.R. Figueiras-Vidal, Post-aggregation of classifier ensembles, Inf. Fusion 26 (2015) 96–102.
[38] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1997) 119–139.
[39] T.K. Ho, Random decision forests, in: Proceedings of the 3rd International Conference on Document Analysis and Recognition, IEEE, 1995, pp. 278–282.
[40] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: A new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1619–1630.
[41] G. Louppe, P. Geurts, Ensembles on random patches, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2012, pp. 346–361.
[42] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[43] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Front. Comput. Sci. 14 (2020) 241–258.
[44] Y. Grandvalet, Bagging can stabilize without reducing variance, in: International Conference on Artificial Neural Networks, Springer, 2001, pp. 49–56.
[45] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[46] M. Kearns, L. Valiant, Cryptographic limitations on learning boolean formulae and finite automata, J. ACM 41 (1994) 67–95.
[47] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[48] X. Lin, S. Yacoub, J. Burns, S. Simske, Performance analysis of pattern classifier combination by plurality voting, Pattern Recognit. Lett. 24 (2003) 1959–1969.
[49] J. Fraenkel, B. Grofman, The borda count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia, Aust. J. Political Sci. 49 (2014).
[50] A.H. Copeland, A 'reasonable' social welfare function, 1951, Unpublished.
[51] S. Brams, P.C. Fishburn, Approval Voting, Springer Science & Business Media, 2007.
[52] E. Chzhen, C. Denis, M. Hebiri, T. Lorieul, Set-valued classification–overview via a unified framework, 2021, arXiv preprint arXiv:2102.12318.
[53] J. Lienen, E. Hüllermeier, From label smoothing to label relaxation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 8583–8591.
[54] M. Lukasik, S. Bhojanapalli, A. Menon, S. Kumar, Does label smoothing mitigate label noise? in: International Conference on Machine Learning, PMLR, 2020, pp. 6448–6458.
[55] A. Campagner, F. Cabitza, P. Berjano, D. Ciucci, Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches, Inform. Sci. 579 (2021a) 347–367.
[56] G. Cherubin, Majority vote ensembles of conformal predictors, Mach. Learn. 108 (2019) 475–488.
[57] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: Rough Sets and Knowledge Technology, Springer, 2009, pp. 642–649.

[58] D. Dubois, H. Prade, Possibility Theory: An Approach To Computerized Processing of Uncertainty, Springer Science & Business Media, 2012.
[59] L. Carlsson, M. Eklund, U. Norinder, Aggregated conformal prediction, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2014, pp. 231–240.
[60] T. Denoeux, Logistic regression, neural networks and dempster–shafer theory: A new perspective, Knowl.-Based Syst. 176 (2019) 54–67.
[61] D. Dua, C. Graff, UCI machine learning repository, 2017, URL http://archive.ics.uci.edu/ml.
[62] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (1937) 675–701.
[63] P. Nemenyi, Distribution-Free Multiple Comparisons (Ph.D. thesis), Princeton University, 1963.
[64] A. Vezhnevets, O. Barinova, Avoiding boosting overfitting by removing confusing samples, in: European Conference on Machine Learning, Springer, 2007, pp. 430–441.
[65] J. Bootkrajang, A. Kabán, Boosting in the presence of label noise, 2013, arXiv preprint arXiv:1309.6818.
[66] D. Nielsen, Tree boosting with XGBoost, in: Why Does XGBoost Wins "Every" Machine Learning Competition? (Master's thesis), Norwegian University of Science and Technology, 2016.
[67] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Inf. Fusion (2021).
[68] A. Gómez-Ríos, J. Luengo, F. Herrera, A study on the noise label influence in boosting algorithms: Adaboost, gbm and xgboost, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2017, pp. 268–280.
[69] R.K. Vinayak, R. Gilad-Bachrach, Dart: Dropouts meet multiple additive regression trees, in: Artificial Intelligence and Statistics, PMLR, 2015, pp. 489–497.
[70] P.J. Boland, Majority systems and the condorcet jury theorem, J. Roy. Statist. Soc. Ser. D 38 (1989) 181–189.
[71] W.V. Gehrlein, Condorcet's paradox and the likelihood of its occurrence: different perspectives on balanced preferences, Theory and Decision 52 (2002) 171–199.
[72] B. Han, Q. Yao, T. Liu, G. Niu, I.W. Tsang, J.T. Kwok, M. Sugiyama, A survey of label-noise representation learning: Past, present and future, 2020, arXiv preprint arXiv:2011.04406.
[73] S. Sinha, H. Bharadhwaj, A. Goyal, H. Larochelle, A. Garg, F. Shkurti, Dibs: Diversity inducing information bottleneck in model ensembles, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 9666–9674.
[74] P. Pérez-Gállego, J.R. Quevedo, J.J. del Coz, Using ensembles for problems with characterizable changes in data distribution: A case study on quantification, Inf. Fusion 34 (2017) 87–100.
[75] A. Ross, W. Pan, L. Celi, F. Doshi-Velez, Ensembles of locally independent prediction models, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 5527–5536.
[76] F. Ranzato, M. Zanella, Abstract interpretation of decision tree ensemble classifiers, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 5478–5486.

Evidential Predictors: Evidential Combination of Conformal Predictors for Multivariate Time Series Classification

Andrea Campagner[1], **Marília Barandas**[2], **Duarte Folgado**[2], **Hugo Gamboa**[2,3], **Federico Cabitza**[1,4]

[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[2] Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, 4200-135, Porto, Portugal
[3] Laboratório de Instrumentação, Engenharia Biomédica e Física da Radiação (LIBPhys-UNL), Departamento de Física, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516, Caparica, Portugal
[4] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

Abstract

In this article we propose a conceptual framework to study ensembles of conformal predictors, inspired by the application of imprecise probabilities in information fusion, that we call Evidential Predictors. We study their theoretical properties, focusing on four combination rules, namely: the minimum, maximum, weighted mean and Dempster's rules of combination. We also illustrate the applicability of the proposed methods in the setting of multivariate time-series classification, showing that these methods provide better performance (in terms of both accuracy and efficiency) than both standard classification algorithms and other combination rules proposed in the literature, on a large set of benchmarks from the UCR time series archive.

Introduction

Multivariate time series classification (MTSC) refers to the task in which the instances to be classified are represented as multi-dimensional vectors of time series. This type of task has become increasingly more relevant in practical contexts, due to the widespread use of data sources and sensors in various applications ranging from physical rehabilitation (Pereira et al. 2019) to activity recognition (Barandas et al. 2022), from clinical data monitoring (Song et al. 2018) to control systems (Susto, Cenedese, and Terzi 2018).

While *bespoke methods*, i.e. approaches that directly operate on the multivariate data representation, have been considered in the literature with promising results (Ruiz et al. 2021), their applicability may be limited due to their inherent complexity and limited evaluation as compared to state-of-the-art univariate algorithms, as well as due to their inappropriateness in settings like multi-source or edge computing-based learning (Pantiskas et al. 2022).

A simple, but remarkably effective, approach to address these limitations is to train univariate models on the one-dimensional time series. The predictions provided by these classifiers, which are equated to information hints about the same object provided by different and potentially unreliable sources, are then combined into a single one by applying information fusion (IF) methods (Ruiz et al. 2021). To this aim, several methods have been considered, including approaches based on Dempster-Shafer (DS) theory (Jin et al. 2021) and possibility theory (Le Carrer and Ferson 2021) as well as more traditional ensemble learning techniques (Dhariyal et al. 2020). Such combinations methods have shown performance comparable with state-of-the-art bespoke models (Ruiz et al. 2021) together with increased interpretability (Ismail et al. 2020).

While the aforementioned methods have been extensively used over the last several years, their properties are not well understood and, in general, they do not satisfy desirable validity or calibration guarantees. This issue is particularly relevant in critical domains, where reliability guarantees that go beyond empirical validity can be desirable (Kompa, Snoek, and Beam 2021).

For this reason, in the recent years several studies have investigated the application of the Conformal Prediction (CP) framework (Vovk, Gammerman, and Shafer 2005) to the analysis of time series (Xu and Xie 2021; Zaffran et al. 2022). CP provides a set of techniques that can be applied to improve the accuracy and calibration of any underlying model with provable guarantees on the aforementioned properties that hold under weak assumptions on the data-generating distribution. Existing approaches, however, focus mainly on the forecasting setting and cannot thus be directly applied to the problem of MTSC.

In analogy with our previous discussion concerning the application of univariate methods to MTSC, methods for aggregating base CPs (Linusson, Johansson, and Boström 2020) have recently been studied, so as to broaden the applicability of CP methods to scenarios where constructing global CP may be unfeasible. In this sense, two main settings have been considered: the ensemble learning (or, resampling-based) setting (Linusson et al. 2017) and the more general information fusion (IF)[1] one (Balasubramanian, Chakraborty, and Panchanathan 2015). In this latter setting, in particular, the data used to generate the base CPs to be aggregated are not assumed to be related and may come from different sources whose dependency structure is not known a priori. It is then

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] More specifically, the IF setting focuses on the *late fusion* problem, i.e. the problem of combining the predictions of several models that are separately trained on each of the different sources: in the CP setting, this corresponds to the case of combining the p-value functions for the single CPs. This is different and of independent interest to the *early fusion* problem, where the combination is performed before the application of CP.

clear that the IF setting could be particularly interesting for two main reasons: 1) as a generalization of the resampling-based setting whose theoretical study could help clarifying the properties of ensembles of CP 2) more relevantly, in regard to its application to MTSC tasks. Nonetheless, compared to the ensemble learning setting, the IF setting (including its application to MTSC) has received much less attention.

Drawing from MTSC as source of inspiration, and grounding on recent works relating CP and DS theory (Cella and Martin 2021), in this work we propose and study Evidential Predictors (EP), a conceptual framework to study methods for aggregating CPs inspired by IF rules proposed in DS and related theories, and we illustrate their application to MTSC tasks. Thus, our main contributions to the study of CP and their applications are as follows: 1) Compared to previous works studying aggregation of CPs in the IF setting, we do not assume the independence of the CPs to be combined, which is unrealistic for the MTSC setting (Dubois et al. 2020); instead, we adopt a flexible approach based on copulas (Nelsen 2007) to model their dependence structure; 2) To the best of our knowledge, we provide the first theoretical analysis of combination methods for CP in the IF setting, studying conditions for validity of a wide selection of aggregation methods as well as providing analytical formula to compute their theoretical error rate; 3) Focusing on the MTSC task as a practically relevant and paradigmatic example of IF tasks, we provide a a comprehensive empirical validation of the proposed approach, based on a large set of standard benchmarks from the UEA/UCR archive (Dau et al. 2018; Bagnall et al. 2018), showing that EPs significantly out-perform other CP combination methods as well as state-of-the-art MTSC algorithms.

Background and Related Work

Let X and Y be two sets, and let Pr be an exchangeable distribution defined on Z^*[2], where $Z = X \times Y$. Let $f : Z^* \times Z \mapsto [0,1]$ be a permutation invariant function, that is a function satisfying $f(\mathbb{z}) = f(\pi(\mathbb{z}))$ for every $\mathbb{z} \in Z^*$ and permutation π[3]. In the literature, f is usually called a *non-conformity measure*. A (smoothed) conformal predictor (CP) (Vovk, Gammerman, and Shafer 2005) is a function $\Gamma_f : Z^* \times X \times [0,1] \mapsto 2^Y$ satisfying $\Gamma^\epsilon_f(B,x) = \{y \in Y : p^y > \epsilon\}$, where $B \in Z^*$ is a bag of examples and:

$$p^y = \frac{|\{i : f(B_i, z_i) > f(B,(x,y)\}|}{n+1} + \theta\frac{|\{i : f(B_i, z_i) = f(B,(x,y)\}|}{n+1} \quad (1)$$

with $B_i = B \setminus \{z_i\} \cup \{(x,y)\}$ and θ is a random variable. If θ is constant, then any CP is *conservatively valid*, i.e. $Pr(y \notin \Gamma^\epsilon_f(B,x)) \leq \epsilon$. Furthermore, if $\theta \sim U[0,1]$, then any CP is *strongly valid*, in the sense that $Pr(y \notin \Gamma^\epsilon_f(B,x)) = \epsilon$. Intuitively, (strong) validity means that the confidence level $1-\epsilon$ associated with $\Gamma^\epsilon_f(B,x)$ can be interpreted as the probability that the true label y lies in the set.

The function $p : Z^* \times Z \mapsto [0,1]$ is called *p-value function*. In the article, we will only consider (base) CPs whose p-value function is *normalized*, i.e. satisfies $\forall B \in Z^*, x \in X, \exists y \in Y$ s.t. $p^y = 1$[4]. Furthermore, since our results can be easily shown to hold both in the online and inductive split CP settings without distinction, we will adopt an abstract definition that encompasses both the above mentioned settings.

As a recent topic attracting considerable theoretical interest, increasing attention has been devoted to the question of how to combine multiple CPs (Toccaceli and Gammerman 2017, 2019), and in particular their p-value functions, as well as the validity of such combination methods (Linusson 2021). We note that while the interest in this topic within the setting of CP is recent, the question of how to aggregate p-value functions is not, and in fact has been studied extensively in the fields of multiple hypothesis testing (Loughin 2004) and statistical meta-analysis (Guerra et al. 1999), where several combination methods have been studied (Loughin 2004). In the CP setting, the combination issue amounts to, given n CPs $\Gamma_1, ..., \Gamma_n$, asking two questions: 1) How to combine $\Gamma_1, ..., \Gamma_n$ so as to obtain a new CP $\hat{\Gamma}$?; 2) What are the properties (w.r.t. validity and efficiency) of the combined CP?

Drawing from the above mentioned statistical literature, these questions have mainly focused on the late-fusion scenario, where combination is performed at the p-value functions level, and have been investigated in two main settings. In the ensemble learning setting, the bags $B_1, ..., B_n$ are obtained from a single bag B via re-sampling: examples of this setting include cross-conformal prediction (Vovk 2015) and aggregated conformal prediction (Carlsson, Eklund, and Norinder 2014). By contrast, in the information fusion (IF) setting, formally introduced in (Balasubramanian, Chakraborty, and Panchanathan 2015), the bags $B_1, ..., B_n$ are not assumed to be related in any way and may be obtained from differently distributed data or different sources. Thus, the IF setting can be considered both a generalization of, and more complex than, the ensemble learning one.

In the former setting, (Carlsson, Eklund, and Norinder 2014) studied aggregated CPs, which are defined by the averaged p-value function $\hat{p}^y = \frac{1}{n}\sum_i p^y_i$, showing that the resulting predictors are valid as long as $B_1, ..., B_n$ are obtained by means of a resampling procedure that is *consistent* w.r.t. f. However, it is not clear whether (and how) consistent resampling procedures can be obtained in practice, and indeed aggregated CPs generally fail to be valid (Linusson et al. 2017). Similar results hold also for other resampling-based aggregation schemes, such as out-of-bag CP (Linusson, Johansson, and Boström 2020). Different combination methods were also proposed in (Cherubin 2019), based on majority voting, and (Toccaceli and Gammerman 2017, 2019), based on either learning methods or approaches inspired by meth-

[2]Given a set A, A^* is the collection of all finite sequences in A, that is $A^* = A \cup (A \times A) \cup \ldots$.

[3]Given a tuple $t = (z_1, ..., z_n)$, is a function $\pi : Z^n \mapsto Z^n$ that permutes the elements in t.

[4]This assumption is not too restrictive: let $h : X \times Y \mapsto [0,1]$ be a scoring classifier. Then the non-conformity score $f(B,(x,y)) = \max_{y' \in Y} h(x,y') - h(x,y)$ satisfies the above mentioned property. In any case, as will be clarified later, this assumption is required to hold only for the base CPs and not necessarily for their combination.

ods for combination of p-values.

In the IF setting, by contrast, (Balasubramanian, Chakraborty, and Panchanathan 2015) and (Spjuth et al. 2019) evaluated the application of different aggregation rules, inspired by either IF or techniques inspired by multiple hypothesis testing. In particular, studied techniques were mostly based on order statistics (Davidov 2011), and in particular the minimum and the maximum, or quantile combination methods (Zaykin et al. 2007). While in preliminary evaluations the latter approaches reported promising results, however, in further studies (Linusson 2021) the same methods were shown to violate validity in more comprehensive benchmarks. As an additional limitation, we note that previous work in the IF setting assumed the CPs to be combined (and the corresponding p-value functions) to be independent, an assumption which is not realistic in IF tasks (Dubois et al. 2016, 2020). Furthermore, the theoretical properties of these techniques have not been studied.

In the rest of this article, we will focus on the IF setting since, as described in the Introduction, it is more relevant to the study of MTSC tasks. We will study, in particular, approaches for combining CPs that are inspired by the application of DS and possibility theory to IF (Dubois et al. 2020): for this reason, here we recall the basic notions of these theories. Let (X, Ω) be a pair where X is a set and Ω is a σ-algebra over X. A mass function (Dempster 1967) is a function $m : \Omega \mapsto [0, 1]$ s.t. $\sum_{A\in\Omega} m(A) = 1$. Given a mass function m, we can define two set function, namely the belief and plausibility measures (Shafer 1976):

$$Bel(A) = \sum_{B\subseteq A} m(B), \quad Pl(A) = \sum_{B\cap A\neq\emptyset} m(B). \quad (2)$$

The focal sets of m are defined as $\mathcal{F}(m) = \{A \in \Omega : m(A) > 0\}$. A mass function is *consonant* if $\forall A, B \in \mathcal{F}(m)$ either $A \subseteq B$ or $B \subseteq A$: in this case, the plausibility Pl is a possibility measure (Dubois and Prade 1988). More in general, given any mass function, its *contour function*, which is defined as $pl_m(x) = Pl(\{x\})$ for every $x \in X$, is always guaranteed to be a possibility measure and can be used as a consonant approximation of m, usually called the *consonant projection* (Dubois and Prade 1990). A consonant mass function m is normalized iff $\exists x \in S$ s.t. $pl_m(x) = 1$. DS theory (and imprecise probabilities more in general) has been widely applied in the field of IF, and several combination rules have been proposed (Dubois et al. 2020). Let m_1, m_2 be two mass functions, Dempster's combination rule is defined as:

$$m_1 \oplus m_2(C) = \frac{1}{1-k} \sum_{A\cap B=C} m_1(A)m_2(B), \quad (3)$$

where $k = \sum_{A\cap B=\emptyset} m_1(A)m_2(B)$ is the degree of conflict. We note that the Dempster's combination of two mass function m_1, m_2 is not necessarily consonant (Dubois and Prade 1988), not even if m_1, m_2 themselves are consonant or if $\oplus$ is applied to their consonant projections. The following rules, which are defined to operate on the consonant projections of m_1, m_2 by contrast, are guaranteed to give a consonant mass function as a result:

$$\max\{pl_1(x), pl_2(x)\}, \quad (4)$$

$$\min\{pl_1(x), pl_2(x)\}, \quad (5)$$

$$w * pl_1(x) + (1 - w) * pl_2(x), \quad (6)$$

for every $w \in [0, 1]$. Indeed, the result of any such operation, when applied point-wise to possibility measures, is a possibility measure and hence the contour function of a consonant mass function[5].

The following result, due to (Cella and Martin 2021), shows that CP p-value functions can be understood as a special instance of consonant plausibility measures:

Theorem 1 ((Cella and Martin 2021)). *Let Γ_f be a CP s.t. the associated p-value function p is normalized. Then, p is the contour function of a consonant plausibility measure.*

Finally, we recall the basic notions on copula theory. A *n-dimensional copula* (Nelsen 2007) is a function $C : [0, 1]^n \mapsto [0, 1]$ satisfying: 1) $\forall i \in [n]$, $C(u_1, ..., u_{i-1}, 0, u_{i+1}, u_n) = 0$; 2) $\forall i \in [n], \forall u_i \in [0, 1], C(1, ..., 1, u_i, 1, 1) = u_i$; 3) C is d-non-decreasing.

Copulas are useful as they can represent a joint cumulative distribution function (CDF) F given only its marginals F_i:

Theorem 2 ((Sklar 1959)). *Let $F : X^n \mapsto [0, 1]$ be a joint CDF whose marginals $F_i : X \mapsto [0, 1]$ are continuous. Then, it exists a (unique) copula C_F, s.t.*

$$F(X_1 \leq x_1, ..., X_n \leq x_n) = C_F(F_1(x_1), ...F_n(x_n)).$$

Notably, copulas have recently been applied in the context of CP, see e.g. (Messoudi, Destercke, and Rousseau 2021, 2022) to study multivariate aspects in regression problems, focusing in particular on the problem of multi-target regression (Johnstone and Cox 2021) and multi-label classification (Cauchois, Gupta, and Duchi 2021). Differently from our work, we note that these approaches used copulas to study the joint dependency structure of the nonconformity scores and obtain a corresponding, single confidence level ϵ: thus, this setting is more akin to the early fusion problem mentioned above. By contrast, in the following, copulas will be used to study the joint dependency structure of the p-value functions for different CPs to be combined, in the late fusion setting typically considered in the application of CP to IF.

We also recall the Fréchet–Hoeffding copula bounds:

Proposition 1 ((Genest et al. 1999)). *For any given copula C, $W \leq C \leq M$, where $W(u_1, ..., u_n) = \max\{1 - d\sum u_i, 0\}$ and $M(u_1, ..., u_n) = \min\{u_1, ..., u_n\}$.*

Finally, the *independence* copula is defined as $I(u_1, ..., u_n) = \Pi_i u_i$, i.e. $F_1, ..., F_n$ are independent.

Methods

In light of the correspondence between p-value function and consonant plausibility measures established by means of Theorem 1, a remarkable consequence is that IF methods proposed in DS and related theories can be applied to the aim of

[5]For the min and weighted average operators the combined mass function is generally not normalized. As mentioned previously, we require that normalization applies to the base CPs to be ensembled, and not necessarily so to the result of their combination.

combining CPs. In this section we introduce Evidential Predictors, that is a class of combination rule for CPs inspired by the application of DS theory and possibility theory to IF tasks. More in detail, we will consider EPs based on the *min*, *max*, *weighted mean* and *Dempster's combination* rules. While some of these rules have already been proposed in the CP setting (Balasubramanian, Chakraborty, and Panchanathan 2015), such as the min and max rules, as we mentioned previously, their theoretical properties have not been studied.

Let $\Gamma_1, ..., \Gamma_n$ be n CPs[6,7]. Given a collection of bags B_i, for each CP Γ_i, and an instance x, let p_i be the p-value function associated with $\Gamma_i(B_i, x)$[8]. In the following, we will study the properties of four different EPs:

$$\Gamma_{\min}(x, \epsilon) = \{y \in Y : \min\{p_i^y\} \geq \epsilon\}, \quad (7)$$

$$\Gamma_{\max}(x, \epsilon) = \{y \in Y : \max\{p_i^y\} \geq \epsilon\}, \quad (8)$$

$$\Gamma_{\mathbf{w}}(x, \epsilon) = \{y \in Y : \sum w_i * p_i^y \geq \epsilon\}, \quad (9)$$

$$\Gamma_D(x, \epsilon) = \{y \in Y : p^y_{m_1 \oplus ... \oplus m_n} \geq \epsilon\}, \quad (10)$$

where $p_{m_1 \oplus ... \oplus m_n}$ is the contour function of the consonant approximation (Dubois and Prade 1990) of $m_1 \oplus ... \oplus m_n$, and $\mathbf{w} \in [0,1]^n$ s.t. $\sum w_i = 1$.

In contrast to the previous literature, the approach we adopt in our analysis does not assume independence, but rather grounds on copula theory. This allows to model in a compact way the dependency structure of of the problem, and further provides two additional advantages. First, it allows to clearly characterize the assumptions needed to ensure validity in terms of properties of the defining copula. Second, assuming a specific family of copulas, the dependence structure can be inferred from the observed error rates (e.g., by maximum likelihood estimation or generative learning methods (Ling, Fang, and Kolter 2020; Ng et al. 2021)), allowing to check for inconsistencies between theoretical and empirical validity.

Given $\Gamma_1, ..., \Gamma_n$, let Q be the joint CDF given by $Q(\epsilon_1, ..., \epsilon_n) = Pr(\forall i, p_i^y \leq \epsilon_i) = Pr(\forall i, y \notin \Gamma_i^{\epsilon_i})$, with continuous uniform marginals $Q_i(\epsilon_i) = Pr(p_i^y \leq \epsilon_i) = Pr(y \notin \Gamma_i^\epsilon) = \epsilon$, where y denotes the true (unknown) label. By Theorem 2, then, it exists a copula C_Q s.t. $Q(\epsilon_1, ..., \epsilon_n) = C(Q_1(\epsilon_1), ..., Q_n(\epsilon)) = C_Q(\epsilon_1, ..., \epsilon_n)$, by definition of smoothed CP.

As a first remark, we note that for any of $\Gamma_{\min}, \Gamma_{\max}, \Gamma_w$ it is easy to observe that if $C_Q = M$ then the corresponding EP can be guaranteed to be conservatively valid. Indeed, in this case, the marginals Q_i of the combined CPs are *co-monotone*, hence, assuming y is the true label, it holds that $\exists i \in \{1, \ldots, n\}$ s.t. $p_i^y \geq \epsilon$ implies $\forall i \in \{1, \ldots, n\} p_i^y \geq \epsilon$, and hence $Pr(y \notin \Gamma_k) \leq \epsilon$, for $k \in \{\min, \max, w\}$. We note, in particular, that this property holds irrespective of whether the combined p-value functions corresponding to the three different combination rules coincide or not, and only derives from the comonotonicity of the marginals.

Based on the previous observations, we then study the properties of the above mentioned EPs, starting from $\Gamma_{\min}$.

Theorem 3. *$\Gamma_{\min}$ is strongly valid iff $C_Q = M$ (i.e., the p-value functions p_i are co-monotone). Furthermore, if C_Q is order-independent, then $Pr(y \notin \Gamma^\epsilon_{\min}) = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} C(\underbrace{\epsilon, ..., \epsilon}_{k}, \underbrace{1, ..., 1}_{n-k})$. In particular, if $C_Q(\epsilon_1, ..., \epsilon_n) \leq I$ (i.e., the Q_i are independent or have negative dependence), then $Pr(y \notin \Gamma^\epsilon_{\min})$ increases exponentially with n, for all $\epsilon \in (0, 1)$.*

Proof. See Appendix. □

The previous result provides a formula for computing the error rate of the min-based EP for a large collection of copulas: in particular, all associative copulas (including Archimedean copulas (Ling, Fang, and Kolter 2020)) are order-independent. Furthermore, the result also shows that the min-based EP is strongly valid iff the marginal error rates are co-monotone, while the error rate generally increases with the number of CPs to be combined (i.e., n). By contrast, the following result shows that the min-based EP improves the efficiency w.r.t. the combined CPs.

Proposition 2. *$\Gamma_{\min}$ is more efficient than Γ_i, for all i. Furthermore, $\mathbb{E}[|\Gamma^\epsilon_{\min}|]$ is non-increasing with n, for all ϵ.*

Proof. The result follows by noting that $\forall i, \forall y, \min_j\{p_j^y\} \leq p_i^y$ implies $\Gamma^\epsilon_{\min} \subseteq \bigcap_i \Gamma_i^\epsilon$. □

Corollary 1. *$\forall i, \epsilon, Pr(\Gamma^\epsilon_{\min} = \emptyset) \geq Pr(\Gamma_i^\epsilon = \emptyset)$ and $Pr(\Gamma^\epsilon_{\min} = \emptyset))$ is non-decreasing with n.*

As a consequence of the two previous results, while min-based EP's efficiency improves with n, this comes at the price of a corresponding reduction in validity. Nonetheless, if we interpret the event $\Gamma_{\min} = \emptyset$ as a rejection event, rather than an error, then Corollary 1 ensures that the probability of genuine errors (i.e., predicting a set larger than $\emptyset$ that does not contain the true label y) is decreasing with n. This interpretation is coherent with the common understanding of the p-value function (Campagner et al. 2021; Hüllermeier and Waegeman 2021): if $\Gamma^\epsilon_{\min}(x) = \emptyset$, then x was an outlier for at least one of the combined CPs, thus, since all combined CPs are assumed to be equally reliable, then x could be conservatively deemed to be an outlier also for the EP $\Gamma_{\min}$.

Next, we study the properties of the max-based EP.

Theorem 4. *$\Gamma_{\max}$ is conservatively valid. It is strongly valid iff $C_Q = M$. Furthermore, $Pr(y \notin \Gamma^\epsilon_{\max}) = C_Q(\epsilon, ..., \epsilon)$. Thus, as n grows, $Pr(y \notin \Gamma^\epsilon_{\max})$ is non-increasing.*

Proof. See Appendix. □

Proposition 3. *$\Gamma_{\max}$ is less efficient than Γ_i, for all i. Furthermore, $\mathbb{E}[|\Gamma^\epsilon_{\max}|]$ is non-decreasing with n, for all ϵ.*

Proof. The result follows by noting that $\forall i, \forall y, \max_j\{p_j^y\} \geq p_i^y$ implies $\bigcup_i \Gamma_i^\epsilon \subseteq \Gamma^\epsilon_{\max}$. □

[6] In the following, we will omit reference to the non-conformity measure f, as the results we prove do not directly depend on f.

[7] As we mentioned in Section , our results hold equivalently in the online CP and in the inductive split-CP settings. For this reason, we will adopt an abstract formulation of CP that encompasses both the above mentioned settings.

[8] We note that p_i is a $|Y|$-dimensional random vector, whose realization depends on both the bag $B_i \in Z^*$ and the newly observed example $(x, y) \in Z$. For ease of notation, in the rest of the paper, we will hold constant, and thus omit, B_i and x.

The previous result shows that the min-based and max-based EPs are opposite in regard to their validity/efficiency behavior: while $\Gamma_{\max}$ is (conservatively) valid (irrespective of the joint CDF Q), its efficiency decreases with the the number n of combined CPs. Remarkably, $\Gamma_{\max}$ satisfies also a stronger definition of validity, what in (Cella and Martin 2021) is called *type-2 validity*. Furthermore, in general, it is the only combination rule satisfying this notion of validity. We state this result in the Appendix, for brevity's sake.

Next, we study the weighted mean EP[9]. First, we quantify the error rate of $\Gamma_{\mathbf{w}}$:

Theorem 5. *Let* $\mathbf{w} \in [0,1]^n$*, s.t.* $\sum_i w_i = 1$*. Further, assume that* C_Q *admits a density.Then* $Pr(y \notin \Gamma^{\epsilon}_{\mathbf{w}}) = \int_{[0,1]^{n-1}} \frac{\partial C}{\partial u_1 ... \partial u_{n-1}}(u_1, ..., u_{n-1}, Q_n(\frac{\epsilon - \sum_{i=1}^{n} w_i u_i}{w_n}))\ d\lambda$.

Proof. See Appendix. □

Similarly to Theorems 3 and 4, also Theorem 5 provides a formula to compute the error rate of the weighted mean EP as a function of ϵ (and, implicitly, of n). However, while the formula in Theorem 5 can be evaluated numerically (e.g. by Monte Carlo methods (Scherer and Mai 2017)), it does not provide a closed-form characterization of the copulas for which the weighted mean EP is valid. In the Appendix we provide such a characterization, showing that, similarly to the $\min$ and $\max$ EPs, comonotonocity of the marginals is a necessary condition for Γ_w to be strongly valid.

Finally, we consider the EP based on Dempster's combination. We note that the analysis of this latter EP is more complex, due to the presence of $1-k$ as a normalization factor, which implicitly influences both its validity and efficiency. To simplify the treatment, we note that if no pair of p-value functions p_i^y, p_j^y is totally in conflict (i.e., $k \neq 1$), the following relation holds (Denœux 2019) $p^y_{m_1 \oplus ... \oplus m_n} = \frac{\Pi_i p l_i^y}{K}$, where $K = \Pi_{i=2}^{n}(1 - k_{\{1,...,i-1\},i})$ is the total degree of agreement[10]. Thus, we prove the following result:

Theorem 6. *Let* $K \neq 0$ *be the total degree of agreement between the* p_i^y*. Further, assume that* C_Q *admits a density. Then* $Pr(y \notin \Gamma^{\epsilon}_{D}) = \int_{[0,1]^{n-1}} \frac{\partial C}{\partial u_1 ... \partial u_{n-1}}(u_1, ..., u_{n-1}, Q_n(\frac{K\epsilon}{\Pi_{i=1}^{n-1} u_i}))\ d\lambda$.

Proof. See Appendix. □

Remarkably, when $n = 2$, for no copula C_Q the EP Γ_D is unconditionally valid. We prove this result in the Appendix.

Finally, we conclude with a remark on the significance of our results in MTSC. As we previously noted, MTSC methods based on the combination of multiple univariate models can be interpreted as IF methods, as they combine multi-source, potentially conflicting, information hints (i.e. the confidence scores predicted by the univariate models) about the assignment of instances to classes: thus, our results apply to MTSC as a special case. The connection, however, goes deeper. In MTSC, and in time series classification more in general, feature engineering represents one of the most critical steps in model development (Baydogan and Runger 2015). Methods based on the combination of univariate models can then be understood also as feature extraction methods: indeed, the confidence scores to combined are high-level features, and the aim is then to define a simple classification rule based on the above mentioned features. However, in general, such features do not provide any type of guarantee w.r.t. their representativeness: for example, the underlying classifiers could be poorly accurate, or mis-calibrated. EPs, then, can be seen as a way to avoid this limitation: instead of using generic classifiers' confidence scores as predictive features, we employ the p-value functions obtained by the CPs to be aggregated. In this sense, the results we proved in this section characterize the relations which should hold among these high-level predictive features (the output of the CPs) to guarantee that their combination preserves validity.

Experiments

In this section, we describe the results of experiments we performed to assess the empirical performance of the studied EPs against both standard state-of-the-art ensemble algorithms for the MTSC task (Ruiz et al. 2021), as well as other combination techniques for CP proposed in the literature. The aim of these experiments is threefold: first, to compare the effectiveness of EPs against alternative CP aggregation methods; second, to evaluate whether univariate-based combination of CPs can be as effective as bespoke and state-of-the-art MTSC methods; third, to analyze the translation of our theoretical results concerning the validity of EPs in practical and empirical settings. For this purpose, we considered the collection of 24 MTSC benchmark datasets from the public UEA/UCR archive (Dau et al. 2018; Bagnall et al. 2018), the largest and most popular benchmark repository for MTSC tasks.

In terms of algorithms, we considered as baseline comparison the following MTSC classifiers: ROCKET (Dempster, Petitjean, and Webb 2020), HIVECOTE (Bagnall et al. 2020), Canonical Interval Forest (CIF) (Middlehurst, Large, and Bagnall 2020). ROCKET and CIF are bespoke algorithms, as they are based on the extraction of features from the multivariate time signals. Specifically, ROCKET employs a large number of convolutional kernels as random features that are subsequently used to train a linear classifier ; by comparison, CIF employs the catch22 features (Lubba et al. 2019), as well as summary statistics, to build a forest of decision trees. By contrast, HIVECOTE is not a bespoke algorithm, but rather is obtained as the ensemble of several state-of-the-art univariate methods, each of which is trained on each dimension of the time series under analysis. In this sense, HIVECOTE is more similar to the approach also adopted in the proposed EP method. These algorithms were selected as they were shown to be the best 3 MTSC algorithms by (Ruiz et al. 2021), in what is the most extensive benchmark study

[9] We note that when $\forall i, w_i = \frac{1}{n}$, $\Gamma_{\mathbf{w}}$ generalizes the aggregated CP (Carlsson, Eklund, and Norinder 2014). In particular, the aggregated CP coincides with $\Gamma_{\mathbf{w}}$ when the bags B_i are re-sampled from a given bag B. In the Appendix, specifically in Theorems 11 and 4, we show that this connection and our theoretical analysis enables us to clarify and generalize the theoretical results in (Carlsson, Eklund, and Norinder 2014).

[10] In the definition of K, the order in which the p-value functions are combined is irrelevant, since Dempster's rule of combination is both associative and commutative under the stated assumptions.

of MTSC classification methods.

In terms of combination rules for CP, we considered the EPs studied in this work (that is: $\min$, $\max$, two different settings of $\mathbf{w}$, and Dempster's rule, denoted respectively as IEP(min), IEP(max), IEP(mean), IEP(weighted), IEP(dempster)) as well as two methods previously discussed in the CP literature, namely the majority vote rule (Cherubin 2019) (ICP(count)) and Fisher's Extended Chi-Square Function (ICP(fisher)) method (Balasubramanian, Chakraborty, and Panchanathan 2015). We selected these latter two approaches, in particular, as they were shown in previous studies to be state-of-the-art CP combination methods in the resampling-based and IF settings, respectively. Furthermore, we also considered a non-ensemble-based bespoke CP model as a baseline comparison (denoted as CP(base)), to evaluate the effectiveness of the above mentioned ensemble techniques as compared with a bespoke implementation of CP.

For each of the dataset, we set the number of CPs to be combined to $n = d$, where d is the number of dimensions for the dataset. In regard to the weighted mean EP, we considered two different settings: *uniformly weighted* (IEP(mean)), i.e. $\forall i, w_i = \frac{1}{n}$; and *confidence weighted* (IEP(weighted)), i.e. $w_i = \frac{p_i^{y_1} - p_i^{y_2}}{\sum_i w_i}$, where $p_i^{y_j}$ is the j-highest p-value for CP i. The rationale for this latter weighting scheme is to weigh more the CPs with higher *confidence* (Gammerman and Vovk 2002; Hüllermeier and Waegeman 2021).

All CP methods were implemented by means of inductive split-CP (Papadopoulos, Vovk, and Gammerman 2002), using ROCKET as the baseline classifier. In particular, the ensemble-based methods used a uni-variate version of ROCKET for each of the dimensions in the datasets at hand, while CP(base) used a bespoke, hence directly multi-variate, version of ROCKET. In regard to the standard MTSC classifiers, we consider as reference the results reported in (Ruiz et al. 2021), while for the CP combinations methods we used a custom implementation in Python. All code was implemented using Python v. 3.10.3, numpy v. 1.22.4, scikit-learn v. and sktime v. 0.13.1. For CP methods we used RockerClassifier as base classifier, as implemented in the sktime library (Löning et al. 2019), with default hyper-parameters. All code, including experiments and setup, is available at anonymized-github, and was executed on a PC equipped with a i7 11700k, RTX 3060 and 16 GB of RAM.

For each dataset, we considered the default train-test split defined in the UEA/UCR archive, to ensure reproducibility and comparability of results. In particular, for CP methods, since we adopted a split CP design, for each dataset 75% of the training set was used as training set proper, while the remaining 25% as calibration set. All random seeds were set to 0 to ensure reproducibility. Algorithms were evaluated in terms of Rejection-Discounted Accuracy (RA) $\frac{|\{(x,y) \in S : y \in h(x)\}|}{|S|}$ and Efficiency $1 - \frac{1}{|S|} \sum_{(x,y) \in S} \frac{|h(x)|-1}{|Y|-1}$. Intuitively, RA measures the ability of a model to predict the correct class and coincides with the standard notion of accuracy employed in CP literature[11]; Efficiency is instead defined as the complement of the average size of the predicted sets of labels. Results in the main paper are reported relative to the threshold value $\epsilon = .25$ (which corresponds to a confidence level of .75), to avoid overly large prediction sets. The overall results, as well as their statistical analysis (in terms of Critical Distance diagrams, based on Friedman omnibus test and Wilcoxon post-hoc procedure (Benavoli, Corani, and Mangili 2016)), are reported in Figures 1-4, while additional results are available in the Appendix. In particular, the validity and efficiency of the considered CP combination methods were considered, based on validity and efficiency curves and deviation from nominal strong validity: these are reported in the Appendices, in Figures A11-A22.

[11] We use the term *rejection-discounted* to make it explicit that we consider rejection events, i.e. cases in which $\Gamma^\epsilon = \emptyset$, as errors.

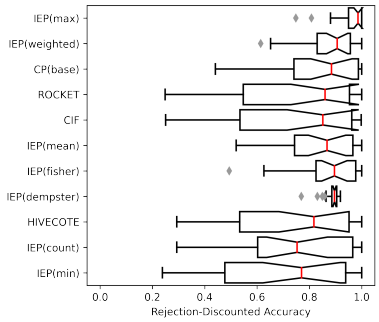


Figure 1: Notched boxplots for the performance of the considered classifiers, in terms of rejection-discounted accuracy.

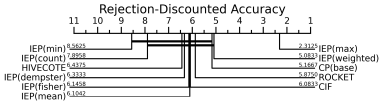


Figure 2: Critical difference plots of the average ranks of the considered classifiers, in terms of rejection-discounted accuracy.

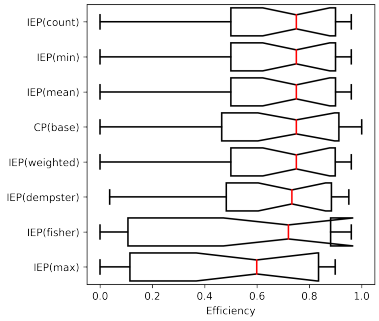


Figure 3: Notched boxplots for the performance of the considered classifiers, in terms of efficiency.

As shown in the Figures (see also the Appendix), in terms of RA, IEP(max) significantly out-performed all other algorithms: this results is not unexpected, since following our theoretical development it is known that IEP(max) is the only

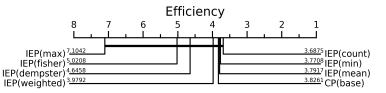


Figure 4: Critical difference plots of the average ranks of the considered classifiers, in terms of efficiency.

EP which is guaranteed to be conservatively valid. The high accuracy of IEP(max), however, comes at the price of its efficiency: indeed, IEP(max) was the worst method in terms of efficiency, even though its efficiency was not significantly worse than that of IEP(dempster) and ICP(fisher). Compared to IEP(max), also IEP(weighted) reported good results in terms of RA, being the second best algorithm w.r.t. to this metric. However, even if IEP(weighted) was better than the remaining algorithms, the difference was not statistically significant w.r.t. neither the baseline classifiers, CP(base) or the ICP(fisher) methods. By contrast, both IEP(min) and IEP(dempter) were the worst-performing methods in terms of RA: this is not surprising, as the high accuracy of the first EP largely stems from their high rejection rates (see Corollary 1), while in regard to IEP(dempster) the observed high error rate stems from the approximation approach we adopted to recover consonance (i.e. approximating a general mass function by the corresponding consonant projection).

In regard to efficiency, CP(base), IEP(min), IEP(mean), IEP(weighted) and ICP(count) reported the best efficiency. By contrast, IEP(max) and IEP(dempster) reported the worst performance in terms of efficiency. Thus, in particular, we claim that IEP(weighted) offers the best trade-off between accuracy and efficiency and could thus be considered as the best out-of-the-box combination method for CPs. Also in comparison with the standard classifiers and CP(base) (see also the Appendix, in Figures A1-A6), IEP(weighted) was, along with IEP(max), the best performing algorithm in terms of accuracy, while having much better efficiency than the latter. Even though IEP(weighted) was not significantly better than the best baseline classifier (i.e. ROCKET) or the corresponding CP-based correction (i.e. CP(base)), we remark that EPs only rely on univariate information: this is in contrast with CP(base), ROCKET and CIF (the third, fourth and fifth best performing algorithms), which are bespoke algorithms. Thus, this result highlights how EP methods in general, and IEP(weighted) in particular, can reach performances comparable with, or even better than, state-of-the-art multivariate models, while only relying on univariate information. Furthermore, more in general, it can be easily seen from Figure 1, that EPs exhibited a significantly lower variance in performance than all other methods, and in particular than CP(base): in particular, IEP(dempster), IEP(max) and IEP(weighted) had the smallest variance among the considered methods. This result highlights how the use of EPs could lead to an improvement in stability as compared with the application of CP methods based on bespoke models.

Focusing on the comparison between CP combination methods (see Figures A11-A15 and Table A2), it can be easily seen that none of the considered methods were strongly valid. This result is not surprising, since in our theoretical analysis we showed that co-monotonicity is a necessary condition for achieving strong validity. This property, however, is a rather strong constraint which could not be expected to hold in general settings. Nonetheless, we note that all the considered CP combination methods, with the exception of IEP(min), were empirically conservatively valid on most datasets at the selected threshold $\epsilon = .25$. In particular, IEP(max) and IEP(weighted) (as well as CP(base)) were conservatively valid at all confidence levels in Figures A11-A15. While the other combination methods were not in general conservatively valid at all thresholds levels, IEP(mean) and ICP(fisher) were nonetheless approximately conservatively valid on most of the considered datasets. At the same time, IEP(weighted) was the method, after CP(base), which had the smallest deviation from strong validity, as shown in Figures A21 and A22. These results, also in combination with the efficiency analysis reported in Figures A16-A20, confirms the practical efficacy of IEP(weighted) and, although to a lower degree due to its reduced efficiency and larger deviation from validity, IEP(max). As we stated previously, we believe that due to their superior performance, which we remark grounds only on univariate feature information in contrast with state-of-the-art MTSC methods, these two EPs can be effectively applied in general MTSC tasks. Then, selection among these two methods depends on the constraints of the specific application considered. In general settings, where high accuracy is desirable but nominal validity and efficiency are of greatest importance, IEP(weighted) should be preferred. However, if the considered application demands accuracy at all costs (including sacrificing efficiency and nominal coverage), as it would happen in any critical domain such as medicine, then IEP(max) should be preferred.

Conclusion

In this article, we introduced a class of combination methods in the Conformal Prediction framework, that we called Evidential Predictors. In particular, we studied the theoretical properties of several EPs providing formulas for their theoretical error rate as well as conditions for their validity. To our knowledge, this marks a first in the specialized literature. We also assessed the effectiveness of EPs in MTSC, by means of an extensive set of experiments through which we compared the proposed EPs against both state-of-the-art classifiers and other CP combination methods, reporting promising results: in particular, we showed that the weighted-average and maximum EPs can obtain performance comparable and on average better than other state-of-the-art algorithms, while relying only on univariate information.

In light of these results, we believe that the following open problems could be of interest: 1) Many other IF methods have been proposed in the specialized literature (e.g., generalized integrals), whose properties could thus be studied in the EP setting; 2) In this article we have shown that co-monotonicity is a sufficient and necessary condition for validity. Therefore, further research should study non-conformity measures or design mechanisms that can ensure that this property holds; 3) Finally, we believe that the application of EPs in more general, multi-modal tasks should be pursued and evaluated.

References

Bagnall, A.; Dau, H. A.; Lines, J.; Flynn, M.; Large, J.; Bostrom, A.; Southam, P.; and Keogh, E. 2018. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*.

Bagnall, A.; Flynn, M.; Large, J.; Lines, J.; and Middlehurst, M. 2020. On the usage and performance of HIVE-COTE v1. 0. In *Proceedings of the 5th workshop on advances analytics and learning on temporal data, lecture notes in artificial intelligence*, volume 12588.

Balasubramanian, V. N.; Chakraborty, S.; and Panchanathan, S. 2015. Conformal predictions for information fusion. *Annals of Mathematics and Artificial Intelligence* 74(1):45–65.

Barandas, M.; Folgado, D.; Santos, R.; Simão, R.; and Gamboa, H. 2022. Uncertainty-based rejection in machine learning: Implications for model development and interpretability. *Electronics* 11(3):396.

Baydogan, M. G., and Runger, G. 2015. Learning a symbolic representation for multivariate time series classification. *Data Mining and Knowledge Discovery* 29(2):400–422.

Benavoli, A.; Corani, G.; and Mangili, F. 2016. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research* 17(1):152–161.

Campagner, A.; Cabitza, F.; Berjano, P.; and Ciucci, D. 2021. Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches. *Information Sciences* 579:347–367.

Carlsson, L.; Eklund, M.; and Norinder, U. 2014. Aggregated conformal prediction. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 231–240. Springer.

Cauchois, M.; Gupta, S.; and Duchi, J. C. 2021. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.* 22:81–1.

Cella, L., and Martin, R. 2021. Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*.

Cherubin, G. 2019. Majority vote ensembles of conformal predictors. *Machine Learning* 108(3):475–488.

Dau, H. A.; Keogh, E.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; Yanping, C.; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; Batista, G.; and Hexagon-ML. 2018. The ucr time series classification archive. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Davidov, O. 2011. Combining p-values using order-based methods. *Computational Statistics & Data Analysis* 55(7):2433–2444.

Dempster, A.; Petitjean, F.; and Webb, G. I. 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery* 34(5):1454–1495.

Dempster, A. 1967. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* 38(2):325–339.

Denœux, T. 2019. Logistic regression, neural networks and dempster–shafer theory: A new perspective. *Knowledge-Based Systems* 176:54–67.

Dhariyal, B.; Le Nguyen, T.; Gsponer, S.; and Ifrim, G. 2020. An examination of the state-of-the-art for multivariate time series classification. In *2020 International Conference on Data Mining Workshops*, 243–250. IEEE.

Dubois, D., and Prade, H. 1988. Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence* 4(3):244–264.

Dubois, D., and Prade, H. 1990. Consonant approximations of belief functions. *International Journal of Approximate Reasoning* 4(5-6):419–449.

Dubois, D.; Liu, W.; Ma, J.; and Prade, H. 2016. The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks. *Information Fusion* 32:12–39.

Dubois, D.; Everaere, P.; Konieczny, S.; and Papini, O. 2020. Main issues in belief revision, belief merging and information fusion. In *A Guided Tour of Artificial Intelligence Research*. Springer. 441–485.

Gammerman, A., and Vovk, V. 2002. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science* 287(1):209–217.

Genest, C.; Molina, J. Q.; Lallena, J. R.; and Sempi, C. 1999. A characterization of quasi-copulas. *Journal of Multivariate Analysis* 69(2):193–205.

Gijbels, I., and Herrmann, K. 2014. On the distribution of sums of random variables with copula-induced dependence. *Insurance: Mathematics and Economics* 59:27–44.

Guerra, R.; Etzel, C. J.; Goldstein, D. R.; and Sain, S. R. 1999. Meta-analysis by combining p-values: simulated linkage studies. *Genetic epidemiology* 17(S1):S605–S609.

Hüllermeier, E., and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* 110(3):457–506.

Ismail, A. A.; Gunady, M.; Corrada Bravo, H.; and Feizi, S. 2020. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems* 33:6441–6452.

Jin, X.-B.; Yang, A.; Su, T.; Kong, J.-L.; and Bai, Y. 2021. Multi-channel fusion classification method based on time-series data. *Sensors* 21(13):4391.

Johnstone, C., and Cox, B. 2021. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, 72–90. PMLR.

Kaas, R.; Dhaene, J.; Vyncke, D.; Goovaerts, M. J.; and Denuit, M. 2002. A simple geometric proof that comonotonic risks have the convex-largest sum. *ASTIN Bulletin: The Journal of the IAA* 32(1):71–80.

Kompa, B.; Snoek, J.; and Beam, A. L. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4(1):1–6.

Le Carrer, N., and Ferson, S. 2021. Beyond probabilities: A possibilistic framework to interpret ensemble predictions and fuse imperfect sources of information. *Quarterly Journal of the Royal Meteorological Society* 147(739):3410–3433.

Ling, C. K.; Fang, F.; and Kolter, J. Z. 2020. Deep archimedean copulas. *Advances in Neural Information Processing Systems* 33:1535–1545.

Linusson, H.; Norinder, U.; Boström, H.; Johansson, U.; and Löfström, T. 2017. On the calibration of aggregated conformal predictors. In *Conformal and probabilistic prediction and applications*, 154–173. PMLR.

Linusson, H.; Johansson, U.; and Boström, H. 2020. Efficient conformal predictor ensembles. *Neurocomputing* 397:266–278.

Linusson, H. 2021. *Nonconformity Measures and Ensemble Strategies: An Analysis of Conformal Predictor Efficiency and Validity*. Ph.D. Dissertation, Department of Computer and Systems Sciences, Stockholm University.

Löning, M.; Bagnall, A.; Ganesh, S.; Kazakov, V.; Lines, J.; and Király, F. J. 2019. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872*.

Loughin, T. M. 2004. A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis* 47(3):467–485.

Lubba, C. H.; Sethi, S. S.; Knaute, P.; Schultz, S. R.; Fulcher, B. D.; and Jones, N. S. 2019. catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery* 33(6):1821–1852.

Ly, S.; Pho, K.-H.; Ly, S.; and Wong, W.-K. 2019. Determining distribution for the product of random variables by using copulas. *Risks* 7(1):23.

Messoudi, S.; Destercke, S.; and Rousseau, S. 2021. Copula-based conformal prediction for multi-target regression. *Pattern Recognition* 120:108101.

Messoudi, S.; Destercke, S.; and Rousseau, S. 2022. Ellipsoidal conformal inference for multi-target regression. *Proceedings of Machine Learning Research* 179:1–13.

Middlehurst, M.; Large, J.; and Bagnall, A. 2020. The canonical interval forest (CIF) classifier for time series classification. In *2020 IEEE international conference on big data (big data)*, 188–195. IEEE.

Nelsen, R. B. 2007. *An introduction to copulas*. Springer Science & Business Media.

Ng, Y.; Hasan, A.; Elkhalil, K.; and Tarokh, V. 2021. Generative archimedean copulas. In *Uncertainty in Artificial Intelligence*, 643–653. PMLR.

Orponen, P. 1990. Dempster's rule of combination is# p-complete. *Artificial Intelligence* 44(1-2):245–253.

Pantiskas, L.; Verstoep, K.; Hoogendoorn, M.; and Bal, H. 2022. Taking rocket on an efficiency mission: Multivariate time series classification with lightwaves. *arXiv preprint arXiv:2204.01379*.

Papadopoulos, H.; Vovk, V.; and Gammerman, A. 2002. Qualified prediction for large data sets in the case of pattern recognition. In *Proceedings of the 2002 Internation Conference on Machine Learning and Applications*, 159–163.

Pereira, A.; Folgado, D.; Nunes, F.; Almeida, J.; and Sousa, I. 2019. Using inertial sensors to evaluate exercise correctness in electromyography-based home rehabilitation systems. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 1–6. IEEE.

Ruiz, A. P.; Flynn, M.; Large, J.; Middlehurst, M.; and Bagnall, A. 2021. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 35(2):401–449.

Scherer, M., and Mai, J.-f. 2017. *Simulating Copulas: Stochastic Models, Sampling Algorithms, And Applications*, volume 6. World Scientific.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Sklar, M. 1959. Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8:229–231.

Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*.

Spjuth, O.; Brännström, R. C.; Carlsson, L.; and Gauraha, N. 2019. Combining prediction intervals on multi-source non-disclosed regression datasets. In *Conformal and Probabilistic Prediction and Applications*, 53–65. PMLR.

Susto, G. A.; Cenedese, A.; and Terzi, M. 2018. Time-series classification methods: Review and applications to power systems data. *Big data application in power systems* 179–220.

Toccaceli, P., and Gammerman, A. 2017. Combination of conformal predictors for classification. In *Conformal and Probabilistic Prediction and Applications*, 39–61. PMLR.

Toccaceli, P., and Gammerman, A. 2019. Combination of inductive mondrian conformal predictors. *Machine Learning* 108(3):489–510.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V. 2015. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* 74(1):9–28.

Xu, C., and Xie, Y. 2021. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, 11559–11569. PMLR.

Zaffran, M.; Féron, O.; Goude, Y.; Josse, J.; and Dieuleveut, A. 2022. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, 25834–25866. PMLR.

Zaykin, D. V.; Zhivotovsky, L. A.; Czika, W.; Shao, S.; and Wolfinger, R. D. 2007. Combining p-values in large-scale genomics experiments. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 6(3):217–226.

Supplementary Materials

Proofs and Additional Theorems

Proof of Theorem 3. We consider the case $n = 2$, the general result follows by induction. First, we show that, when $C_Q(x, y) = M(x, y)$, $\Gamma_{\min}$ is strongly valid. By definition of $\Gamma_{\min}$ and standard probability calculus, $Pr(y \notin \Gamma^{\epsilon}_{\min}(x)) = Pr(\min\{p^x_1(y), p^x_2(y)\} < \epsilon) = C_Q(\epsilon, 1) + C_Q(1, \epsilon) - C_Q(\epsilon, \epsilon) = 2\epsilon - M(\epsilon, \epsilon) = \epsilon$.

For the other part, note that in the previous proof we used $C_Q = M$ only in the last equivalence. Therefore, the EP $\Gamma_{\min}$ is conservatively valid iff $\forall \epsilon, C_Q(\epsilon, \epsilon) \geq \epsilon$. By the Frechet-Hoeffding bounds, this happens iff $C_Q = M$.

Finally, note that if $C_Q(\epsilon_1, ..., \epsilon_n) \leq I$ then $\forall y$, $Pr(y \notin \Gamma^{\epsilon}_{\min}(x)) \geq n\epsilon - \sum_{k=2}^{n}(-1)^k \binom{n}{k}\epsilon^k$. □

Proof of Theorem 4. $\Gamma_{\max}$ is conservatively valid iff $\forall \epsilon$ it holds that $Pr(y \notin \Gamma^{\epsilon}_{\max}(x)) = Pr(\max_i\{p^y_i\} < \epsilon) = Pr(\forall i, p^y_i < \epsilon) = C_Q(\epsilon, ..., \epsilon)$. For $C_Q = M$ it holds that the latter quantity is equal to ϵ. Furthermore, by Frechet-Hoeffding bounds, for every other copula C_Q it holds that $C_Q < M$. Finally, for $C_Q < M$, $C_Q(\underbrace{\epsilon, ..., \epsilon}_{n})$ is a decreasing function of n (since $\epsilon \in [0, 1]$, and $C_Q(\epsilon, ..., \epsilon) < \epsilon$). □

Theorem 7. *The EP $\Gamma_{\max}$ is type-2 valid, that is $Pr(\max_{y' \in A} \max_i\{p^{y'}_i\} \leq \epsilon, y \in A) \leq \epsilon$. Furthermore, without any assumption on the joint CDF Q, it is the only EP satisfying the previous property.*

Proof. By (Cella and Martin 2021), Theorem 1, a CP is type-2 valid iff the corresponding p-value function is normalized. Since, by assumption, this holds for every Γ_i, it follows that the same holds also for $\Gamma_{\max}$. For the second statement, note that, for every other combination rule r, $\max_y r(p^y_1, ..., p^y_n) \leq \max_y \max_i\{p^y_i\}$. □

Proof of Theorem 5. The result follows (Gijbels and Herrmann 2014), Proposition 2.1 and Theorem 2.2. Indeed, let $Z = \sum_i w_i p^y_i$ be the weighted mean random variable, with CDF F_Z. Let $A_\epsilon = \{x \in [0, 1]^n : \sum_i w_i x_i \leq \epsilon\}$. Then, $Pr(y \notin \Gamma^{\epsilon}_{\mathbf{w}}) = F_Z(\epsilon) = \int_{\phi(A_\epsilon)} c \, \mathrm{d}\lambda$, where c is the density of C_Q (which exists by assumption), $\phi(x) = (Q_1(x_1), ..., Q_n(x_n))$ and λ is Lebesgue measure. By the definition of Q_i, it holds that

$$\phi(x)_i = \begin{cases} 1 & x_i \geq 1 \\ 0 & x_i \leq 0 \\ x_i & otherwise \end{cases}.$$

Then, $Pr(y \notin \Gamma^{\epsilon}_{\mathbf{w}}) = \int_{A_\epsilon} c d\lambda$ and the result follows. □

Theorem 8. *Let $\mathbf{w} \in [0, 1]^n$, s.t. $\sum_i w_i = 1$. Then the EP $\Gamma_{\mathbf{w}}$ is strongly valid iff $C_Q = M$, it is conservatively valid iff $\mathbb{1}_{Z \leq \epsilon}$ is convex for every $\epsilon \in [0, 1]$, where $Z = w_1 Q_1 + ... + w_n Q_n$.*

Proof. First, we show that $C_Q = M$ implies $\Gamma_{\mathbf{w}}$ is strongly valid. By (Kaas et al. 2002), Theorem 7, $F^{-1}_Z(\epsilon) = w_1 Q^{-1}_1(\epsilon) + ... + w_n Q^{-1}_n(\epsilon) = \epsilon$. For the converse, by (Kaas et al. 2002), Lemma 1, for every C_Q it holds that $P(y \notin \Gamma^{\epsilon}_{\mathbf{w}}) = \epsilon$ iff $\mathbb{1}_{Z \leq \epsilon}$ is both convex and concave, which, by (Kaas et al. 2002) Theorem 6, holds only if $C_Q = M$. In regard to conservative validity, note that, if for a certain copula C_Q $\mathbb{1}_{Z \leq \epsilon}$ is convex for every ϵ, it holds that $P(y \notin \Gamma^{\epsilon}_{\mathbf{w}}) \leq \epsilon$ by (Kaas et al. 2002), Theorem 7 and Lemma 1. □

Proof of Theorem 6. The result follows from (Ly et al. 2019), Theorem 2. Indeed, let $Z = \Pi_i p^y_i$ be the product random variabile, with CDF F_Z. Let $A_\epsilon = \{x \in [0, 1]^n : \Pi_i \leq K\epsilon\}$. Then, $Pr(y \notin \Gamma^{\epsilon}_D) = F_Z(\epsilon) = \int_{\phi(A_\epsilon)} c \, \mathrm{d}\lambda$, where c and $\phi(x)$ are as in Theorem 5. By (Ly et al. 2019) Theorem 2, $\int_{\phi(A_\epsilon)} c \, \mathrm{d}\lambda = V^- + \int_{[0,1]^{n-1}} sign\left(\Pi^{n-1}_{i=1} Q^{-1}_i(u_i)\right) \tau(Q_n(\frac{K\epsilon}{\Pi^{n-1}_{i=1} u_i})) \, \mathrm{d}\lambda$, where $V = \int_{\{u_1,...,u_n \in [0,1]^{n-1}; \Pi^{n-1}_{i=1} Q^{-1}_i(u_i) < 0\}} \tau(1) d\lambda$ and $\tau(z) = \frac{\partial C}{\partial u_1 ... \partial u_{n-1}}(u_1, ..., u_{n-1}, z)$. Since $V = 0$ and $sign\left(\Pi^{n-1}_{i=1} Q^{-1}_i(u_i)\right) = 1$, the result follows. □

Theorem 9. *For every copula C_Q, n and ϵ, Γ_D is not (strongly, conservatively) valid. Furthermore, if $n = 2$, there is a value $K_Q(\epsilon, n)$ s.t. if $K \leq K_Q(\epsilon,)$ then $Pr(y \notin \Gamma^{\epsilon}_D) \leq \epsilon$, and if $K > K_Q(\epsilon, n)$ then $Pr(y \notin \Gamma^{\epsilon}_D) > \epsilon$.*

Proof. The first statement follows directly by noting that, for every copula C_Q, $Pr(y \notin \Gamma^{\epsilon}_D) \geq K\epsilon$: thus, in particular, if $K = 1$ (i.e. there is no conflict among the CPs), Γ_D cannot be unconditionally valid. For the second statement, by Theorem 6, for the copula W it holds that $Pr(y \notin \Gamma^{\epsilon}_D) \leq \epsilon$ iff $K \leq \frac{2-\epsilon}{4}$; furthermore, if $K\epsilon > \frac{1}{4}$, then $Pr(y \notin \Gamma^{\epsilon}_D) = 1$. Similarly, for copula M, it holds that $Pr(y \notin \Gamma^{\epsilon}_D) = \sqrt{K\epsilon}$, which is smaller than ϵ iff $K \leq \sqrt{\epsilon}$. Finally, for copula I, it holds that $Pr(y \notin \Gamma^{\epsilon}_D) = K\epsilon - K\epsilon \ln(K\epsilon)$. Since, depending on the value of K and ϵ, one of the three above mentioned copulas upper bounds all the other copulas (Ly et al. 2019), the result follows. □

Theorem 10. *Assume for simplicity that $|B_1| = ... = |B_n| = r$. Further, let t_f the cost of evaluating the nonconformity measure f. Then, in the inductive split-conformal setting, the complexity of the $\Gamma_{\min}$, $\Gamma_{\max}$ and $\Gamma_{\mathbf{w}}$ is $O(n \cdot |Y| \cdot t_f \cdot log(t_c))$. In the same setting, the complexity of Γ_D is in $o(n \cdot 2^{|Y|} \cdot t_f \cdot log(t_c))$.*

Proof. For the EPs $\Gamma_{\min}$, $\Gamma_{\max}$ and $\Gamma_{\mathbf{w}}$, the results simply follows by noting that the corresponding combination rules have complexity linear in the number of CPs to be aggregated. For the EP Γ_D, on the other hand, the result follows from the characterization of the complexity of Dempster's rule of combination due to (Orponen 1990). □

While the previous results focused on the study of the properties of EPs in the general, IF setting, the following result, instead, will connect the properties of EPs with EL setting. To this aim, we recall the notion of a *consistent sampling process* (CSP) (Carlsson, Eklund, and Norinder 2014):

Definition 1. *Let $M : Z^* \to \mathbb{R}$ be a function, S be a set and $R : S \to [0, 1]$ a probability measure on S, called* resampling procedure*, s.t. given $z_1, \ldots, z_n \in S$ and any permutation $\pi : [n] \to [n]$, $R(\{z_1, \ldots, z_n\}) = R(\{z_{\pi(1)}, \ldots, z_{\pi(n)}\})$. Then, R is a CSP w.r.t. M if:*

$$\lim_{n\to\infty, |S|\to\infty} \sup_m |G(m) - G^*(m)| = 0, \qquad (11)$$

where G is the distribution of M under the empirical distribution over S, and G^* is the distribution of M under R.

Theorem 11. *Let R be a CSP, B be a bag, $B_1, \ldots, B_n$ be sampled from S according to R, with $\forall i, |B_i| = |S|$. Let m be a non-conformity measure, and $\Gamma_1, \ldots, \Gamma_n$ be the CPs obtained by the non-conformity measure m on, respectively $B_1, \ldots, B_n$. Then, asymptotically, $C_Q = M$.*

Proof. Following (Carlsson, Eklund, and Norinder 2014), when R is a CSP, each of the marginals Q_i is valid iff all the others are. From this follows their co-monotonicity and hence $C_Q = M$. □

Thus, the previous result, together with the results in Section Methods, allow to generalize the result from (Carlsson, Eklund, and Norinder 2014), in which it is shown that, under the conditions of Theorem 11, the EP Γ_w with $w = \langle \frac{1}{n}, \ldots, \frac{1}{n} \rangle$. Indeed, the following corollary directly follows from Theorem 11:

Corollary 2. *Under the conditions of Theorem 11, $\Gamma_{min}, \Gamma_{max}, \Gamma_w$, for any w, are strongly valid.*

One of the main limitations of the above results concerns the observation, originally made in (Linusson et al. 2017), that defining CSPs in practical settings may be difficult, as the properties of such a resampling procedure depend both on the resampling procedure itself as well as on the selected non-conformity measure. To conclude this section, we provide a more practical instantiation of the above results, focusing on the case where R is the empirical distribution on the given bag B. To start, we define a notion of stability (*asymptotic bootstrap scoring stability* (ABSS)) for a given class of scoring classifiers:

Definition 2. *Let $\mathcal{H}$ be a class of scoring classifiers, and $A : Z^* \mapsto \mathcal{H}$ be a learning algorithm. Then, A is ABSS iff for each $y \in Y$:*

$$\lim_{n\to\infty} \sup_{S\subset Z:|S|=n} Pr\left(|A(S)(y) - A(S^B)(y)| > \epsilon\right) = 0,$$

where the probability is w.r.t. to the uniform resampling of a set S^B from S.

Intuitively, the output of an ABSS algorithm does not change much when its training set is sufficiently large. We remark that, even though the notion of an ABSS seems relatively strong, any asymptotically consistent learning algorithm (and, in particular, any asymptotically Bayes learning algorithm) can be easily shown to be ABSS. The following result, combined with Theorem 11, then shows that if one uses a non-conformity measure derived from an ABSS, the empirical distribution over the given bag B is a CSP, providing a simpler setting in which Theorem 11 holds (asymptotically):

Proposition 4. *Let A be ABSS, S be a training set, R the empirical distribution on S. Let $m_{S'}(x, y) = \max_{y' \in Y} A(S')(x, y') - A(S')(x, y)$. Then, asymptotically, R is CSP w.r.t. m_S.*

Proof. The result directly follows from the definitions of an ABSS and a CSP. □

List of Benchmark Datasets

The complete list of considered datasets, described in terms of train and test size, number of dimensions, time-series length and number of classes, is reported in Table A1. 5 different dataset categories can be distinguished (Ruiz et al. 2021), based on the source of the data. The considered datasets can be considered sufficiently representative of MTSC tasks, as they encompass different application domains, different number of instances $(15 - 7494)$, different number of classes $(2 - 26)$, different series lengths $(8 - 3000)$ and different numbers of dimensions $(2 - 963)$.

Additional Results

Complete Tables of Results The numerical results, in terms of the different considered metrics, are reported in Tables A2, A3, A4. Interestingly, we note that on 4 datasets (AtrialFibrillation, Handwriting, MotorImagery and StandWalkJump), all the evaluated EPs and CPs reported nil efficiency (that is, they always predicted the full set of class labels), see Table A3. This result, however, is not surprising, as the baseline classifiers reported low accuracy on these datasets (e.g., the accuracies reported by ROCKET were, respectively, 25%, 57%, 53% and 46%): indeed, since the EP and CP algorithms grounded on a univariate version of ROCKET as a baseline classifier, poor accuracy of the latter corresponds to poor representativeness of the corresponding non-conformity measure.

Pairwise Diagrams and Comparison with Standard Classifiers In order to compare the considered CP combination methods and EPs with standard MTSC classifiers, we considered an evaluation based on a novel metric, that we call Discounted Accuracy (DA), defined as:

$$F_2(RA, E) = 5\frac{RA * E}{4E + RA} \qquad (12)$$

DA measures the trade-off between accuracy and efficiency, by using the F_2 score (i.e., RA is weighed twice as much as efficiency). This latter metrics is used solely to compare the results of CP methods with those of traditional baseline classifiers: indeed, a comparison based solely on accuracy could be misleading. Thus, in the definition of DA we adopted the F_2 as a trade-off criterion between accuracy and efficiency since in the CP literature validity (hence, accuracy) is considered to be more important than efficiency.

Pairwise comparisons of IEP(max) and IEP(weighted) against the best standard classifier (i.e., ROCKET), in terms of RA and DA, are reported in Figures A3, A4, A5 and A6. It can be easily observed that both IEP(max) and IEP(weighted) out-performed ROCKET on all datasets from the ElectricBiosignals category (the most represented one), the Audio

Table A1: List of Benchmark Datasets from the UEA/UCR MTSC archive.

Dataset	TrainSize	TestSize	NumDimensions	SeriesLength	NumClasses	Type
ArticularyWordRecognition	275	300	9	144	25	Coordinates
AtrialFibrillation	15	15	2	640	3	ElectricBiosignals
BasicMotions	40	40	6	100	4	AccelerometerGyroscope
Cricket	108	72	6	1197	12	AccelerometerGyroscope
EigenWorms	128	131	6	17984	5	Other
Epilepsy	137	138	3	206	4	AccelerometerGyroscope
EthanolConcentration	261	263	3	1751	4	Other
ERing	30	270	4	65	6	Other
FaceDetection	5890	3524	144	62	2	ElectricBiosignals
FingerMovements	316	100	28	50	2	ElectricBiosignals
HandMovementDirection	160	74	10	400	4	ElectricBiosignals
Handwriting	150	850	3	152	26	AccelerometerGyroscope
Heartbeat	204	205	61	405	2	Audio
Libras	180	180	2	45	15	Coordinates
LSST	2459	2466	6	36	14	Other
MotorImagery	278	100	64	3000	2	ElectricBiosignals
NATOPS	180	180	24	51	6	AccelerometerGyroscope
PenDigits	7494	3498	2	8	10	Coordinates
PEMS-SF	267	173	963	144	7	Other
RacketSports	151	152	6	30	4	AccelerometerGyroscope
SelfRegulationSCP1	268	293	6	896	2	ElectricBiosignals
SelfRegulationSCP2	200	180	7	1152	2	ElectricBiosignals
StandWalkJump	12	15	4	2500	3	ElectricBiosignals
UWaveGestureLibrary	120	320	3	315	8	AccelerometerGyroscope

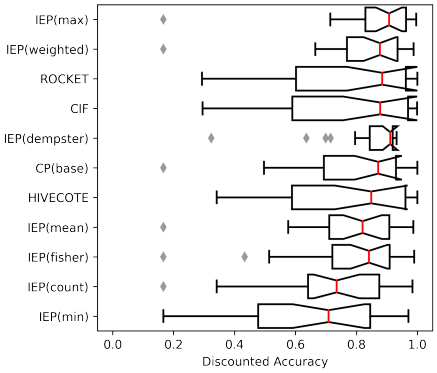


Figure A1: Notched boxplots for the performance of the considered classifiers, in terms of discounted accuracy.

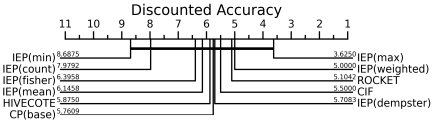


Figure A2: Critical difference plots of the average ranks of the considered classifiers, in terms of discounted accuracy.

category and on most datasets from the Other category. By contrast, ROCKET was better than IEP(weighted) on the Accelerometer/Gyroscope category, as well as on the Coordinates category. This shows that the improvement reported by EP methods over ROCKET was relatively consistent across different types of time series.

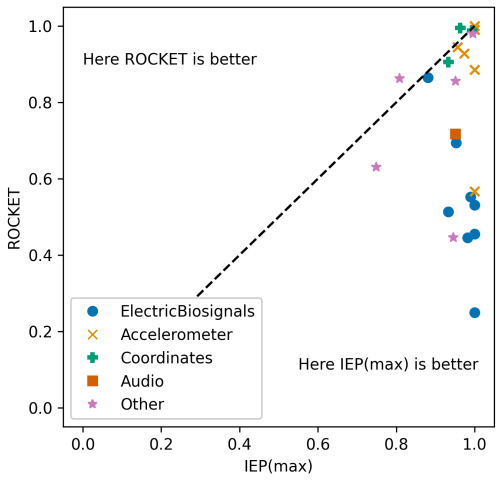


Figure A3: Comparison between IEP(max) and ROCKET, in terms of rejection-discounted accuracy.

Pairwise comparisons of IEP(max) and IEP(weighted) against the third-best CP combination method (i.e., ICP(fisher), in terms of RA and efficiency, are reported in Figures A7, A8, A9, A10. IEP(max) was consistently better than ICP(fisher), on all datasets, in respect to RA but consistently worse in terms of efficiency. By contrast, the comparison between IEP(weighted) and ICP(fisher) was more balanced: while IEP(weighted) reported better efficiency on

Table A2: The results of the experimental analysis, in terms of rejection-discounted accuracy.

	CIF	CP(base)	HIVECOTE	IEP(count)	IEP(dempster)	IEP(fisher)	IEP(max)	IEP(mean)	IEP(min)	IEP(weighted)	ROCKET
ArticularyWordRecognition	0.98	0.99	0.98	0.73	0.90	0.98	0.99	0.87	0.43	0.92	0.99
AtrialFibrillation	0.25	1.00	0.29	1.00	0.89	1.00	1.00	1.00	1.00	1.00	0.25
BasicMotions	1.00	1.00	1.00	0.98	0.91	0.99	1.00	0.97	0.93	0.98	0.99
Cricket	0.98	1.00	0.99	0.95	0.90	0.98	1.00	0.95	0.78	0.95	1.00
EigenWorms	0.90	0.87	0.78	0.71	0.89	0.70	0.81	0.85	0.76	0.85	0.86
Epilepsy	0.98	0.99	1.00	0.94	0.92	0.93	1.00	0.95	0.93	0.93	0.99
EthanolConcentration	0.73	0.44	0.81	0.44	0.90	0.86	0.95	0.88	0.55	0.88	0.45
ERing	0.96	0.90	0.94	0.98	0.87	0.91	1.00	0.98	0.90	0.99	0.98
FaceDetection	0.69	0.69	0.69	0.60	0.90	0.73	0.95	0.64	0.49	0.71	0.69
FingerMovements	0.54	0.54	0.54	0.42	0.89	0.91	0.99	0.62	0.28	0.61	0.55
HandMovementDirection	0.52	0.46	0.38	0.29	0.92	0.96	0.98	0.52	0.24	0.76	0.45
Handwriting	0.35	1.00	0.50	1.00	0.91	1.00	1.00	1.00	1.00	1.00	0.57
Heartbeat	0.77	0.76	0.72	0.41	0.86	0.49	0.95	0.75	0.39	0.77	0.72
Libras	0.92	0.84	0.90	0.67	0.83	0.86	0.93	0.65	0.63	0.93	0.91
LSST	0.56	0.62	0.54	0.55	0.86	0.63	0.75	0.65	0.41	0.65	0.63
MotorImagery	0.52	0.51	0.52	1.00	0.90	1.00	1.00	1.00	1.00	1.00	0.53
NATOPS	0.84	0.89	0.83	0.60	0.92	0.83	1.00	0.73	0.34	0.73	0.89
PenDigits	0.99	0.98	0.98	0.95	0.90	0.96	0.96	0.96	0.96	0.95	1.00
PEMS-SF	1.00	0.88	0.98	0.96	0.90	0.88	0.95	0.91	0.96	0.91	0.86
RacketSports	0.89	0.93	0.91	0.74	0.89	0.88	0.99	0.75	0.52	0.85	0.93
SelfRegulationSCP1	0.86	0.87	0.86	0.82	0.77	0.82	0.88	0.87	0.86	0.87	0.87
SelfRegulationSCP2	0.49	0.76	0.52	0.76	0.85	0.82	0.93	0.84	0.77	0.86	0.51
StandWalkJump	0.45	1.00	0.41	1.00	0.89	1.00	1.00	1.00	1.00	1.00	0.46
UWaveGestureLibrary	0.92	0.96	0.91	0.64	0.91	0.83	0.96	0.77	0.65	0.90	0.94

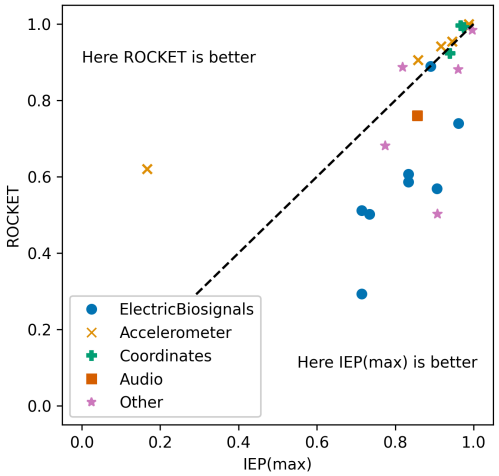


Figure A4: Comparison between IEP(max) and ROCKET, in terms of discounted accuracy.

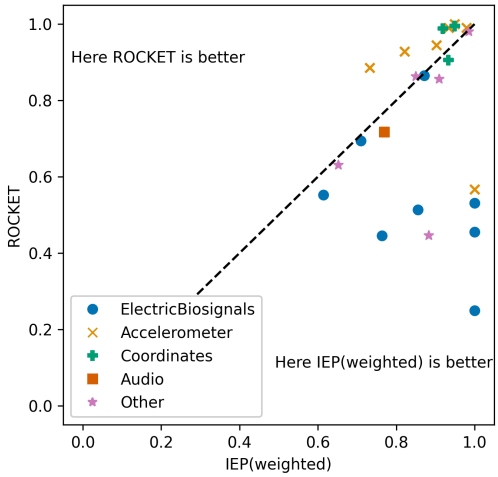


Figure A5: Comparison between IEP(weighted) and ROCKET, in terms of rejection-discounted accuracy.

most datasets, no specific trend could be found with respect to RA, while IEP(weighted) was on average better than ICP(fisher).

Validity and Efficiency Analysis In Figures A11, A12, A13, A14 and A15 and in Figures A16, A17, A18, A19 and A20 we report, respectively, on the validity and efficienct diagrams for the considered EP and CP combination models, for the 5 datasets (one for each dataset category) having higher variance in terms of discounted accuracy (that is, the datasets with more variability in models' performance). In Figures A21 and A22, by contrast, we show the deviation from strong validity for all the considered EP and CP methods, both in terms of boxplots, as well as ranks.

Remarkably, none of the considered EPs and CP combination methods were strongly valid: in light of the theoretical results in Section , the observed lack of validity could be explained as arising from the p-value functions of the aggregated CPs not being co-monotone. Nonetheless, all EPs except IEP(min) and IEP(dempster) were conservatively valid, as their accuracy was always higher than $1 - \epsilon$, for every threshold value ϵ. In particular, IEP(max) reported the highest accuracy values across all considered values of ϵ: this result is not surprising in light of Theorem 4, which shows that the max-based EP is always conservatively valid.

By contrast, in regard to efficiency, IEP(min) reported the best performance across all considered datasets, having the highest level of efficiency among the considered CP combination methods at all values of the threshold ϵ. While in general less efficient than IEP(min), also IEP(mean) and

Table A3: The results of the experimental analysis, in terms of efficiency.

	IEP(min)	IEP(max)	IEP(mean)	IEP(weighted)	IEP(dempster)	IEP(count)	IEP(fisher)	CP(base)
ArticularyWordRecognition	0.96	0.87	0.96	0.96	0.95	0.96	0.96	1.00
AtrialFibrillation	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00
BasicMotions	0.75	0.67	0.75	0.75	0.72	0.75	0.75	0.69
Cricket	0.92	0.86	0.92	0.92	0.94	0.92	0.92	1.00
EigenWorms	0.80	0.66	0.85	0.87	0.79	0.80	0.66	0.89
Epilepsy	0.75	0.69	0.75	0.75	0.74	0.75	0.75	NaN
EthanolConcentration	0.75	0.53	0.75	0.73	0.73	0.75	0.69	0.75
ERing	0.92	0.86	0.92	0.92	0.90	0.92	0.92	1.00
FaceDetection	0.75	0.70	0.75	0.75	0.72	0.75	0.75	0.75
FingerMovements	0.50	0.01	0.50	0.50	0.50	0.50	0.09	0.50
HandMovementDirection	0.00	0.11	0.75	0.75	0.04	0.75	0.12	1.00
Handwriting	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
Heartbeat	0.50	0.11	0.50	0.55	0.50	0.50	0.11	0.25
Libras	0.93	0.90	0.93	0.90	0.91	0.93	0.93	0.93
LSST	0.93	0.83	0.93	0.93	0.93	0.93	0.91	1.00
MotorImagery	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
NATOPS	0.83	0.38	0.83	0.83	0.82	0.83	0.80	0.82
PenDigits	0.90	0.88	0.90	0.90	0.90	0.90	0.90	0.88
PEMS-SF	0.90	0.88	0.90	0.90	0.88	0.90	0.00	0.75
RacketSports	0.75	0.49	0.75	0.75	0.75	0.75	0.75	0.75
SelfRegulationSCP1	0.50	0.43	0.50	0.50	0.48	0.50	0.50	0.43
SelfRegulationSCP2	0.50	0.31	0.49	0.47	0.48	0.50	0.23	0.50
StandWalkJump	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00
UWaveGestureLibrary	0.88	0.78	0.88	0.88	0.85	0.88	0.88	0.88

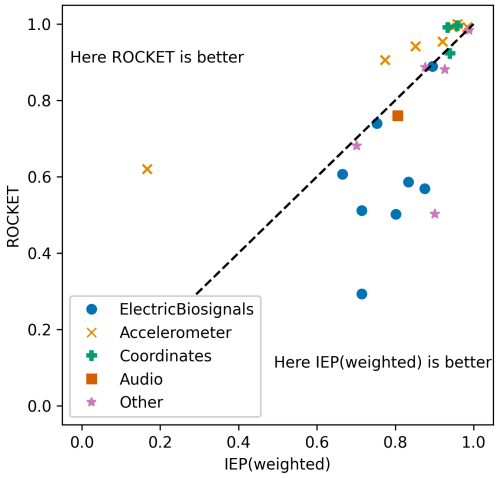


Figure A6: Comparison between IEP(weighted) and ROCKET, in terms of discounted accuracy.

IEP(weighted) reported good efficiency, comparable with that of ICP(count) and greater than ICP(fisher) and IEP(max).

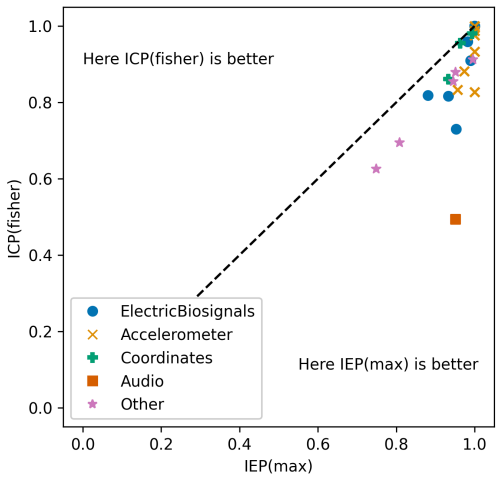


Figure A7: Comparison between IEP(max) and ICP(fisher), in terms of rejection-discounted accuracy.

Table A4: The results of the experimental analysis, in terms of discounted accuracy.

	CIF	CP(base)	HIVECOTE	IEP(count)	IEP(dempster)	IEP(fisher)	IEP(max)	IEP(mean)	IEP(min)	IEP(weighted)	ROCKET
ArticularyWordRecognition	0.98	0.99	0.98	0.77	0.91	0.98	0.98	0.89	0.48	0.93	0.99
AtrialFibrillation	0.30	0.71	0.34	0.71	0.70	0.71	0.71	0.71	0.71	0.71	0.29
BasicMotions	1.00	0.99	1.00	0.98	0.92	0.99	0.98	0.98	0.94	0.98	0.99
Cricket	0.99	1.00	0.99	0.96	0.92	0.98	0.99	0.96	0.82	0.96	1.00
EigenWorms	0.92	0.89	0.82	0.76	0.91	0.72	0.82	0.88	0.80	0.88	0.89
Epilepsy	0.99	NaN	1.00	0.95	0.93	0.95	0.99	0.96	0.95	0.95	0.99
EthanolConcentration	0.77	0.50	0.84	0.50	0.92	0.87	0.91	0.90	0.61	0.90	0.50
ERing	0.96	0.92	0.95	0.98	0.89	0.93	1.00	0.99	0.92	0.99	0.98
FaceDetection	0.73	0.74	0.74	0.65	0.92	0.77	0.96	0.69	0.55	0.75	0.74
FingerMovements	0.59	0.59	0.59	0.47	0.91	0.82	0.83	0.67	0.33	0.67	0.61
HandMovementDirection	0.58	0.51	0.43	0.34	0.64	0.73	0.73	0.58	0.24	0.80	0.50
Handwriting	0.40	0.17	0.56	0.17	0.32	0.17	0.17	0.17	0.17	0.17	0.62
Heartbeat	0.80	0.76	0.76	0.46	0.89	0.51	0.86	0.79	0.45	0.81	0.76
Libras	0.93	0.87	0.92	0.71	0.86	0.89	0.94	0.70	0.68	0.94	0.92
LSST	0.62	0.67	0.59	0.61	0.88	0.68	0.77	0.70	0.47	0.70	0.68
MotorImagery	0.57	0.51	0.58	0.83	0.80	0.83	0.83	0.83	0.83	0.83	0.59
NATOPS	0.87	0.91	0.86	0.66	0.93	0.85	0.86	0.77	0.39	0.77	0.91
PenDigits	0.99	0.98	0.98	0.96	0.92	0.96	0.97	0.97	0.97	0.96	1.00
PEMS-SF	1.00	0.88	0.98	0.97	0.92	0.43	0.96	0.93	0.97	0.93	0.88
RacketSports	0.91	0.94	0.92	0.78	0.91	0.90	0.93	0.79	0.57	0.88	0.94
SelfRegulationSCP1	0.88	0.88	0.88	0.85	0.80	0.85	0.89	0.89	0.88	0.89	0.89
SelfRegulationSCP2	0.54	0.80	0.57	0.80	0.87	0.80	0.91	0.87	0.81	0.88	0.57
StandWalkJump	0.51	0.71	0.46	0.71	0.72	0.71	0.71	0.71	0.71	0.71	0.51
UWaveGestureLibrary	0.94	0.96	0.93	0.69	0.92	0.86	0.95	0.81	0.70	0.92	0.95

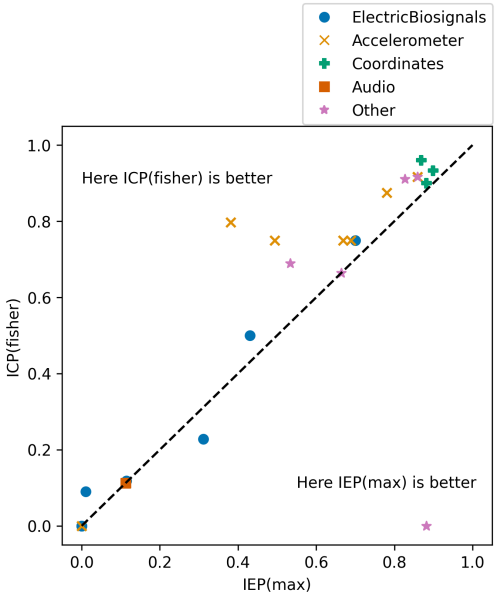


Figure A8: Comparison between IEP(max) and ICP(fisher), in terms of efficiency.

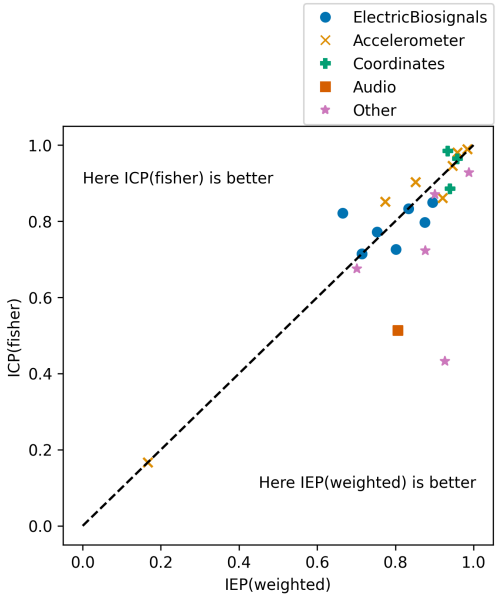


Figure A9: Comparison between IEP(weighted) and ICP(fisher), in terms of rejection-discounted accuracy.

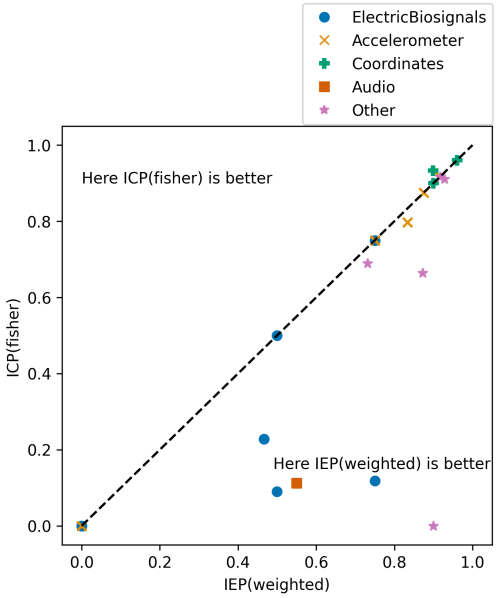


Figure A10: Comparison between IEP(weighted) and ICP(fisher), in terms of efficiency.

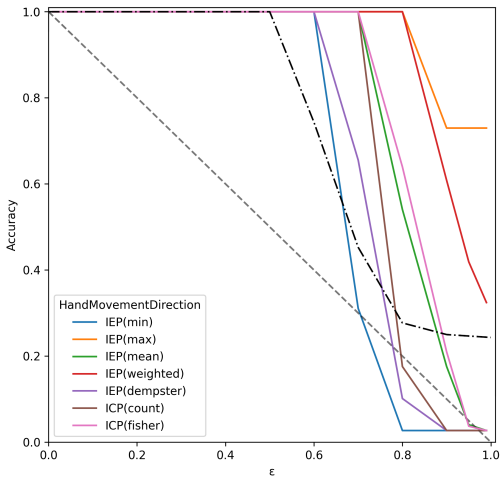


Figure A12: Validity diagram for the HandMovementDirection dataset: for each EP and CP combination method, we report the accuracy as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

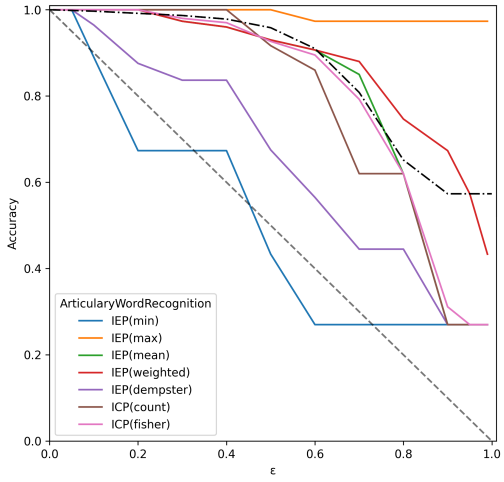


Figure A11: Validity diagram for the ArticularyWordRecognition dataset: for each EP and CP combination method, we report the accuracy as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

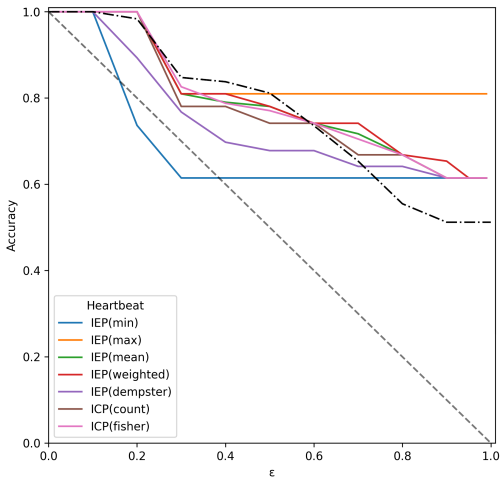


Figure A13: Validity diagram for the Heartbeat dataset: for each EP and CP combination method, we report the accuracy as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

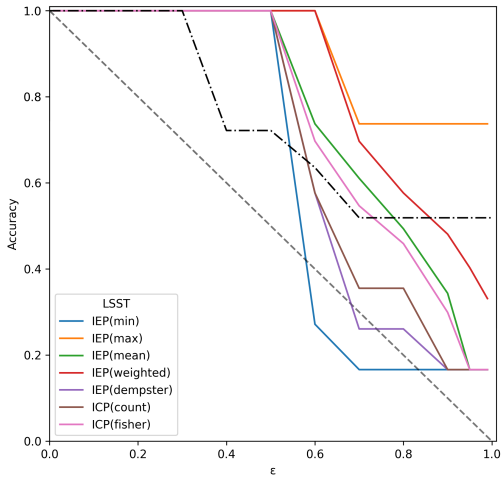


Figure A14: Validity diagram for the LSST dataset: for each EP and CP combination method, we report the accuracy as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

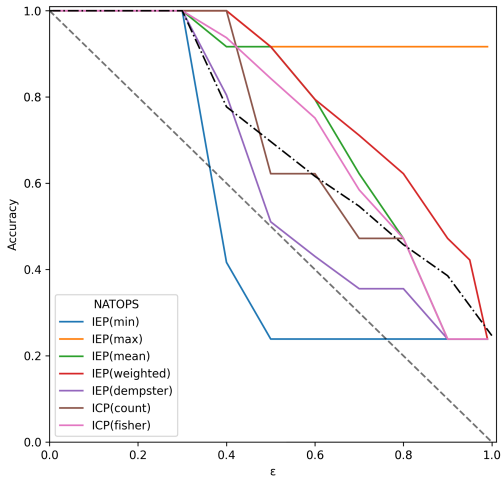


Figure A15: Validity diagram for the NATOPS dataset: for each EP and CP combination method, we report the accuracy as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

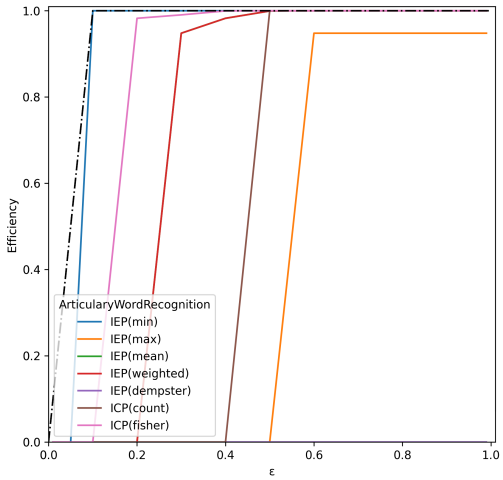


Figure A16: Efficiency diagram for the ArticularyWordRecognition dataset: for each EP and CP combination method, we report the efficiency as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

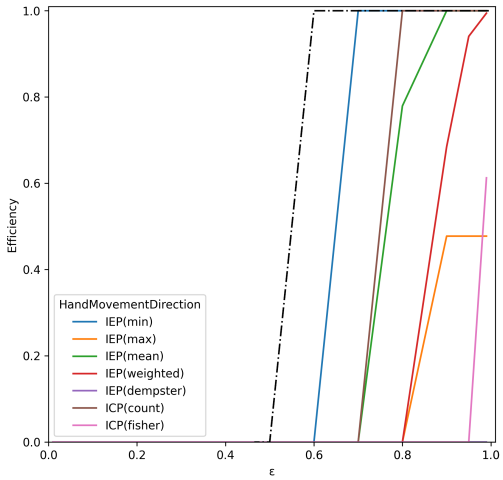


Figure A17: Efficiency diagram for the HandMovementDirection dataset: for each EP and CP combination method, we report the efficiency as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

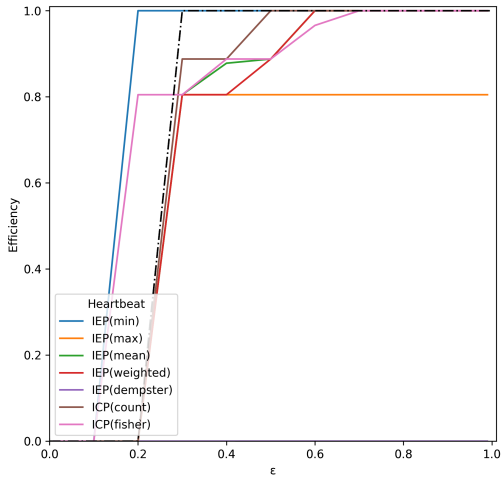


Figure A18: Efficiency diagram for the Heartbeat dataset: for each EP and CP combination method, we report the efficiency as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

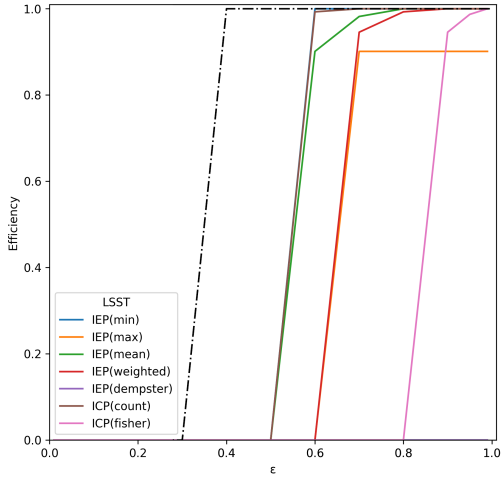


Figure A19: Efficiency diagram for the LSST dataset: for each EP and CP combination method, we report the efficiency as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

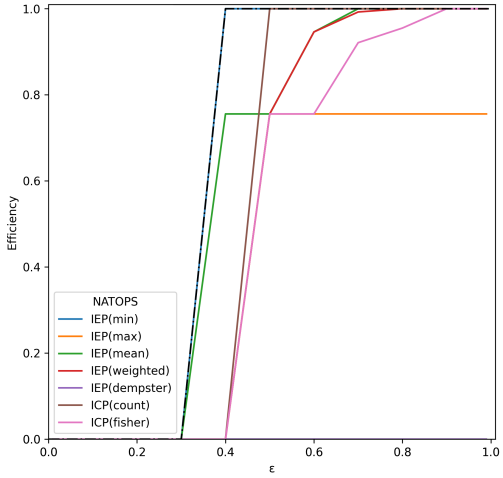


Figure A20: Efficiency diagram for the NATOPS dataset: for each EP and CP combination method, we report the efficiency as a function of ϵ, the complement of the confidence level. The dash-dotted line represents the baseline CP(base).

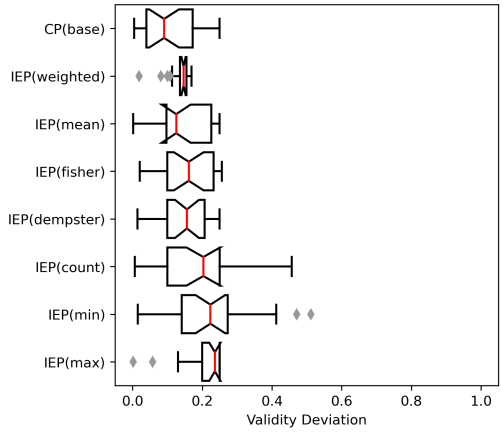


Figure A21: Notched boxplots for the performance of the con- sidered classifiers, in terms of deviation from strong validity.

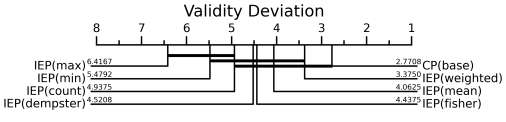


Figure A22: Critical difference plots of the average ranks of the considered classifiers in terms of deviation from strong validity.

Chapter 7

Conclusions

In this thesis, the main aim has been to study methods and techniques to handle imprecision in ML, focusing both on setting where imprecision affects the input data of a ML pipeline (what has been denoted with the term *learning from imprecise data*) as well as on settings where imprecision is adopted as a form of uncertainty quantification (what has been denoted with the term *cautious inference*).

In the first setting, first a theoretical characterization of a relevant instance of the problem of learning from imprecise data, namely learning from fuzzy labels, has been proposed. In this setting, the obtained results have provided a characterization of the learning properties of three algorithms belonging to widely adopted ML paradigms, namely generalized risk minimization, instance-based learning, and pseudo-label learning. In particular, it has been shown that learnability in this setting, compared with the standard supervised one, is only possible conditional on some properties of the data generating distribution, which characterize the hardness of learning problem instance. Furthermore, a novel pseudo-label learning algorithm, called RRL, has been proposed and has been shown to improve the classification performance as compared with other state-of-the-art ML methods for learning from fuzzy labels. Then, the problem of feature selection and dimensionality reduction with imprecise data has been studied, with the aim of overcoming the curse of dimensionality and improve the generalization of ML methods in this setting. To this aim, a novel feature selection approach, based on Rough Set theory, has been proposed and

it has been shown that such an approach allows to improve the accuracy of ML algorithms trained on the reduced datasets, compared with the state-of-the-art. Finally, to provide a better context for the novelty and usefulness of the developed techniques and approaches, the application of the proposed techniques in three paradigmatic real-world settings, arising within the medical domain, has been illustrated.

In the second setting, on the other hand, the first aim has been to study the theoretical properties of a cautious inference method, three-way decision, as well as its relationship with two other such paradigms, namely selective prediction and conformal prediction. Secondly, the ensembling of cautious inference models has been studied. In particular, contributions in this setting encompass theoretical results on the validity and efficiency properties of ensemble methods for cautious inference, in a general setting that extends previous results in the existing literature, as well as an extensive experimental analysis which showed the effectiveness of such techniques in comparison with state-of-the-art ensemble models in general benchmarks as well as in relevant application tasks, focusing on the task of multi-variate time series classification. Furthermore, some initial, proof-of-concept results devoted to the evaluation of cautious inference methods from the point of view of human-AI interaction has been presented, providing promising indications in this sense.

As a further remark and contribution, in order to support further research in the above mentioned areas, as well as to motivate the development of tools and frameworks oriented to the real use in applications, all code developed and employed within this thesis has been made publicly available, under open source licenses, mainly through two libraries:

- scikit-weak, available at the URL `https://github.com/AndreaCampagner/scikit-weak`, and devoted to applications of ML in the setting of learning from imprecise data;

- scikit-cautious, available at the URL `https://github.com/AndreaCampagner/scikit-cautious`, and devoted to uncertainty quantification methods and cautious inference in ML.

Similarly, all datasets that have been collected and used for the articles appearing this thesis, have been publicly released on the open access archive Zenodo `https://zenodo.org/`. Aside from a personal belief in the principles of open science, the hope is that the availability of these tools and datasets will stimulate further research related to the representation and management of imprecision, and more generally uncertainty, in ML, as well as relevant applications thereof.

Future Research Directions

This thesis opens up to different research directions and possible future work:

- As already mentioned, Chapter 2 focused on the problem of learning from fuzzy labels, a practically relevant but specific instance of the more general problem of learning from imprecise data. The extension of the theoretical results proven in this work to more general forms of imprecise data would be of interest for further understanding the limits and characteristics of learnability in these settings. To this aim, two particularly promising research directions would be to investigate the problem of learning from fuzzy data, which has been discussed in Sections 4.2 and 4.3, as well as the problem of learning from comparative probabilities [54]. On the one hand, the study of the problem of learning from fuzzy data would extend the applicability of the proposed RRL algorithm, as well as other state-of-the-art methods for learning from fuzzy labels, to more general settings in which imprecision affects not only the target supervision but also the feature values. On the other hand, the problem of learning from comparative probabilities represents a particularly interesting generalization from a complexity-theoretic perspective, due to the relationships between comparative probabilities and the theory of credal sets [170], which would enable the analysis of learning from imprecise labels problems within the framework of convex optimization [35], one of the most effective computational paradigms in modern ML theory [219];

- Chapter 5 studied the relationships between three-way decision, selective pre-

diction and conformal prediction: further work should be devoted at exploring the connections between cautious inference techniques as well as neighboring methods for uncertainty quantification and management. On the one hand, selective prediction has a rich theory, with deep connections with the fields of active learning and machine teaching [111], as well as with the problem of learning over-parameterized models [5] and adversarial learning [113]. Such relationships could be further investigated in light of the correspondence between selective prediction and three-way decision (or, more in general, decision-theoretic cautious inference methods). Similarly, the correspondence between conformal prediction and uncertainty quantification frameworks such as game-theoretic probability [218] and imprecise probabilities [61, 62] could motivate the study of relationships between cautious inference methods based on decision-theoretic methods and those based on other uncertainty representation theories, as well as the investigation of the theoretical properties of hybrid approaches that employ the extension of decision-theoretic principles to settings more general than probability theory [166, 263];

- Finally, in Chapter 5, a proof-of-concept, small-sample study to assess the practical utility of cautious inference from a user-oriented perspective has been presented. A particularly interesting research direction would be to further investigate these issues and, more in general, the impact of the introduction of uncertainty quantification methods in socio-technical systems, not only in terms of improved accuracy [12, 32], but also as it relates to more psychometric dimensions, such as user acceptance and appropriation, or pragmatic utility.

Bibliography

[1] Aasne K Aarsand, Ann Helen Kristoffersen, Sverre Sandberg, et al. "The European Biological Variation Study (EuBIVAS): Biological Variation Data for Coagulation Markers Estimated by a Bayesian Model". In: *Clinical Chemistry* 67.9 (2021), pp. 1259–1270.

[2] Görkem Algan and Ilkay Ulusoy. "Image classification with deep learning in the presence of noisy labels: A survey". In: *Knowl Based Syst* (2021), p. 106771.

[3] Rahman Ali, Muhammad Hameed Siddiqi, and Sungyoung Lee. "Rough set-based approaches for discretization: a compact review". In: *Artificial Intelligence Review* 44.2 (2015), pp. 235–263.

[4] Roohallah Alizadehsani et al. "Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020)". In: *Annals of Operations Research* (2021), pp. 1–42.

[5] Noga Alon et al. "A theory of PAC learnability of partial concept classes". In: *arXiv preprint arXiv:2107.08444* (2021).

[6] Jonathan Alvarsson et al. "Predicting with confidence: using conformal prediction in drug discovery". In: *Journal of Pharmaceutical Sciences* 110.1 (2021), pp. 42–49.

[7] Anastasios N Angelopoulos and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: *arXiv preprint arXiv:2107.07511* (2021).

[8] Dana Angluin and Philip Laird. "Learning from noisy examples". In: *Machine Learning* 2.4 (1988), pp. 343–370.

[9] Richard Arratia and Louis Gordon. "Tutorial on large deviations for the binomial distribution". In: *Bulletin of mathematical biology* 51.1 (1989), pp. 125–131.

[10] Thomas M Atkinson et al. "What Do "None,"‘Mild,"‘Moderate,"‘Severe," and "Very Severe" Mean to Patients With Cancer? Content Validity of PRO-CTCAE™ Response Scales". In: *Journal of pain and symptom management* 55.3 (2018), e3–e6.

[11] Arkar Min Aung and Jacob Whitehill. "Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 166–170.

[12] Varun Babbar, Umang Bhatt, and Adrian Weller. "On the Utility of Prediction Sets in Human-AI Teams". In: *arXiv preprint arXiv:2205.01411* (2022).

[13] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[14] Vineeth N Balasubramanian, Shayok Chakraborty, and Sethuraman Panchanathan. "Conformal predictions for information fusion". In: *Annals of Mathematics and Artificial Intelligence* 74.1 (2015), pp. 45–65.

[15] Maria-Florina Balcan and Avrim Blum. "A discriminative model for semi-supervised learning". In: *Journal of the ACM (JACM)* 57.3 (2010), pp. 1–46.

[16] Wei-Xuan Bao, Jun-Yi Hang, and Min-Ling Zhang. "Partial label dimensionality reduction via confidence-based dependence maximization". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 46–54.

[17] Wei-Xuan Bao, Jun-Yi Hang, and Min-Ling Zhang. "Submodular Feature Selection for Partial Label Learning". In: KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 26–34. ISBN: 9781450393850.

[18] Marılia Barandas et al. "Uncertainty-Based Rejection in Machine Learning: Implications for Model Development and Interpretability". In: *Electronics* 11.3 (2022), p. 396.

[19] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[20] Valerio Basile. "It's the End of the Gold Standard as We Know It". In: *International Conference of the Italian Association for Artificial Intelligence*. Springer. 2020, pp. 441–453.

[21] Valerio Basile et al. "Toward a perspectivist turn in ground truthing for predictive computing". In: *arXiv preprint arXiv:2109.04270* (2021).

[22] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." In: *Journal of machine learning research* 7.11 (2006).

[23] Rafael Bello and Rafael Falcon. "Rough sets in machine learning: A review". In: *Thriving Rough Sets*. Springer, 2017, pp. 87–118.

[24] Alessio Benavoli, Giorgio Corani, and Francesca Mangili. "Should we really use post-hoc tests based on mean-ranks?" In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 152–161.

[25] Kristin Bennett and Ayhan Demiriz. "Semi-supervised support vector machines". In: *Advances in Neural Information processing systems* 11 (1998).

[26] James C Bezdek, Siew K Chuah, and David Leep. "Generalized k-nearest neighbor rules". In: *Fuzzy Sets and Systems* 18.3 (1986), pp. 237–256.

[27] Avradeep Bhowmik. "Learning from aggregated data". PhD thesis. The University of Texas at Austin, 2019.

[28] Gérard Biau, Luc Devroye, and Gäbor Lugosi. "Consistency of random forests and other averaging classifiers." In: *Journal of Machine Learning Research* 9.9 (2008).

[29] Emily Black, Klas Leino, and Matt Fredrikson. "Selective Ensembles for Consistent Predictions". In: *arXiv preprint arXiv:2111.08230* (2021).

[30] Avrim Blum and Maria-Florina Balcan. "Open problems in efficient semi-supervised PAC learning". In: *International Conference on Computational Learning Theory*. Springer. 2007, pp. 622–624.

[31] Philip J Boland. "Majority systems and the Condorcet jury theorem". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 38.3 (1989), pp. 181–189.

[32] Elizabeth Bondi et al. "Role of Human-AI Interaction in Selective Prediction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 5286–5294.

[33] Nicolas Bosc et al. "Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery". In: *Journal of cheminformatics* 11.1 (2019), pp. 1–16.

[34] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[35] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[36] Daren C Brabham. *Crowdsourcing*. Mit Press, 2013.

[37] Steven Brams and Peter C Fishburn. *Approval voting*. Springer Science & Business Media, 2007.

[38] Vivien A Cabannes, Francis Bach, and Alessandro Rudi. "Disambiguation of Weak Supervision leading to Exponential Convergence rates". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1147–1157.

[39] Federico Cabitza. “Cobra AI: Exploring Some Unintended Consequences”. In: *Machines We Trust: Perspectives on Dependable AI* (2021), p. 87.

[40] Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. “As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI”. In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–21.

[41] Federico Cabitza and Davide Ciucci. “Fuzzification of Ordinal Classes. The Case of the HL7 Severity Grading”. In: *International Conference on Scalable Uncertainty Management*. Springer. 2018, pp. 64–77.

[42] Federico Cabitza, Davide Ciucci, and Raffaele Rasoini. “A giant with feet of clay: On the validity of the data that feed machine learning in medicine”. In: *Organizing for the digital world*. Springer, 2019, pp. 121–136.

[43] Federico Cabitza et al. “Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests”. In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 59.2 (2021), pp. 421–431.

[44] Joyce Cahoon and Ryan Martin. “Generalized inferential models for censored data”. In: *International Journal of Approximate Reasoning* 137 (2021), pp. 51–66.

[45] Andrea Campagner, Federico Cabitza, and Davide Ciucci. “The three-way-in and three-way-out framework to treat and exploit ambiguity in data”. In: *International Journal of Approximate Reasoning* 119 (2020), pp. 292–312.

[46] Andrea Campagner, Federico Cabitza, and Davide Ciucci. “Three-way decision for handling uncertainty in machine learning: a narrative review”. In: *International Joint Conference on Rough Sets*. Springer. 2020, pp. 137–152.

[47] Andrea Campagner, Federico Cabitza, and Davide Ciucci. “Three–Way Decision for Handling Uncertainty in Machine Learning: a Narrative Review”. In:

Proceedings of International Joint Conference on Rough Sets 2020. Vol. 12179. LNCS. Springer, 2020, pp. 137–152.

[48] Andrea Campagner, Davide Ciucci, and Federico Cabitza. "Ensemble Learning, Social Choice and Collective Intelligence". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Springer. 2020, pp. 53–65.

[49] Andrea Campagner, Davide Ciucci, and Thierry Denœux. "Belief functions and rough sets: Survey and new insights". In: *International Journal of Approximate Reasoning* 143 (2022), pp. 192–215.

[50] Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, et al. "Ground truthing from multi-rater labeling with three-way decision and possibility theory". In: *Information Sciences* 545 (2021), pp. 771–790.

[51] Andrea Campagner et al. "Assessment and prediction of spine surgery invasiveness with machine learning techniques". In: *Computers in Biology and Medicine* 121 (2020), p. 103796.

[52] Andrea Campagner et al. "Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches". In: *Information Sciences* 579 (2021), pp. 347–367.

[53] Cassio P de Campos and Alessandro Antonucci. "Imprecision in Machine Learning and AI". In: *IEEE Intelligent Informatics Bulletin* 16.1 (2015), pp. 20–23.

[54] Andrea Capotorti and Andrea Formisano. "Comparative uncertainty: theory and automation". In: *Mathematical Structures in Computer Science* 18.1 (2008), pp. 57–79.

[55] Algo Carè et al. "A study on majority-voting classifiers with guarantees on the probability of error". In: *IFAC-PapersOnLine* 53.2 (2020), pp. 1013–1018.

[56] Giuseppe Carleo et al. "Machine learning and the physical sciences". In: *Reviews of Modern Physics* 91.4 (2019), p. 045002.

[57] Lars Carlsson, Martin Eklund, and Ulf Norinder. "Aggregated conformal prediction". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer. 2014, pp. 231–240.

[58] Anna Carobene, Andrea Campagner, Christian Uccheddu, et al. "The multicenter European Biological Variation Study (EuBIVAS): a new glance provided by the Principal Component Analysis (PCA), a machine learning unsupervised algorithms, based on the basic metabolic panel linked measurands". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* (2021).

[59] Marco Cattaneo. "M-Estimation with Imprecise Data". In: *Ninth International Symposium on Imprecise Probability: Theories and Applications*. 2015, p. 335.

[60] Marco Cattaneo and Andrea Wiencierz. "On the validity of minimin and minimax methods for Support Vector Regression with interval data". In: *Ninth International Symposium on Imprecise Probability: Theories and Applications*. 2015.

[61] Leonardo Cella and Ryan Martin. "Valid inferential models for prediction in supervised learning problems". In: *International Journal of Approximate Reasoning* 150 (2022), pp. 1–18.

[62] Leonardo Cella and Ryan Martin. "Validity, consonant plausibility measures, and conformal prediction". In: *International Journal of Approximate Reasoning* 141 (2022), pp. 110–130.

[63] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. 2006.

[64] Zohra L Cherfi et al. "Partially supervised independent factor analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis". In: *Soft computing* 16.5 (2012), pp. 741–754.

[65] Giovanni Cherubin. "Majority vote ensembles of conformal predictors". In: *Machine Learning* 108.3 (2019), pp. 475–488.

[66] C Chow. "On optimum recognition error and reject tradeoff". In: *IEEE Transactions on information theory* 16.1 (1970), pp. 41–46.

[67] Evgenii Chzhen et al. "Set-valued classification–overview via a unified framework". In: *arXiv preprint arXiv:2102.12318* (2021).

[68] Davide Ciucci and Ivan Forcati. "Certainty-based rough sets". In: *International Joint Conference on Rough Sets*. Springer. 2017, pp. 43–55.

[69] William W Cohen. "Learning trees and rules with set-valued features". In: *AAAI/IAAI, Vol. 1*. 1996, pp. 709–716.

[70] Etienne Côme et al. "Learning from partially supervised data using mixture models and belief functions". In: *Pattern recognition* 42.3 (2009), pp. 334–348.

[71] Bryan Conroy et al. "A dynamic ensemble approach to robust classification in the presence of missing data". In: *Machine Learning* 102.3 (2016), pp. 443–463.

[72] Giorgio Corani and Alessandro Antonucci. "Credal ensembles of classifiers". In: *Computational Statistics & Data Analysis* 71 (2014), pp. 818–831.

[73] Giorgio Corani and Marco Zaffalon. "Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2." In: *Journal of Machine Learning Research* 9.4 (2008).

[74] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. "Boosting with abstention". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1660–1668.

[75] Timothee Cour, Ben Sapp, and Ben Taskar. "Learning from partial labels". In: *The Journal of Machine Learning Research* 12 (2011), pp. 1501–1536.

[76] Ines Couso and Luciano Sánchez. "Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach". In: *Information Sciences* 358 (2016), pp. 129–150.

[77] Ines Couso et al. “Fuzzy sets in data analysis: From statistical foundations to machine learning”. In: *IEEE Computational Intelligence Magazine* 14.1 (2019), pp. 31–44.

[78] Inés Couso and Didier Dubois. “A general framework for maximizing likelihood under incomplete data”. In: *International Journal of Approximate Reasoning* 93 (2018), pp. 238–260.

[79] Inés Couso and Didier Dubois. “Statistical reasoning with set-valued information: Ontic vs. epistemic views”. In: *International Journal of Approximate Reasoning* 55.7 (2014), pp. 1502–1518.

[80] Inés Couso, Didier Dubois, and Eyke Hüllermeier. “Maximum likelihood estimation and coarse data”. In: *International Conference on Scalable Uncertainty Management*. Springer. 2017, pp. 3–16.

[81] Inés Couso, Didier Dubois, and Luciano Sánchez. “Random sets and random fuzzy sets as ill-perceived random variables”. In: *SpringerBriefs Computat Intell* (2014).

[82] Alexander D’Amour et al. “Underspecification presents challenges for credibility in modern machine learning”. In: *arXiv preprint arXiv:2011.03395* (2020).

[83] Amit Daniely et al. “Multiclass learnability and the erm principle”. In: *Proceedings of COLT 2011*. JMLR Workshop and Conference Proceedings. 2011, pp. 207–232.

[84] Juan José Del Coz, Jorge Dıez, and Antonio Bahamonde. “Learning Nondeterministic Classifiers.” In: *Journal of Machine Learning Research* 10.10 (2009).

[85] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *The Journal of Machine learning research* 7 (2006), pp. 1–30.

[86] Thierry Denœux. “Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence”. In: *Fuzzy Sets Syst* (2020).

[87] Thierry Denœux, Didier Dubois, and Henri Prade. "Representations of uncertainty in artificial intelligence: Probability and possibility". In: *A Guided Tour of Artificial Intelligence Research*. Springer, 2020, pp. 69–117.

[88] Thierry Denoeux. "A k-nearest neighbor classification rule based on Dempster-Shafer theory". In: *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008, pp. 737–760.

[89] Thierry Denoeux. "Maximum likelihood estimation from fuzzy data using the EM algorithm". In: *Fuzzy Sets Syst* 183.1 (2011), pp. 72–91.

[90] Joaquın Derrac, Salvador Garcıa, and Francisco Herrera. "Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects". In: *Inf Sci* 260 (2014), pp. 98–119.

[91] Joaquın Derrac et al. "Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets". In: *Information Sciences* 329 (2016), pp. 144–163.

[92] Sébastien Destercke. "Uncertain data in learning: challenges and opportunities". In: *Conformal and Probabilistic Prediction with Applications* (2022), pp. 322–332.

[93] Raul Diaz and Amit Marathe. "Soft labels for ordinal regression". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4738–4747.

[94] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114.

[95] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in Finance*. Vol. 1170. Springer, 2020.

[96] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[97] Didier Dubois and Henri Prade. "Ontic vs. epistemic fuzzy sets in modeling and data processing tasks". In: *3rd International Joint Conference on Computational Intelligence*. SciTePress. 2011.

[98] Didier Dubois, Henri Prade, and Sandra Sandri. "On possibility/probability transformations". In: *Fuzzy logic*. Springer, 1993, pp. 103–112.

[99] Charles Elkan. "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd. 2001, pp. 973–978.

[100] Jean Feng et al. "Selective prediction-set models with coverage rate guarantees". In: *Biometrics* (2021).

[101] Lei Feng and Bo An. "Leveraging Latent Label Distributions for Partial Label Learning." In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 2018, pp. 2107–2113.

[102] Lei Feng and Bo An. "Partial label learning with self-guided retraining". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3542–3549.

[103] Lei Feng et al. "Provably consistent partial-label learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10948–10960.

[104] José A Ferreira. "Models under which random forests perform badly; consequences for applications". In: *Computational Statistics* (2022), pp. 1–16.

[105] Cèsar Ferri and José Hernández-Orallo. "Cautious Classifiers." In: *ROCAI* 4 (2004), pp. 27–36.

[106] Callum G Fraser. *Biological variation: from principles to practice*. American Association for Clinical Chemistry, 2001.

[107] Gallum G Fraser and Eugene K Harris. "Generation and application of data on biological variation in clinical chemistry". In: *Critical reviews in clinical laboratory sciences* 27.5 (1989), pp. 409–437.

[108] Mariem Gandouz, Hajo Holzmann, and Dominik Heider. “Machine learning with asymmetric abstention for biomedical decision-making”. In: *BMC medical informatics and decision making* 21.1 (2021), pp. 1–11.

[109] Tommaso Gastaldi. “Optimal reconstruction of a generally censored sample”. In: *Statistics & probability letters* 14.5 (1992), pp. 393–399.

[110] Yonatan Geifman and Ran El-Yaniv. “Selective classification for deep neural networks”. In: *Advances in neural information processing systems*. 2017, pp. 4878–4887.

[111] Roei Gelbhart and Ran El-Yaniv. “The Relationship Between Agnostic Selective Classification, Active Learning and the Disagreement Coefficient”. In: *J. Mach. Learn. Res.* 20.33 (2019), pp. 1–38.

[112] Sally A. Goldman and Robert H. Sloan. “Can PAC learning algorithms tolerate random attribute noise?” In: *Algorithmica* 14.1 (1995), pp. 70–84.

[113] Shafi Goldwasser et al. “Beyond perturbations: Learning guarantees with arbitrary adversarial test examples”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15859–15870.

[114] Matteo Golfarelli, Dario Maio, and D Malton. “On the error-reject trade-off in biometric verification systems”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), pp. 786–796.

[115] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[116] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. “Why do tree-based models still outperform deep learning on tabular data?” In: *arXiv preprint arXiv:2207.08815* (2022).

[117] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[118] Marek Grzegorowski et al. “On the role of feature space granulation in feature selection processes”. In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 1806–1815.

[119] Rudy Guerra et al. “Meta-analysis by combining p-values: simulated linkage studies”. In: *Genetic epidemiology* 17.S1 (1999), S605–S609.

[120] Romain Guillaume and Didier Dubois. “A maximum likelihood approach to inference under coarse data based on minimax regret”. In: *International Conference Series on Soft Methods in Probability and Statistics*. Springer. 2018, pp. 99–106.

[121] Romain Guillaume and Didier Dubois. “Robust parameter estimation of density functions under fuzzy interval observations”. In: *9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA’15)*. 2015, pp. 147–156.

[122] Lars Kai Hansen, Christian Liisberg, and Peter Salamon. “The error-reject tradeoff”. In: *Open Systems & Information Dynamics* 4.2 (1997), pp. 159–184.

[123] Reihaneh H Hariri, Erik M Fredericks, and Kate M Bowers. “Uncertainty in big data analytics: survey, opportunities, and challenges”. In: *Journal of Big Data* 6.1 (2019), pp. 1–16.

[124] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[125] Kilian Hendrickx et al. “Machine learning with a reject option: A survey”. In: *arXiv preprint arXiv:2107.11277* (2021).

[126] James J Higgins. *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA, 2004.

[127] Andreas Holzinger. “Interactive machine learning for health informatics: when do we need the human-in-the-loop?” In: *Brain Informatics* 3.2 (2016), pp. 119–131.

[128] Max Hopkins et al. “Realizable learning is all you need”. In: *Conference on Learning Theory*. PMLR. 2022, pp. 3015–3069.

[129] Eyke Hüllermeier. “Does machine learning need fuzzy logic?” In: *Fuzzy Sets and Systems* 281 (2015), pp. 292–299.

[130] Eyke Hüllermeier. “Fuzzy methods in machine learning and data mining: Status and prospects”. In: *Fuzzy sets and Systems* 156.3 (2005), pp. 387–406.

[131] Eyke Hüllermeier. “Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization”. In: *International Journal of Approximate Reasoning* 55.7 (2014), pp. 1519–1534.

[132] Eyke Hüllermeier, Sébastien Destercke, and Ines Couso. “Learning from imprecise data: Adjustments of optimistic and pessimistic variants”. In: *Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compiègne, France, December 16-18, 2019, Proceedings*. Ed. by Nahla Ben Amor, Benjamin Quost, and Martin Theobald. Vol. 11940. Lecture Notes in Computer Science. Springer, 2019, pp. 266–279.

[133] Eyke Hüllermeier and Willem Waegeman. “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine Learning* 110.3 (2021), pp. 457–506.

[134] Andrzej Janusz and Dominik Ślęzak. “Rough set methods for attribute clustering and selection”. In: *Applied Artificial Intelligence* 28.3 (2014), pp. 220–242.

[135] Richard Jensen. “Rough set-based feature selection: A review”. In: *Rough computing: theories, technologies and applications* (2008), pp. 70–107.

[136] Wei Ji et al. “Learning calibrated medical image segmentation via multi-rater agreement modeling”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12341–12351.

[137] Rong Jin and Zoubin Ghahramani. “Learning with multiple labels”. In: *Advances in neural information processing systems*. 2003, pp. 921–928.

[138] Ulf Johansson et al. "Regression conformal prediction with random forests". In: *Machine learning* 97.1 (2014), pp. 155–176.

[139] Michael J Kearns. *The computational complexity of machine learning*. MIT press, 1990.

[140] Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[141] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. "Second opinion needed: communicating uncertainty in medical machine learning". In: *NPJ Digital Medicine* 4.1 (2021), pp. 1–6.

[142] Guy Kornowski and Ohad Shamir. "Oracle complexity in nonsmooth nonconvex optimization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 324–334.

[143] Bartosz Krawczyk and Alberto Cano. "Online ensemble learning with abstaining classifiers for drifting and noisy data streams". In: *Applied Soft Computing* 68 (2018), pp. 677–692.

[144] Rudolf Kruse and Klaus Dieter Meyer. *Statistics with vague data*. Vol. 6. Springer Science & Business Media, 2012.

[145] Ludmila Kuncheva. *Fuzzy classifier design*. Vol. 49. Springer Science & Business Media, 2000.

[146] Dong-Hyun Lee. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, International Conference on Machine Learning*. Vol. 3. 2013.

[147] Yonghoon Lee and Rina Barber. "Distribution-free inference for regression: discrete, continuous, and in between". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7448–7459.

[148] Jing Lei, James Robins, and Larry Wasserman. "Distribution-free prediction sets". In: *Journal of the American Statistical Association* 108.501 (2013), pp. 278–287.

[149] Jian-hua Li. "Cyber security meets artificial intelligence: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.12 (2018), pp. 1462–1474.

[150] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2018.

[151] Julian Lienen and Eyke Hüllermeier. "Credal Self-Supervised Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[152] Julian Lienen and Eyke Hüllermeier. "Instance weighting through data imprecisiation". In: *International Journal of Approximate Reasoning* 134 (2021), pp. 1–14.

[153] Henrik Linusson. "Nonconformity Measures and Ensemble Strategies: An Analysis of Conformal Predictor Efficiency and Validity". PhD thesis. Department of Computer and Systems Sciences, Stockholm University, 2021.

[154] Henrik Linusson, Ulf Johansson, and Henrik Boström. "Efficient conformal predictor ensembles". In: *Neurocomputing* 397 (2020), pp. 266–278.

[155] Henrik Linusson et al. "On the calibration of aggregated conformal predictors". In: *Conformal and probabilistic prediction and applications*. PMLR. 2017, pp. 154–173.

[156] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[157] Dun Liu. "The effectiveness of three-way classification with interpretable perspective". In: *Information Sciences* 567 (2021), pp. 237–255.

[158] Liping Liu and Thomas Dietterich. "Learnability of the superset label learning problem". In: *Proceedings of ICML 2014*. 2014, pp. 1629–1637.

[159] Liping Liu and Thomas G Dietterich. "A conditional multinomial mixture model for superset label learning". In: *Advances in neural information processing systems*. 2012, pp. 548–556.

[160] Sheng Liu et al. "Early-learning regularization prevents memorization of noisy labels". In: *Advances in neural information processing systems* 33 (2020), pp. 20331–20342.

[161] Zhun-Ga Liu et al. "Credal classification rule for uncertain data based on belief functions". In: *Pattern Recognition* 47.7 (2014), pp. 2532–2541.

[162] MIchael Loizos. "Learning from Partial Observations." In: *IJCAI*. 2007, pp. 968–974.

[163] Thomas M Loughin. "A systematic comparison of methods for combining p-values from independent tests". In: *Computational statistics & data analysis* 47.3 (2004), pp. 467–485.

[164] Michal Lukasik et al. "Does label smoothing mitigate label noise?" In: *ICML*. PMLR. 2020, pp. 6448–6458.

[165] Jiaqi Lv et al. "Progressive identification of true labels for partial-label learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6500–6510.

[166] Liyao Ma and Thierry Denoeux. "Partial classification in the belief function framework". In: *Knowledge-Based Systems* 214 (2021), p. 106742.

[167] Giorgio Manganini, Alessandro Falsone, and Maria Prandini. "A majority voting classifier with probabilistic guarantees". In: *2015 IEEE Conference on Control Applications (CCA)*. IEEE. 2015, pp. 1084–1089.

[168] Barbara Marszał-Paszek and Piotr Paszek. "Classifiers based on nondeterministic decision rules". In: *Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam*. Springer, 2013, pp. 445–454.

[169] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[170] Enrique Miranda and Sébastien Destercke. "Extreme points of the credal sets generated by comparative probabilities". In: *Journal of Mathematical Psychology* 64 (2015), pp. 44–57.

[171] Tom M Mitchell. "Version spaces: A candidate elimination approach to rule learning". In: *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*. 1977, pp. 305–310.

[172] Serafın Moral-Garcıa et al. "Bagging of credal decision trees for imprecise classification". In: *Expert Systems with Applications* 141 (2020), p. 112944.

[173] Thomas Mortier et al. "Efficient set-valued prediction in multi-class classification". In: *Data Mining and Knowledge Discovery* 35.4 (2021), pp. 1435–1469.

[174] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. "Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option". In: *Machine Learning in Systems Biology*. PMLR. 2009, pp. 65–81.

[175] Michinori Nakata and Hiroshi Sakai. "Rule Induction Based on Rough Sets from Possibilistic Data Tables". In: *Lecture Notes in Computer Science*. Vol. 11471. Springer. 2019, pp. 86–97.

[176] Balas K Natarajan. "On learning sets and functions". In: *Mach Learn* 4.1 (1989), pp. 67–97.

[177] Brady Neal. "On the bias-variance tradeoff: Textbooks need an update". In: *arXiv preprint arXiv:1912.08286* (2019).

[178] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[179] Vu-Linh Nguyen. "Imprecision in Machine Learning Problems". PhD thesis. Université de Technologie de Compiègne, 2018.

[180] Vu-Linh Nguyen, Sébastien Destercke, and Marie-Hélène Masson. "Partial data querying through racing algorithms". In: *International Journal of Approximate Reasoning* 96 (2018), pp. 36–55.

[181] Vu-Linh Nguyen et al. "Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty". In: *27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*. 2018, pp. 5089–5095.

[182] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. "Pervasive label errors in test sets destabilize machine learning benchmarks". In: *arXiv preprint arXiv:2103.14749* (2021).

[183] Harris Papadopoulos. "Inductive conformal prediction: Theory and application to neural networks". In: *Tools in artificial intelligence*. Citeseer, 2008.

[184] Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. "Conformal prediction with neural networks". In: *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*. Vol. 2. IEEE. 2007, pp. 388–395.

[185] Zdzisław Pawlak. "Rough sets". In: *International Journal of Computer & Information Sciences* 11.5 (1982), pp. 341–356.

[186] Vladimir Pestov. "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" In: *Computers & Mathematics with Applications* 65.10 (2013), pp. 1427–1437.

[187] Dmitry Pidan and Ran El-Yaniv. "Selective prediction of financial trends with hidden markov models". In: *Advances in Neural Information Processing Systems* 24 (2011).

[188] Catarina Pires et al. "Towards Knowledge Uncertainty Estimation for Open Set Recognition". In: *Machine Learning and Knowledge Extraction* 2.4 (2020), pp. 505–532.

[189] Mario Plebani, Andrea Padoan, and Giuseppe Lippi. "Biological variation: back to basics". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 53.2 (2015), pp. 155–156.

[190] Rafael Poyiadzi et al. "The Weak Supervision Landscape". In: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2022, pp. 218–223.

[191] Arnu Pretorius, Surette Bierman, and Sarel J Steel. "A bias-variance analysis of ensemble learning for classification". In: *Annual Proceedings of the South African Statistical Association Conference*. Vol. 2016. con-1. South African Statistical Association (SASA). 2016, pp. 57–64.

[192] Mingda Qiao and Gregory Valiant. "A theory of selective prediction". In: *Conference on Learning Theory*. PMLR. 2019, pp. 2580–2594.

[193] Benjamin Quost, Thierry Denoeux, and Shoumei Li. "Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression". In: *Advances in Data Analysis and Classification* 11.4 (2017), pp. 659–690.

[194] Benjamin Quost, Marie-Hélène Masson, and Thierry Denœux. "Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules". In: *International Journal of Approximate Reasoning* 52.3 (2011), pp. 353–374.

[195] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. "Failing loudly: An empirical study of methods for detecting dataset shift". In: *Advances in Neural Information Processing Systems* 32 (2019).

[196] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in neural information processing systems* 21 (2008).

[197] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. "Machine learning in medicine". In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.

[198] Vikas C. Raykar et al. "Learning From Crowds". In: *Journal of Machine Learning Research* 11 (2010), pp. 1297–1322. ISSN: 1532-4435.

[199] Ievgen Redko et al. *Advances in domain adaptation theory*. Elsevier, 2019.

[200] Patrick Riley. "Three pitfalls to avoid in machine learning". In: *Nature* 572 (2019), pp. 27–29.

[201] Mamshad Nayeem Rizve et al. "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning". In: *arXiv preprint arXiv:2101.06329* (2021).

[202] Donald B Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[203] Mauricio Sadinle, Jing Lei, and Larry Wasserman. "Least ambiguous set-valued classifiers with bounded error levels". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 223–234.

[204] Omer Sagi and Lior Rokach. "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1249.

[205] Hiroshi Sakai, Michinori Nakata, and Dominik Ślęzak. "A prototype system for rule generation in Lipski's incomplete information databases". In: *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. Springer. 2011, pp. 175–182.

[206] Hiroshi Sakai, Mao Wu, and Michinori Nakata. "Apriori-based rule generation in incomplete information databases and non-deterministic information systems". In: *Fundamenta Informaticae* 130.3 (2014), pp. 343–376.

[207] Hiroshi Sakai et al. "A proposal of a privacy-preserving questionnaire by non-deterministic information and its analysis". In: *IEEE Big Data 2016*. IEEE. 2016, pp. 1956–1965.

[208] Hiroshi Sakai et al. "Rough sets-based machine learning over non-deterministic data: a brief survey". In: *International Conference on Advanced Machine Learning Technologies and Applications*. Springer. 2012, pp. 3–12.

[209] Luciano Sánchez et al. "Informed Weak Supervision for Battery Deterioration Level Labeling". In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer. 2022, pp. 748–760.

[210] Sverre Sandberg, Anna Carobene, and Aasne K Aarsand. "Biological variation–eight years after the 1st Strategic Conference of EFLM". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* (2022).

[211] Fadil Santosa and William W Symes. "Linear inversion of band-limited reflection seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1307–1330.

[212] Rishi Saripalle, Christopher Runyan, and Mitchell Russell. "Using HL7 FHIR to achieve interoperability in patient health record". In: *Journal of biomedical informatics* 94 (2019), p. 103188.

[213] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT Press, 2013.

[214] Lars Schmarje et al. "Beyond Cats and Dogs: Semi-supervised Classification of fuzzy labels with overclustering". In: *arXiv preprint arXiv:2012.01768* (2020).

[215] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[216] Dale Schuurmans and Russell Greiner. "Learning to classify incomplete examples". In: *Computational Learning Theory and Natural Learning Systems, Making Learning Systems Practical* 4 (1997), pp. 87–105.

[217] Tore Schweder and Nils Lid Hjort. *Confidence, likelihood, probability*. Vol. 41. Cambridge University Press, 2016.

[218] Glenn Shafer and Vladimir Vovk. “A tutorial on conformal prediction”. In: *Journal of Machine Learning Research* 9.Mar (2008), pp. 371–421.

[219] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[220] Supreeth P Shashikumar et al. “Artificial intelligence sepsis prediction algorithm learns to say “I don’t know””. In: *NPJ digital medicine* 4.1 (2021), pp. 1–9.

[221] PG Shynu, H Md Shayan, and Chiranji Lal Chowdhary. “A fuzzy based data perturbation technique for privacy preserved data mining”. In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE. 2020, pp. 1–4.

[222] Dominik Slezak and Soma Dutta. “Dynamic and Discernibility Characteristics of Different Attribute Reduction Criteria”. In: *Rough Sets - International Joint Conference, IJCRS 2018, Quy Nhon, Vietnam, August 20-24, 2018, Proceedings*. 2018, pp. 628–643.

[223] Ola Spjuth et al. “Combining prediction intervals on multi-source non-disclosed regression datasets”. In: *Conformal and Probabilistic Prediction and Applications*. PMLR. 2019, pp. 53–65.

[224] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[225] Carole H Sudre et al. “Let’s agree to disagree: Learning highly debatable multirater labelling”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 665–673.

[226] Carl-Magnus Svensson, Ron Hübler, and Marc Thilo Figge. “Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance”. In: *Journal of immunology research* 2015 (2015).

[227] Carl-Magnus Svensson et al. "Automated detection of circulating tumor cells with naive Bayesian classifiers". In: *Cytometry Part A* 85.6 (2014), pp. 501–511.

[228] Roman W Swiniarski and Andrzej Skowron. "Rough set methods in feature selection and recognition". In: *Pattern recognition letters* 24.6 (2003), pp. 833–849.

[229] Yu-Ru Syau, Churn-Jung Liau, and En-Bing Lin. "On Variable Precision Generalized Rough Sets and Incomplete Decision Tables". In: *Fundamenta Informaticae* 179.1 (2021), pp. 75–92.

[230] O Teytaud. *Bayesian learning/Structural Risk Minimization*. Tech. rep. Citeseer, 2000.

[231] K Thangavel and A Pethalakshmi. "Dimensionality reduction based on rough set theory: A review". In: *Applied Soft Computing* 9.1 (2009), pp. 1–12.

[232] Christian Thiel. "Classification on soft labels is robust against label noise". In: *Proceedings of KES-2008*. Springer. 2008, pp. 65–73.

[233] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[234] Andrey Nikolayevich Tikhonov. "On the stability of inverse problems". In: *Doklady Akademii Nauk USSR*. Vol. 39. 1943, pp. 195–198.

[235] Paolo Toccaceli. "Conformal and Venn Predictors for large, imbalanced and sparse chemoinformatics data". PhD thesis. Royal Holloway, University of London, 2021.

[236] Paolo Toccaceli and Alexander Gammerman. "Combination of conformal predictors for classification". In: *Conformal and Probabilistic Prediction and Applications*. PMLR. 2017, pp. 39–61.

[237] Paolo Toccaceli and Alexander Gammerman. "Combination of inductive mondrian conformal predictors". In: *Machine Learning* 108.3 (2019), pp. 489–510.

[238] Ilya O Tolstikhin and Yevgeny Seldin. "PAC-Bayes-empirical-Bernstein inequality". In: *Advances in Neural Information Processing Systems* 26 (2013).

[239] Isaac Triguero, Salvador Garcıa, and Francisco Herrera. "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study". In: *Knowledge and Information systems* 42.2 (2015), pp. 245–284.

[240] Alexandra N Uma et al. "Learning from disagreement: A survey". In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 1385–1470.

[241] Hafeez Ur Rehman et al. "A three-way approach for protein function classification". In: *PloS one* 12.2 (2017), e0171702.

[242] Ryan J Urbanowicz et al. "Relief-based feature selection: Introduction and review". In: *Journal of biomedical informatics* 85 (2018), pp. 189–203.

[243] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[244] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[245] Patrick Vannoorenberghe and Philippe Smets. "Partially supervised learning by a credal EM approach". In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer. 2005, pp. 956–967.

[246] Vladimir Vapnik. "Principles of risk minimization for learning theory". In: *Advances in neural information processing systems*. 1992, pp. 831–838.

[247] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[248] Vladimir Vapnik and Alexey Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Doklady Akademii Nauk USSR*. Vol. 181. 4. 1968, pp. 781–787.

[249] Michel Verleysen and Damien François. "The curse of dimensionality in data mining and time series prediction". In: *International work-conference on artificial neural networks*. Springer. 2005, pp. 758–770.

[250] Vladimir Vovk. "Cross-conformal predictors". In: *Annals of Mathematics and Artificial Intelligence* 74.1 (2015), pp. 9–28.

[251] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[252] Vladimir Vovk et al. "Criteria of efficiency for set-valued classification". In: *Annals of Mathematics and Artificial Intelligence* 81.1-2 (2017), pp. 21–46.

[253] Jerzy W. Grzymala-Busse. "Rough Set Strategies to Data with Missing Attribute Values". In: *Proceedings of ISMIS 2005*. Vol. 542. 2005, pp. 197–212.

[254] Nicolas Wagner et al. "Fuzzy k-NN Based Classifiers for Time Series with Soft Labels". In: *Proceedings of IPMU 2020*. Springer. 2020, pp. 578–589.

[255] Changzhong Wang et al. "Attribute reduction based on k-nearest neighborhood rough sets". In: *International Journal of Approximate Reasoning* 106 (2019), pp. 18–31.

[256] Peter Washington et al. "Training affective computer vision models by crowdsourcing soft-target labels". In: *Cognitive Computation* 13.5 (2021), pp. 1363–1373.

[257] Yair Wiener and Ran El-Yaniv. "Theoretical foundations of selective prediction". PhD thesis. Computer Science Department, Technion, 2013.

[258] Alan HB Wu. "Biological and analytical variation of clinical biomarker testing: implications for biomarker-guided therapy". In: *Current heart failure reports* 10.4 (2013), pp. 434–440.

[259] Jing-Han Wu and Min-Ling Zhang. "Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 416–424.

[260] Wei-Zhi Wu. "Attribute reduction based on evidence theory in incomplete decision systems". In: *Information Sciences* 178.5 (2008), pp. 1355–1371.

[261] Ji Xin et al. “The art of abstention: Selective prediction and error regularization for natural language processing”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 1040–1051.

[262] Ning Xu et al. “Progressive Purification for Instance-Dependent Partial Label Learning”. In: *arXiv preprint arXiv:2206.00830* (2022).

[263] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. “Cautious classification with nested dichotomies and imprecise probabilities”. In: *Soft Computing* 21.24 (2017), pp. 7447–7462.

[264] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. “The costs of indeterminacy: how to determine them?” In: *IEEE transactions on cybernetics* 47.12 (2016), pp. 4316–4327.

[265] Ran El-Yaniv. “On the Foundations of Noise-free Selective Classification.” In: *Journal of Machine Learning Research* 11.5 (2010).

[266] JingTao Yao and Nouman Azam. “Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets”. In: *IEEE Transactions on Fuzzy Systems* 23.1 (2014), pp. 3–15.

[267] Yiyu Yao. “An Outline of a Theory of Three-Way Decisions”. In: *Rough Sets and Current Trends in Computing - 8th International Conference, RSCTC 2012, Chengdu, China, August 17-20, 2012. Proceedings*. Ed. by Jingtao Yao et al. Vol. 7413. Lecture Notes in Computer Science. Springer, 2012, pp. 1–17.

[268] Yiyu Yao. “Three-way decision: an interpretation of rules in rough set theory”. In: *International Conference on Rough Sets and Knowledge Technology*. Springer. 2009, pp. 642–649.

[269] YY Yao. “Generalized rough set models”. In: *Rough sets in knowledge discovery* 1 (1998), pp. 286–318.

[270] Liu Yong et al. “Quick attribute reduct algorithm for neighborhood rough set model”. In: *Information Sciences* 271 (2014), pp. 65–81.

[271] Marco Zaffalon. “The naive credal classifier”. In: *Journal of statistical planning and inference* 105.1 (2002), pp. 5–21.

[272] Marco Zaffalon, Keith Wesnes, and Orlando Petrini. “Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data”. In: *Artificial intelligence in medicine* 29.1-2 (2003), pp. 61–79.

[273] Mohamed M El-Zahhar and Neamat F El-Gayar. “A semi-supervised learning approach for soft labeled data”. In: *Proceedings of ISDA 2010*. IEEE. 2010, pp. 1136–1141.

[274] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[275] Jing Zhang, Xindong Wu, and Victor S Sheng. “Learning from crowdsourced labeled data: a survey”. In: *Artificial Intelligence Review* 46.4 (2016), pp. 543–576.

[276] Min-Ling Zhang, Jing-Han Wu, and Wei-Xuan Bao. “Disambiguation Enabled Linear Discriminant Analysis for Partial Label Dimensionality Reduction”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.4 (2022), pp. 1–18.

[277] Min-Ling Zhang and Fei Yu. “Solving the partial label learning problem: An instance-based approach”. In: *Twenty-Fourth IJCAI*. 2015.

[278] Shuai Zhang et al. “How does unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis”. In: *arXiv preprint arXiv:2201.08514* (2022).

[279] Kai Zheng, Pui Cheong Fung, and Xiaofang Zhou. “K-nearest neighbor search for fuzzy objects”. In: *Proceedings of the 2010 ACM SIGMOD international conference on Management of data*. 2010, pp. 699–710.

[280] Shi Zhi. "Learning from multiple heterogeneous sources-Handling source trustworthiness and incompleteness". PhD thesis. University of Illinois at Urbana-Champaign, 2020.

[281] Xiong Zhou et al. "Learning with Noisy Labels via Sparse Regularization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 72–81.

[282] Zhi-Hua Zhou. "A brief introduction to weakly supervised learning". In: *National Science Review* 5.1 (2018), pp. 44–53.

[283] Zhi-Hua Zhou. "Bayes Classifiers". In: *Machine Learning*. Springer, 2021, pp. 155–179.

[284] William Zhu. "Topological approaches to covering rough sets". In: *Information sciences* 177.6 (2007), pp. 1499–1508.

Appendix A

Appendix: Code and Data Repositories

In this Appendix, the code and data repositories created and used within this thesis are listed. In regard to the data repositories, all the data collected and used within this thesis is available on the Zenodo archive and GitHub at the following URLs:

- `https://zenodo.org/record/5336525#.YzHB27RBy3A`
- `https://zenodo.org/record/4958146#.YzHB67RBy3A`
- `https://zenodo.org/record/4562597#.YzHCAbRBy3A`
- `https://zenodo.org/record/4081318#.YzHCdLRBy3A`
- `https://github.com/AndreaCampagner/InvasivenessAssessment`

In regard to the code repositories, all used code is available in the following GitHub repositories:

- `https://github.com/AndreaCampagner/scikit-cautious`
- `https://github.com/AndreaCampagner/scikit-weak`
- `https://github.com/AndreaCampagner/Aggregation-Models-in-Ensemble-Learning`

- https://github.com/AndreaCampagner/qualiMLpy
- https://github.com/AndreaCampagner/Development--evaluation--and-validation-of-machine-learning-models-for-COVID-19

The following paper, in particular, provides an extended discussion, as well as documentation, of the scikit-weak library, which encompasses the methods discussed in this thesis for learning from imprecise data.

scikit-weak: A Python Library for Weakly Supervised Machine Learning

Andrea Campagner[1], Julian Lienen[2], Eyke Hüllermeier[3], and Davide Ciucci[1]

[1] Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, Viale Sarca 336/14, 20126 Milano, Italy
[2] Department of Computer Science, Paderborn University,
Warburger Str. 100, 33098 Paderborn, Germany
[3] Institute of Informatics, University of Munich (LMU),
Akademiestr. 7, 80799 Munich, Germany

Abstract. In this article we introduce and describe SCIKIT-WEAK, a Python library inspired by SCIKIT-LEARN and developed to provide an easy-to-use framework for dealing with weakly supervised and imprecise data learning problems, which, despite their importance in real-world settings, cannot be easily managed by existing libraries. We provide a rationale for the development of such a library, then we discuss its design and the currently implemented methods and classes, which encompass several state-of-the-art algorithms.

Keywords: Weakly supervised learning, Imprecise data, Rough sets, Generalized risk minimization, Imprecisiation

1 Introduction

In the recent years, applications of machine learning (ML) have spread into both research and industry. Arguably, one of the major driving forces behind this growth has been the wide availability of a multitude of publicly available ML libraries, chiefly among them the Python ML eco-system [1, 9, 21, 22], centred around the SCIKIT-LEARN library[4] [23]. While such libraries offer a wide array of methods that can be applied to various ML tasks, including supervised, semi-supervised and fully unsupervised learning. By providing high-level APIs not requiring deeper knowledge, they drastically improved the accessibility.

However, not all ML tasks fit neatly into the above mentioned categories. In particular, weakly supervised learning [29] refers to machine learning tasks situated in the spectrum between supervised and unsupervised learning [24], encompassing various tasks such as multiple-instance learning [30], learning from aggregate data [8] and learning from imprecise data [15]. In this latter case, in particular, the data and annotations can be imprecise or partial: Some examples include semi-supervised learning as mentioned above, but also more general tasks

[4] https://scikit-learn.org

such as soft labels learning [10, 11, 25], in which partial labels are represented through belief functions; learning from fuzzy labels [12, 15], in which partial labels are represented through possibility distributions, and superset learning [4, 16, 20], in which partial labels are represented by exclusive sets of alternatives.

Despite the importance and practical relevance of weakly supervised learning in a variety of settings, including learning from anonymized data [26], learning from multi-rater data [8] and self-regularized learning [19], out-of-the-box libraries and frameworks to deal with such tasks are still missing and no libraries currently exist to easily manage this type of data in Python. In this article we introduce SCIKIT-WEAK, the first, to the authors' knowledge, Python library, inspired by and compatible with SCIKIT-LEARN, that provides easy-to-use methods and classes for dealing with weakly supervised learning problems. More in particular, the current version of the library focuses on the implementation of algorithms to deal with imprecise data learning problems. We provide a rationale for the development of such a library, followed by a discussion of its design and the currently implemented methods and classes, which encompass several state-of-the-art algorithms. Furthermore we briefly show the use of SCIKIT-WEAK, highlighting its interoperability with SCIKIT-LEARN, through a purposely simple but illustrative code example.

2 Background and Design Philosophy

In this section, we provide a basic background on weakly supervised learning, and specifically so to learning from imprecise data, describe the general design philosophy of SCIKIT-WEAK and illustrate an exemplary application of the library through a simple code example.

2.1 Background

In the supervised learning setting, a problem instance is defined by an instance space X and a target space Y, along with a probability distribution $\mathcal{D}$ over $X \times Y$. A finite sample of data $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, called *training set*, is assumed to be sampled from $\mathcal{D}$ and to be available for learning. In rough set terminology we can describe S by means of a *decision table*[5], that is a triple (U, Att, Y), where $U \subseteq X$ is a finite set of instances in the instance space X, Att is a set of features with each feature $f : X \to V_f$, and t is a target feature with $t : X \to Y$, where Y denotes the target space. We note that while the definition of t may suggest that the association between instances and target labels is deterministic (hence, a mapping), this is not necessarily the case as the dependency between X and Y is probabilistic and described by the unknown data generating distribution $\mathcal{D}$.

[5] Compared to the usual definition of a training set considered in the ML literature the definition of a decision table in rough set theory distinguishes instances in U from their representation in terms of features.

By contrast, in weakly supervised learning, and more specifically in learning from imprecise labels, the target feature is not assumed to be precisely known, but is instead only given in an imprecise form. In general, instead of the true target t, one can only observe the values of d, that is, a function $d : X \to D(Y)$, where $D(Y)$ is a set of *structures* over Y. As before, more in general, we may assume that instances are sampled from a distribution $\tilde{\mathcal{D}}$ defined over $X \times D(Y)$. As described in the introduction, weakly supervised learning aims at modeling learning problems in which knowledge about the supervision in a learning problem is not precisely or completely specified, but is only given in terms of imprecise beliefs or knowledge. Then, different tasks are defined based on the considered type of structures, for example:

- When $D(Y) = Y \cup \{\bot\}$, that is, each instance x is associated with either a label $y \in Y$ or no label at all ($\bot$), then the corresponding learning problem is called *semi-supervised learning*;
- When $D(Y) = 2^Y$, that is, each instance x is associated with a set of possible labels $\tilde{y} \subset Y$, then the corresponding learning problem is called *superset learning* or *partial-label learning*;
- When $D(Y) = [0,1]^Y$, that is, each instance x is associated with a possibility distribution $\pi_x : Y \to [0,1]$ over Y, then the corresponding learning problem is called *learning from fuzzy labels*;
- When $D(Y) = 2^{\mathbb{P}(Y)}$, that is, each instance x is associated with a set of probability distributions $\mathcal{Q}_x \subseteq \mathbb{P}(Y)$ over Y (that is, a *credal set*), then the corresponding leaning problem is called *credal learning*.

Thus, a weakly supervised problem instance is defined by a *weakly supervised* training set $W = \{(x_1, d_1), \ldots, (x_n, d_n)\}$ and the corresponding weakly supervised decision table $W = (U, Att, d)$, where, as above, $d : X \to D(Y)$. Given a weakly supervised decision table W, an *instantiation* of W is a standard decision table $I = (U, Att, \tilde{t})$, that is *compatible* with W (denoted $I \sim W$). For example:

- If $D(Y) = Y \cup \{\bot\}$, then $I \sim W$ iff $\forall x \in U$, $d(x) \neq \bot \implies \tilde{t}(x) = d(x)$ and $d(x) \neq \bot \implies \tilde{t}(x) \in Y$;
- If $D(Y) = 2^Y$, then $I \sim W$ iff $\forall x \in U$, $\tilde{t}(x) \in d(x)$;
- If $D(Y) = [0,1]^Y$, then $I \sim W$ iff $\forall x \in U$, $\pi_x(\tilde{t}(x)) > 0$.;
- If $D(Y) = 2^{\mathbb{P}(Y)}$, then $I \sim W$ iff $\forall x \in U$, $\exists p \in \mathcal{Q}_x$ s.t. $p(\tilde{t}) > 0$.

Notably, while we gave a binary definition of compatibility, a *graded* notion of compatibility can be defined for the learning from fuzzy labels and credal learning settings. Focusing on the first case for simplicity, for example, given two instantiations I_1, I_2 compatible with W, one could say that I_1 has stronger compatibility than I_2 when $\forall x \in U, \pi_x(\tilde{t}_1(x)) \geq \pi_x(\tilde{t}_2(x))$. See also [6] for possible alternative definitions of graded compatibility.

2.2 Design Philosophy

SCIKIT-WEAK is an open-source library, freely available via GitHub[6] and PyPi[7], that has been designed with two main aims:

- To provide a variety of easy-to-use tools and functionalities to enable data analysis grounding on weakly supervised data;
- To be inter-operable with SCIKIT-LEARN main functionalities and API.

To address the first aim, SCIKIT-WEAK is implemented through a module hierarchy that offers a variety of classes and functions to meet the main needs of a machine learning pipeline: data representation (through the `data_representation` module); pre-processing (through the `utilities` and `feature_selection` modules) and learning (through the `classification` module). Section 3 gives a comprehensive overview over each module.

To address the second aim, SCIKIT-WEAK conforms to the API of SCIKIT-LEARN. For example, classes in SCIKIT-WEAK's `feature_selection` module inherit from `sklearn.base.TransformerMixin` and thus exhibit the usual `fit`, `transform`, `fit_transform` interface. Thus, SCIKIT-WEAK classes can be used anywhere, and in the same way, a corresponding SCIKIT-LEARN class would be used, e.g., inside a `Pipeline`, enabling greater modularity and inter-operability.

Aside from SCIKIT-LEARN compatibility, to further facilitate use, SCIKIT-WEAK documentation, generated using SPHINX[8], is freely available online[9] and the library ships with an integrated suite of unit tests to ensure its correct functionality.

2.3 Code Example

To demonstrate the ease-of-use and the interoperability of SCIKIT-WEAK with SCIKIT-LEARN, consider the following example. First, starting from a standard supervised learning problem, weak supervision is generated (lines 11 – 18) by applying `DiscreteEstimatorSmoother`: this employs an underlying base classifier (in the example, a `KNeighborsClassifier`) to generate fuzzy labels. Then, a weakly supervised kNN model is instantiated (line 21; cf. Section 3.4) and a 5-fold cross validation is computed using the SCIKIT-LEARN implementation (lines 24 – 30), in order to fit and evaluate the weakly supervised model: this step, in particular, shows the interoperability between SCIKIT-WEAK and SCIKIT-LEARN base functionalities.

```
1 from scikit_weak.data_representation import
      DiscreteFuzzyLabel
2 from scikit_weak.classification import
      WeaklySupervisedKNeighborsClassifier
```

[6] https://github.com/AndreaCampagner/scikit-weak
[7] https://pypi.org/project/scikit-weak/
[8] https://sphinx-doc.org/
[9] https://scikit-weak.readthedocs.io

```
3
4  from sklearn.datasets import load_iris
5  from sklearn.neighbors import KNeighborsClassifier
6  from sklearn.model_selection import cross_val_score
7
8  import numpy as np
9
10 # Construct exemplary weak supervision
11 X, y = load_iris(return_X_y=True)
12 smooth = DiscreteEstimatorSmoother(KNeighborsClassifier(
       n_neighbors=10), type="fuzzy")
13 y_fuzzy = smooth.fit_transform(X, y)
14
15 # Instantiate weakly-supervised KNN classifier
16 clf = WeaklySupervisedKNeighborsClassifier(k=5)
17
18 # Accuracy metric
19 def accuracy(estimator, X, y_soft):
20     y_pred = estimator.predict(X)
21     y_true = np.array([np.argmax(y.to_probs()) for y in
       y_soft])
22     return np.mean(y_true == y_pred)
23
24 # Perform 5-fold cross-validation
25 cv_scores = cross_val_score(clf, X, y_soft, cv=5, scoring=
       accuracy)
```

3 Contents and Documentation

In this section, we describe the main sub-modules and classes implemented in the SCIKIT-WEAK library.

3.1 Data Representation

SCIKIT-WEAK offers a flexible set of object classes representing weak target information [13, 15], which can be found in the corresponding `data_representation` module and is depicted in Figure 1.

The basic representation is given by the abstract class `GenericWeakLabel` that defines a standard interface that should be implemented by every concrete class of weak targets, such as the ability to randomly sample an element through the `sample_value` method. SCIKIT-WEAK primarily distinguishes between continuous and discrete weak labels, which are described in the following.

Continuous Weak Labels Continuous weak labels are represented as instances of the abstract class `ContinuousWeakLabel`, whose main concrete sub-class is `IntervalLabel`. An object of this kind represents an interval-valued target specified by its lower and upper bounds l and u, e.g., as often observed in weakly

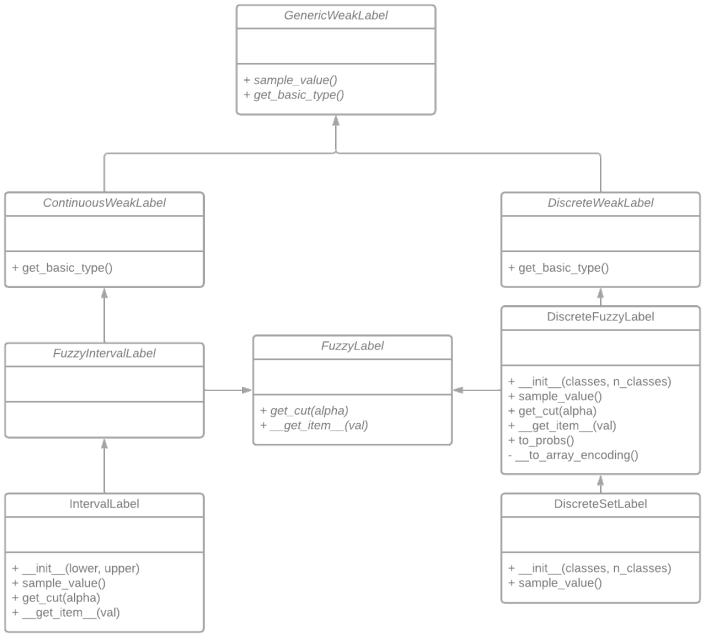


Fig. 1. The class hierarchy of data representation formats included in the module `data_representation`.

supervised regression problems. Without any further specification, each element within $[l, u]$ is considered to be equally plausible. Moreover, this class features to sample an element uniformly within this interval.

Discrete Weak Labels Discrete weak labels can be represented as instances of the abstract class `DiscreteWeakLabel`, whose main concrete sub-classes are `DiscreteFuzzyLabel` and `DiscreteSetLabel`. As discrete target representation, objects of the former class maintain possibilities $\pi_x(y) \in [0, 1]$ over elements Y, e.g., classes as typically considered in classification problems. These possibilities represent upper probabilities of the true underlying probability distribution over Y. `DiscreteFuzzyLabel` supports a sampling mechanism to draw labels according to the possibilities. Moreover, discrete fuzzy labels can represent agnostic label information, i.e., assigning full possibility $\pi_x = (1, \ldots, 1)$ to any value in Y without further distinction. Semi-supervised learning is a typical setting where such data occurs, as parts of the data are completely unlabeled and target information is agnostic. To simplify the management of the type of data that occur frequently in superset and partial-label learning, namely, that a set of elements in Y have full plausibility, while all other elements are totally implausible, SCIKIT-WEAK also implements the `DiscreteSetLabel` class.

3.2 Utilities

The `utilities` module collects general utility functions that can be used for pre-processing, optimization or data checking and analysis. In particular, the module contains the `smoothers` sub-module, that encompasses several methods to transform supervised datasets into weakly supervised datasets; as well as the `losses` sub-module, that contains some commonly used loss functions for model evaluation and optimization-based learning.

DiscreteEstimatorSmoother `DiscreteEstimatorSmoother` is a class to transform a supervised learning problem into a weakly supervised one, which uses an underlying classifier for imprecisiation. The need for this class stems from the fact that most existing benchmark datasets are precise, and hence cannot be used to test weakly supervised learning algorithms. Thus, `DiscreteEstimatorSmoother` allows to convert a standard supervised benchmark into a weakly supervised one. It supports transformation of standard labels to either `DiscreteSetLabel` or `DiscreteFuzzyLabel`. In the case of transformation to `DiscreteFuzzyLabel` objects, the underlying classifier given as input is trained on the supervised data given as input, and the output confidence scores are then normalized and used as values for the corresponding `DiscreteFuzzyLabel`. In the case of transformation to `DiscreteSetLabel` instances, only labels whose normalized confidence scores are greater than a parameterized threshold ϵ are considered as output.

DiscreteRandomSmoother Related to the previous method, instances of the class `DiscreteRandomSmoother` realize the transformation from supervised to weakly supervised problems based on random sampling. Therefore, the class supports transformation of standard labels to either `DiscreteSetLabel` or `DiscreteFuzzyLabel`. To this end, discrete random smoother offers two sampling strategies: either according to the random set model, or according to the random membership model. In the random set model, labels in the corresponding `DiscreteSetLabel` are sampled at random, according either to probability `p_incl` (for the correct label) or `p_err` (for the incorrect labels). Formally, given instance (x, y) and the corresponding set-valued label S, it holds that $P(y \in S) =$ `p_incl` and $\forall y' \neq y, P(y' \in S) =$ `p_err`. In the random membership model, possibility degrees for the labels are sampled uniformly from the set of possible possibility degrees given as input in parameter `prob_ranges`.

3.3 Feature Selection

SCIKIT-WEAK offers a selection of methods to control model complexity and data dimensionality through the `feature_selection` module, which comprises different classes to perform weakly supervised feature selection and dimensionality reduction. In particular, the current version of the library implements two rough set-based feature selection algorithms (namely, classes `RoughSetSelector` and `GeneticRoughSetSelector`) and a dimensionality reduction algorithm (`DELIN`).

RoughSetSelector `RoughSetSelector` performs weakly supervised feature selection using rough set-based reduct search [5, 7]. The class supports datasets whose weakly supervised labels are either instances of the class `DiscreteSetLabel` or `DiscreteFuzzyLabel`, and offers several choices in regard to the search strategy (brute-force or greedy search), the class of reducts to search for (superset reducts, C-reducts, λ-reducts), and the rough set model to be used (k-neighborhood or radius neighborhood rough sets). When the weakly supervised labels are instances of `DiscreteSetLabel`, both brute-force and greedy search aim to find minimal superset reducts. A superset reduct is a reduct for an instantiation of the weakly supervised dataset given as input. The brute-force search strategy examines all subsets of features $R \subseteq Att$ exhaustively to check whether they are superset reducts. The algorithm is guaranteed to return all the minimal-size superset reducts, but, however, the computational complexity is exponential ($O(|X| \cdot 2^{|Att|})$). By contrast, the greedy search strategy starts with the full set of features Att and iteratively removes one feature as long as the remaining set of feature is a superset reduct. The algorithm is not guaranteed to return a minimal-size superset reduct, but global search is supported via random restarts. The complexity of greedy search is $O(|X| \cdot |Att|^2)$. When the weakly supervised labels are instances of `DiscreteFuzzyLabel`, brute-force and greedy search aim to find either C- or λ-reducts. A C-reduct $R \subseteq Att$ is a superset reduct for an instantiation I_R for which $\nexists R' \subseteq Att$ superset reduct for an instantiation $I_{R'}$ such that both $|R'| \leq |R|$ and $\min_{x \in S} \pi_x(\tilde{t}_{I_R}(x)) \leq \min_{x \in S} \pi_x(\tilde{t}_{I_{R'}}(x))$. A λ-reduct $R \subseteq Att$ is a superset reduct for an instantiation I_R that minimizes $(1-\lambda)(\min_{x \in S} \pi_x(\tilde{t}_{I_R}(x))) - \lambda \frac{|R|}{|Att|}$ among all superset reducts. Both brute-force and greedy search perform feature selection by searching for superset reducts on the α-cuts of the fuzzy-labeled dataset given as input, and then selecting among the retrieved reducts those that satisfy the constraints of being either a C-reduct or a λ-reduct. Thus, the complexity of brute-force search is $O(|X|^2 \cdot 2^{|Att|})$ while the complexity of greedy search is $O(|X|^2 \cdot |Att|^2)$.

GeneticRoughSetSelector The class `GeneticRoughSetSelector` offers functionality to perform weakly supervised selection by reduct search using genetic algorithms [6]. The class supports datasets whose weakly supervised labels are instances of `DiscreteFuzzyLabel`. `GeneticRoughSetSelector` aims to find either C-reducts, D-reducts or λ-reducts for the weakly supervised dataset given as input, supporting every type of weakly supervised label. A D-reduct $R \subseteq Att$ is a superset reduct for an instantiation I_R for which $\nexists R' \subseteq Att$ superset reduct for an instantiation $I_{R'}$ s.t. both $|R'| \leq |R|$ and $\exists x \in S, \pi_x(\tilde{t}_{I_R}(x)) < \pi_x(\tilde{t}_{I_{R'}}(x))$. The genetic algorithm-based search is guided by one of three possible fitness functions, corresponding to the above mentioned reduct classes:

$$Fitness_C = \langle r, p \rangle, \tag{1}$$

$$Fitness_\lambda = (1-\lambda)p - \lambda \frac{r}{|Att|}, \tag{2}$$

$$Fitness_D = \langle r, s \rangle, \tag{3}$$

where $p = \min_{x \in S} \pi_x(\tilde{t}_I(x))$, $r = \begin{cases} |A| & F \text{ is a super-reduct} \\ \infty & \text{otherwise} \end{cases}$, and $s \in [0,1]^{|U|}$ is a vector s.t. $s_x = \pi_x(\tilde{t}_I(x))$. Note, in particular, that only $Fitness_\lambda$ is single-valued, while the other two fitness functions are multi-valued. Consequently, for these latter two fitness functions, the implementation employs a multi-objective optimization algorithm. Irrespective of the fitness function adopted, the computational complexity of `GeneticRoughSetSelector` is $O(|X| \cdot |Att|)$. With regard to selection and cross-over, `GeneticRoughSetSelector` employs non-dominated tournament selection and single-point cross-over, respectively. For mutation, candidate reducts are mutated by random addition or deletion of features according to a Bernoulli distribution. By contrast, instantiations are mutated according to a two-step procedure. First, for each instance x, a binary value is randomly sampled from a Bernoulli distribution, then, if the above mentioned value was equal to 1, a new target label is sampled using the method `sample_value` of the corresponding `GenericWeakLabel` instance.

DELIN `DELIN` is a weakly supervised dimensionality reduction algorithm, based on the combination of linear discriminant analysis and weakly supervised k-NN [2, 27, 28]. The class supports datasets whose weakly supervised labels are instances either of the class `DiscreteSetLabel` or `DiscreteFuzzyLabel`. `DELIN` requires one to determine a-priori the number of dimensions to be selected via the parameter `n`. Intuitively, the algorithm works in iterations, each of which consists of two steps: first, `WeaklySupervisedKNeighborsClassifier` is applied to the data, then linear discriminant analysis is applied to the original data w.r.t. the confidence scores given as output of the first step. Compared to the algorithm originally proposed in [27, 28], the `DELIN` class has two main modifications: first, it supports not only `DiscreteSetLabel` but also `DiscreteFuzzyLabel` instances; second, singular value decomposition is used in the computation of linear discriminant analysis to avoid stability issues. The computational complexity of `DELIN` is $O(|X| \cdot |Att|^2)$.

3.4 Classification

Aside from the pre-processing and dimensionality reduction methods described in the previous sections, SCIKIT-WEAK also offers a wide selection of weakly supervised classification algorithms contained in the `classification` module.

WeaklySupervisedKNeighborsClassifier As one of two neighborhood-based methods, `WeaklySupervisedKNeighborsClassifier` is a simple generalization of k-nearest neighbors classification to the setting of weakly supervised data [3, 17], and is compatible with every instance of `DiscreteWeakLabel`. The number of neighbors can be controlled through parameter `k`, while the class supports any metric callable (through the `metric` parameter, default is the Euclidean metric). For efficiency reasons, SCIKIT-LEARN's `NearestNeighbors` is used to speed-up

neighbors search: the computational complexity is $\Omega(|X| \cdot log|X|)$, with an additional complexity of $\Omega(log|X|), O(|X|)$ at inference time.

WeaklySupervisedRadiusClassifier `WeaklySupervisedRadiusClassifier` is yet another simple generalization of radius-based neighbors classification to weakly supervised data [17], and is compatible with every instance of `DiscreteWeakLabel`. The radius within which to search for neighbor instances can be controlled through the `radius` parameter, while the class supports any metric callable (through the `metric` parameter, default is the Euclidean metric). Similarly as for class `WeaklySupervisedKNeighborsClassifier`, `NearestNeighbors` is used to speed-up neighbors search: the computational complexity is $\Omega(|X| \cdot log|X|)$, with an additional complexity of $\Omega(log|X|), O(|X|)$ at inference time.

GRMLinearClassifier `GRMLinearClassifier` is an optimization-based classification method that attempts to directly minimize the generalized risk for a linear model [15]. Currently, it supports instances of `DiscreteFuzzyLabel` and implements two different linear classification algorithms, namely, logistic regression (by setting `loss` parameter to "*logistic*") or linear SVM (by setting `loss` parameter to "*hinge*"). More in detail, given loss function l, `GRMLinearClassifier` attempts to solve the following optimization problem:

$$\underset{W}{\operatorname{argmin}} \frac{1}{|X|} \sum_{(x,\pi) \in S} l_F(\pi, W \cdot x)$$

where $l_F : [0,1]^Y \times \mathbb{R}^Y \to \mathbb{R}$ is the generalized risk [15], defined as

$$l_F(\pi, W \cdot x) = \int_0^1 \min_{y \in \pi^\alpha} l(y, W \cdot x) d\alpha \,. \tag{4}$$

Optimization is implemented by means of gradient descent, relying on TensorFlow[10] for efficient computation. In particular, the class supports every TensorFlow optimizer (through the `optimizer` parameter, default is stochastic gradient descent "*sgd*"). In general, the optimization problem described above is non-convex, thus convergence to a global optimum is not guaranteed and no convergence checking is implemented. Training is performed for a fixed number of iterations (set through parameter `max_epochs`), therefore complexity is on the order of $O(|X| \cdot |Att|)$. To avoid overfitting, `GRMLinearClassifier` supports weight regularization, set through the `regularizer` parameter.

RRLClassifier `RRLClassifier` is an efficient ensemble-based method for weakly supervised classification based on a generalization of tree ensemble-based learning [8]. `RRLClassifier` trains an ensemble of standard supervised classifiers (by default, scikit-learn's `ExtraTreeClassifier` [14], but the type of classifier

[10] https://tensorflow.org

can be set through parameter `estimator`) by drawing random samples from the weakly supervised data given as input. For each instance label Y in the training set, and each classifier h_i to be ensembled, a sample label is obtained by calling `y.sample_value()`. Thus, `RRLClassifier` supports every instance of `GenericWeakLabel`. Optionally, bootstrapping (as in random forests) can be applied (through parameter `resample`, by default set to `False`) to ensure increased diversity among the classifiers in the ensemble. The computational complexity of `RRLClassifier` is $O(k \cdot |Att||X| \cdot log|X|)$, where k is the number of classifiers to be ensembled (set through parameter `n_estimators`).

LabelRelaxationNNClassifier As one example of a credal learning classifier, `LabelRelaxationNNClassifier` provides an implementation of the label relaxation loss [19] to train probabilistic neural network classifiers $H : X \to \mathbb{P}(Y)$ with $\mathbb{P}(Y)$ denoting the space of probability distributions over Y. Commonly, training of such models H involves a gradient-descent based optimization of a probabilistic loss $l : \mathbb{P}(Y) \times \mathbb{P}(Y) \to \mathbb{R}_+$, where degenerate probability distributions p_y with $p_y(y) = 1$ and $p_y(\cdot) = 0$ otherwise are considered as surrogate targets for an observed class labels $y \in Y$, typically resulting into overconfident models. To achieve better calibrated models by a more faithful target modeling, label relaxation replaces the degenerate distribution p_y assigned to an instance x by a credal set $\mathcal{Q}_{\pi_x}$ in accordance with a possibility distribution π_x that assigns a fixed possibility $\pi_x(y') = \alpha \in [0, 1]$ to the labels $y' \neq y$ and $\pi_x(y) = 1$. This credal set $\mathcal{Q}_{\pi_x}$ is then used as target within a generalized loss formulation adopting Eq. (4) to train models, which is implemented in the class `LabelRelaxationLoss`. `LabelRelaxationNNClassifier` allows one to specify the imprecisiation parameter α (parameter `lr_alpha`), as well as hyperparameters related to stochastic gradient descent (SGD) optimization. Moreover, the base network to be trained can be specified by its hidden layer depth and widths. The computational complexity depends on the parameterization of the SGD procedure, resulting in a complexity similar to `GRMLinearClassifier`. As before, we use TensorFlow as optimization framework.

CSSLClassifier Another credal learning method is provided in the class `CSSLClassifier`, which implements so-called credal self-supervised learning (CSSL) [18] to induce probabilistic classifiers in a semi-supervised learning scenario. To this end, CSSL maintains credal sets $\mathcal{Q}_{\pi_x}$ as used in `LabelRelaxationLoss` expressing the model's belief about the true target for previously unlabeled instances, proceeding from agnostic credal sets of the form $\mathcal{Q}_{\pi_x} = \mathbb{P}(Y)$ with $\pi_x(y') = 1 \, \forall y' \in Y$. These credal sets successively shrink with increased training progress and thus higher model confidence by reducing the degree of imprecisiation in π_x. `CSSLClassifier` allows one to specify a base model (parameter `estimator`, e.g., an instance of `LabelRelaxationNNClassifier`), the number of iterations (`n_iterations`), a class prior distribution used within the credal set construction (`p_data`) and the buffer size of the model prediction history also employed in the credal set construction (`p_hist_buffer_size`). In each iteration,

the base model is retrained on the complete data and the credal sets are adjusted according to the updated model.

4 Conclusion

In this article, we introduced SCIKIT-WEAK, a Python library for weakly supervised learning and data analysis, currently focusing on the handling of learning from imprecise data problems. To the authors knowledge, SCIKIT-WEAK is the first library providing such functionality in Python, and thus we believe it could advance the applicability of the Python data science ecosystem to non-standard and weakly supervised learning problems. We described the fundamental design concepts underlying the library and documented the main implemented functionalities and classes. We also illustrated the use of the library by means of a simple example. The SCIKIT-WEAK is an open source project and we hope that additional contributors can help maintain the library as well as implement new functionalities: indeed, being freely and openly available on GitHub, and being implemented completely in Python, we believe developers could easily extend and add new functionalities to the existing library. In particular, we envision the following next steps for the development of the library:

- To extend the suite of implemented weakly supervised data representation, so as to encompass additional and more general learning settings such as those mentioned in the introduction;
- To provide more efficient and robust implementations of the currently implemented classes, e.g., by off-loading time-sensitive routines to low-level or device code, or by implementing more extensive type checking and tests;
- To enrich the library with sample weakly supervised datasets that can be used for prototyping, testing as well as benchmarking purposes.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Bao, W.X., Hang, J.Y., Zhang, M.L.: Partial label dimensionality reduction via confidence-based dependence maximization. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 46–54 (2021)
3. Bezdek, J.C., Chuah, S.K., Leep, D.: Generalized k-nearest neighbor rules. Fuzzy Sets and Systems **18**(3), 237–256 (1986)
4. Cabannes, V., Bach, F., Rudi, A.: Disambiguation of weak supervision with exponential convergence rates. arXiv preprint arXiv:2102.02789 (2021)

5. Campagner, A., Ciucci, D.: Feature selection and disambiguation in learning from fuzzy labels using rough sets. In: International Joint Conference on Rough Sets. pp. 164–179. Springer (2021)
6. Campagner, A., Ciucci, D.: Rough-set based genetic algorithms for weakly supervised feature selection. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 761–773. Springer (2022)
7. Campagner, A., Ciucci, D., Hüllermeier, E.: Rough set-based feature selection for weakly labeled data. International Journal of Approximate Reasoning **136**, 150–167 (2021)
8. Campagner, A., Ciucci, D., Svensson, C.M., Figge, M.T., Cabitza, F.: Ground truthing from multi-rater labeling with three-way decision and possibility theory. Information Sciences **545**, 771–790 (2021)
9. Chollet, F., et al.: Keras. https://keras.io (2015)
10. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern recognition **42**(3), 334–348 (2009)
11. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Transactions on Knowledge and Data Engineering **25**(1), 119–130 (2011)
12. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy sets and systems **122**(3), 409–424 (2001)
13. Destercke, S.: Uncertain data in learning: challenges and opportunities. Conformal and Probabilistic Prediction with Applications pp. 322–332 (2022)
14. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)
15. Hüllermeier, E.: Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. International Journal of Approximate Reasoning **55**(7), 1519–1534 (2014)
16. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intelligent Data Analysis **10**(5), 419–439 (2006)
17. Kuncheva, L.: Fuzzy classifier design, vol. 49. Springer Science & Business Media (2000)
18. Lienen, J., Hüllermeier, E.: Credal self-supervised learning. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 14370–14382 (2021)
19. Lienen, J., Hüllermeier, E.: From label smoothing to label relaxation. In: Proc. of the 35th AAAI Conference on Artificial Intelligence, virtual, February 2-9 (2021)
20. Liu, L., Dietterich, T.G.: A conditional multinomial mixture model for superset label learning. In: Advances in neural information processing systems. pp. 548–556 (2012)
21. Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., Király, F.J.: sktime: A unified interface for machine learning with time series. arXiv preprint arXiv:1909.07872 (2019)
22. McKinney, W., et al.: pandas: a foundational python library for data analysis and statistics. Python for high performance and scientific computing **14**(9), 1–9 (2011)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)

24. Poyiadzi, R., Bacaicoa-Barber, D., Cid-Sueiro, J., Perello-Nieto, M., Flach, P., Santos-Rodriguez, R.: The weak supervision landscape. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). pp. 218–223. IEEE (2022)
25. Quost, B., Denoeux, T., Li, S.: Parametric classification with soft labels using the evidential em algorithm: linear discriminant analysis versus logistic regression. Advances in Data Analysis and Classification **11**(4), 659–690 (2017)
26. Sakai, H., Liu, C., Nakata, M., Tsumoto, S.: A proposal of a privacy-preserving questionnaire by non-deterministic information and its analysis. In: 2016 IEEE International Conference on Big Data (Big Data). pp. 1956–1965. IEEE (2016)
27. Wu, J.H., Zhang, M.L.: Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 416–424 (2019)
28. Zhang, M.L., Wu, J.H., Bao, W.X.: Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. ACM Transactions on Knowledge Discovery from Data (TKDD) **16**(4), 1–18 (2022)
29. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Science Review **5**(1), 44–53 (2018)
30. Zhou, Z.H., Sun, Y.Y., Li, Y.F.: Multi-instance learning by treating instances as non-iid samples. In: Proceedings of the 26th annual international conference on machine learning. pp. 1249–1256 (2009)