Learning Sets of Probabilities Through Ensemble Methods

Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke

Heudiasyc Laboratory, Université de Technologie de Compiègne, France
{vu-linh.nguyen,haifei.zhang,sebastien.destercke}@hds.utc.fr

Abstract. A possible approach to obtain set-valued predictions is to learn for each query instance a probability set (a.k.a. credal set) representing its associated uncertainty. Theoretically founded decision rules extending classical expectation and inducing a partial order between predictions can the be used to derive set-valued predictions. However, obtaining such a credal set by imprecisiating a given learning algorithm is usually computationally challenging, except for simple models such as decision trees or naive Bayes classifiers. In this paper, we propose a simple, easy to use quantile-based framework for estimating credal sets using output of ensemble methods, that can also cope with complex types of data, such as images and mixed/multimodal data, etc. Experiments are conducted to highlight the usefulness of the proposed framework.

Keywords: Ensemble learning · Credal sets approximation· Set-valued prediction · Quantile-based approach.

1 Introduction

Classification algorithms are usually designed to produce, for each instance, a prediction in the form of a unique element of the set of possible outputs. Under the presence of uncertainty, which is often a consequence of model inadequacy and/or data imperfections (in terms of quality and/or quantity), the model can however be uncertain about its predictions and make unreliable precise predictions. In such a case, it might be more desirable to provide imprecise (or indeterminate) set-valued predictions which aims to balance correctness (the true output is an element of the set-valued prediction) and precision (the cardinality of the set-valued prediction) in some appropriate manner [11,24,32,38].

Learning with a reject option is the simplest case of learning set-valued predictions, in which the classifier is allowed to either produce a singleton prediction or refuse to make a prediction for a given query instance. Threshold-based classifiers have been proposed for that purpose, in which a (global/local) threshold will be employed to decide whether a query instance should be rejected or predicted and then a conventional classifier is called only if the instance should be classified [2,5,7,14,16,17]. Threshold-based classifiers have been developed for multi-class classification (MCC) [11,24], when the classifier is allowed to return top (locally/globally) ranked classes. While such classifiers are intuitive and easy to implement, they often require reliable estimates of the class probabilities to be performant, which is hard to ensure when information is lacking.

By considering more expressive uncertainty representations, imprecise probabilistic classifiers [6,8,22,37] can provide, at least in theory, more reliable outputs. They are developed based on the assumption that uncertainty is described by a (not necessarily convex) set of probabilities, i.e., a *credal set* [21], a description to which can then be applied theoretically justified decision rules [19,32] to produce set-valued predictions. Moving from a single distribution to a *credal set* is a natural way to model the lack of information, an aspect that unique probabilities can hardly capture. Unfortunately, imprecise probabilistic classifiers often suffer from the limited use to certain types of (tabular) data, as well as from the high computational cost that represent a credal extension of a given learning method. A solution might be to consider the credal set as a neighbourhood of the initial estimated distribution [23,29], yet ensuring the quality of the initial estimated distribution is a challenge itself.

In this paper, we propose a quantile-based framework for estimating credal sets from the output of ensembles [12]. We specifically seek a correctness-precision trade-off when constructing estimates of credal sets, i.e., the estimates are expected to be informative and at the same time not very large. This shall be done by defining "median" of set of distributions and use the "median" to filter out a proportion of "extreme" distributions before forming credal sets. Moreover, we only require the availability of an ensemble of probabilistic classifiers. Thus, the base learner (ensemble) can be freely chosen according to our needs. This flexibility of the proposed approach is remarkably different from existing imprecise probabilistic classifiers. Therefore, we hope to broaden the use of generalized decision rules [19,32] to applications with complex types of data, such as mixed data [10], image/video [35,36] and multimodal data [26,33].

We provide in Section 2 a minimal description of MCC with sets of probabilities. Our main contribution which is a quantile-based approach for estimating credal sets is presented in Section 3. The inference problem with sets of probabilities is summarized in Section 4. Section 5 presents some preliminary experiments on tabular data sets to motivate the use of the proposed framework. Section 6 concludes this work and sketches out future work.

2 Preliminary

We shall recall basics of classification with sets of probabilities and notations.

2.1 Probabilistic Classification

Let $\mathcal{X}$ denote an instance space, and let $\mathcal{Y} = \{y^1, \dots, y^K\}$ be a finite set of classes. We assume that an instance $\boldsymbol{x} \in \mathcal{X}$ is (probabilistically) associated with members of $\mathcal{Y}$. We denote by $\boldsymbol{p}(Y \mid \boldsymbol{x})$ the conditional distribution of Y given $\mathbf{X} = \boldsymbol{x}$. Given training data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \dots, N\}$ drawn independently from $\boldsymbol{p}(\mathbf{X}, Y)$, the goal in MCC is to learn a classifier $\boldsymbol{h}$, which is a mapping $\mathcal{X} \longrightarrow \mathcal{Y}$ that assigns to each instance $\boldsymbol{x} \in \mathcal{X}$ a class $\hat{y} := \boldsymbol{h}(\boldsymbol{x}) \in \mathcal{Y}$.

To evaluate the performance of a classifier $\boldsymbol{h}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is needed, which compares a prediction $\hat{y}$ with a ground-truth y. Each classifier $\boldsymbol{h}$ is evaluated using its expected loss

$$R(\boldsymbol{h}) := \mathbf{E}\big[\ell(Y, \boldsymbol{h}(\mathbf{X}))\big] = \int \ell(y, \boldsymbol{h}(\boldsymbol{x}))\, d\,\mathbf{P}(\boldsymbol{x}, y)\,,$$

where $\mathbf{P}$ is the joint probability measure on $\mathcal{X} \times \mathcal{Y}$ characterizing the underlying data-generating process. Therefore, the Bayes-optimal classifier is given by

$$\boldsymbol{h}^* := \operatorname*{argmin}_{\boldsymbol{h} \in \mathcal{H}} R(\boldsymbol{h})\,, \tag{1}$$

where $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the hypothesis space. When $\mathcal{H}$ is probabilistic, we can follow maximum likelihood estimation and define the Bayes-optimal classifier as:

$$\hat{\boldsymbol{h}} := \hat{\boldsymbol{p}} := \operatorname*{argmax}_{\boldsymbol{p} \in \mathcal{H}} \mathrm{CLL}(\boldsymbol{p} \,|\, \mathcal{D}) := \operatorname*{argmax}_{\boldsymbol{p} \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \log \boldsymbol{p}(y_n \,|\, \boldsymbol{x}_n)\,. \tag{2}$$

To avoid overfitting, the CLL is often augmented by a regularization term.

Once the classifier (2) is learned from $\mathcal{D}$, we can in principle find an optimal prediction of any loss function ℓ at the prediction time [13,24]. More precisely, assume the classifier (2) is made available, and predicts for each query instance $\boldsymbol{x}$ a probability distribution $\boldsymbol{p}(\cdot \,|\, \boldsymbol{x})$ on the set of labelings $\mathcal{Y}$. The Bayes-optimal prediction (BOP) of any ℓ is then given by the expected loss minimizer

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}(\boldsymbol{x}) \in \operatorname*{argmin}_{\bar{y} \in \mathcal{Y}} \mathbf{E}\big(\ell(\boldsymbol{y}, \bar{\boldsymbol{y}})\big) = \operatorname*{argmin}_{\bar{\boldsymbol{y}} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \ell(y, \bar{y})\, \boldsymbol{p}(y \,|\, \boldsymbol{x})\,. \tag{3}$$

2.2 Classification with Set of Probabilities

Under this setting, we assume that our uncertainty is described by a (not necessarily convex) set of probabilities $\mathcal{P}(\mathcal{Y} \,|\, \boldsymbol{x})$, i.e., a *credal set* [21]. Clearly, the decision rule (3) is no longer directly applicable. Therefore, it is necessary to use some generalized decision rule such as the ones benefiting from strong theoretical justifications [19,32].

Credal sets can arise in different ways, either as a native result of the learning method [1], as the result of an agnostic (with respect to the missingness process) estimation in presence of imprecise data, or as a neighbourhood taken over an initial estimated distribution $\boldsymbol{p}(Y \,|\, \boldsymbol{x})$ [23,29]. These approaches seems to introduce some inconvenience. Native credal classifiers can be hard to learn, and are unavailable for complex inputs such as such as mixed data and images. Approximating $\mathcal{P}(\mathcal{Y} \,|\, \boldsymbol{x})$ as a neighbourhood taken over an initial estimated distribution $\boldsymbol{p}(Y \,|\, \boldsymbol{x})$ does not face this inconvenience, but requires that the initial estimated distribution is well-estimated, a hard to ensure quality.

In the next section, we propose a simple, flexible and easy to use quantile-based framework for estimating credal sets using output of ensemble methods [12]. This is especially designed to make use of the current and future development of both probabilistic classification and generalized decision rules in a unified framework to broaden the application of imprecise probability (IP) to real-world applications with complex data types.

3 Credal Sets Approximation

We assume an ensemble $\mathbf{H} := \{\boldsymbol{h}^m \,|\, m \in [M] := \{1, \ldots, M\}\}$ of M probabilistic classifiers $\boldsymbol{h}^m$, $m \in [M]$ is made available and provides, for each instance $\boldsymbol{x}$, a set of M probabilistic predictions

$$\mathbf{H}(\boldsymbol{x}) := \{\boldsymbol{h}^m(\boldsymbol{x}) \,|\, m \in [M]\} = \{\boldsymbol{p}^m := (p_1^m, p_2^m, \ldots, p_K^m) \;|\, m \in [M]\} \,. \tag{4}$$

Our goal is to aggregate this set of probabilistic predictions into a credal set $\mathcal{P}(\mathcal{Y} \,|\, \boldsymbol{x})$ in some meaningful way.

3.1 A Quantile-Based Approach

The intention of this approach is to seek a correctness-precision trade-off, i.e., the estimations of $\mathcal{P}(\mathcal{Y} \,|\, \boldsymbol{x})$ are expected to be informative and at the same time not very large. We define the reference point of $\mathbf{H}(\boldsymbol{x})$ as follows:

$$\boldsymbol{p}^* = \operatorname*{argmin}_{\boldsymbol{p}:\sum_{k=1}^{K} p_k = 1} \sum_{m=1}^{M} d(\boldsymbol{p}, \boldsymbol{p}^m) \,. \tag{5}$$

where d is some distance defined for pairs of probability distributions.

Once the reference point $\boldsymbol{p}^*$ is made available, it allows us to define a preference order, reflecting how common/weird each distribution in $\mathbf{H}(\boldsymbol{x})$ is:

$$\boldsymbol{p} \succ \boldsymbol{p}' \text{ if } d(\boldsymbol{p}^*, \boldsymbol{p}) < d(\boldsymbol{p}^*, \boldsymbol{p}') \,. \tag{6}$$

Such a preference order in turn allows us to “discard” a given percentage of outliers among elements of $\mathbf{H}(\boldsymbol{x})$.

Let $\alpha \in [0, 1]$ be some threshold. We define $\mathbf{H}_\alpha(\boldsymbol{x})$ as the set of $(1 - \alpha) * 100$ % of closest distributions in $\mathbf{H}(\boldsymbol{x})$ with respect to the preference order (6). We approximate the credal set $\boldsymbol{p}(\mathcal{Y} \,|\, \boldsymbol{x})$ of $\boldsymbol{x}$ by the convex hull of $\mathbf{H}_\alpha(\boldsymbol{x})$. Let $\mathbf{H}_\alpha(\boldsymbol{x}) := \{\boldsymbol{p}^m \,|\, m \in [M_\alpha]\}$. The convex hull is defined as

$$\mathbf{CH}_\alpha(\boldsymbol{x}) := \left\{ \boldsymbol{p} := \sum_{m=1}^{M_\alpha} \alpha_m \, \boldsymbol{p}^m \,|\, \alpha := (\alpha_1, \ldots, \alpha_{M_\alpha}) \geq 0, \sum_{m=1}^{M_\alpha} \alpha_m = 1 \right\} \,. \tag{7}$$

The computational complexity of the problem of determining the reference point (5) can greatly depend on the nature of the distance d. In the next section, we recall commonly used distances. Due to page length limit, we only mention few convex distances and refer to [4,15,20,31] for more distances.

3.2 The Cases of Convex Distances

For completeness, we shall start with few definitions and remarks, which are quite basic and would have appeared in textbooks and papers (see, e.g., [3,9,28]).

Definition 1. *A function $f : \mathbb{R}^K \longmapsto \mathbb{R}$ is convex if for every $\boldsymbol{p}, \boldsymbol{p}' \in \mathbb{R}^K$ and every $\lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$, we have the inequality*

$$f(\lambda_1 \boldsymbol{p} + \lambda_2 \boldsymbol{p}') \leq \lambda_1 f(\boldsymbol{p}) + \lambda_2 f(\boldsymbol{p}') \,. \tag{8}$$

Remark 1. Let $\boldsymbol{z} \in \mathbb{R}^K$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^K$. $f(\boldsymbol{p}) := \|\boldsymbol{p} - \boldsymbol{z}\|$ is convex.

Proof. (Sketch) The convexity of $f(\boldsymbol{p})$ follows consequently from the triangle inequality of norms. □

Remark 2. Conical combinations of convex functions are also convex.

Proof. The proof is trivial. It is enough to multiply the inequalities, one per convex function, by non-negative scalars and sum them up. □

In the following, we show that if $f^m(\boldsymbol{p}) := d(\boldsymbol{p}, \boldsymbol{p}^m)$ is convex, $m \in [M]$, then the problem of finding a reference point (5) of $\mathbf{H}(\boldsymbol{x})$ can be straightforwardly formulated as a convex optimization problem. This is indeed computationally advantageous because with recent advances, convex programming is nearly as straightforward as linear programming [3,30].

Definition 2. *A standard convex optimization problem is of the form*

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad f(\boldsymbol{p}) \qquad \text{subject to} \quad g_i(\boldsymbol{p}) \leq 0 \,, i \in [I] \,,\, h_j(\boldsymbol{p}) = 0 \,, j \in [J] \tag{9}$$

where: $\boldsymbol{p} \in \mathbb{R}^K$ is the optimization variable; The objective function $f : \mathbb{R}^K \longmapsto \mathbb{R}$ is convex; The inequality constraint functions $g_i : \mathbb{R}^K \longmapsto \mathbb{R}$, $i \in [I]$ are convex; The equality constraint functions $h_j : \mathbb{R}^K \longmapsto \mathbb{R}$, $j \in [J]$, are of the form: $h_i(\boldsymbol{p}) = \boldsymbol{a}_j \boldsymbol{p} - b_j$, where $\boldsymbol{a}_j$ is a vector and b_j is a scalar.

We can encode the condition that the reference point must be a valid probability distribution by using K inequality constraint functions g_i and 1 equality constraint function h_1:

$$g_k(\boldsymbol{p}) := -p_k \leq 0 \,, k \in [K] \,,\, h_1(\boldsymbol{p}) := \mathbf{1}_K \boldsymbol{p} - 1 = 0 \,, \tag{10}$$

where $\mathbf{1}_K = (1, \ldots, 1)$. The constraints $p_k \leq 1$, $k \in [K]$, are implicitly enforced by the K constraints g_k (i.e., $p_k \geq 0$, $k \in [K]$) and h_1 (i.e., $\sum_{k=1}^{K} p_k = 1$, $k \in [K]$). Therefore, we can use any existing package to find $\boldsymbol{p}^*$ (5).

Using Remark 1–2, we can verify that different distances (See [4,15,20,31] and elsewhere) are convex. Examples are members of the L_p Minkowski family

$$f_p(\boldsymbol{p}) := L_p(\boldsymbol{p}, \boldsymbol{z}) := \sqrt[p]{\sum_{k=1}^{K} |p_k - z_k|^p} \,, p \geq 1 \,, \tag{11}$$

and Chebyshev distance

$$f_{\text{cheb}}(\boldsymbol{p}) := L_\infty(\boldsymbol{p}, \boldsymbol{z}) := \max_{k\in[K]} |p_k - z_k| \,. \tag{12}$$

Moreover, a closer look at Definition 1 is enough to verify the convexity of some other distances (discussed in [4,15,20,31] and elsewhere). Examples are the Squared Euclidean distance (whose square function allows triangle inequality)

$$f_{\text{sqe}}(\boldsymbol{p}) := d^{\text{sqe}}(\boldsymbol{p}, \boldsymbol{z}) := \sum_{k=1}^{K} (p_k - z_k)^2 \,, \tag{13}$$

and KL divergence (inequality (8) can be verified using the log sum inequality):

$$f_{\text{KL}}(\boldsymbol{p}) := d_{\text{KL}}(\boldsymbol{p}, \boldsymbol{z}) := \sum_{k=1}^{K} p_k \log\left(p_k / z_k\right) \,, \,. \tag{14}$$

To solve the problem (9) efficiently, one should carefully look at the nature of the given convex distance. For example, for any given K, closed-form solution for the f_{Sqe} (13) can be derived (See Proposition 1). This is also a special case where the additional constraints (i.e., $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$, $k \in [K]$) do not change the minimizer. However, it is not always the case. For example, these additional constraints can change the minimizer of f_1 (11) (See Proposition 2). Also, different distances may seek for the same minimizer. Examples of such distances are Topsør and Jensen-Shannon [4]. Moreover, for some distance, such as Inner Product [4], the problem (9) is reduced to a linear program.

Proposition 1. *The reference point* $\boldsymbol{p}^*$ (5) *under Squared Euclidean distance* f_{sqe} (13) *is uniquely defined as*

$$p_k^* = \frac{1}{M} \sum_{m=1}^{M} p_k^m \,, k \in [K] \,. \tag{15}$$

Proof. (Sketch) Since $f_{\text{sqe}}(\boldsymbol{p})$ (13) is strictly convex, its unique minimizer is attained when the partial derivatives are zeros, i.e., $\boldsymbol{p}^*$ is defined in (15). $\boldsymbol{p}^*$ is a valid distribution because the set of possible distributions is a convex set. □

Proposition 2. *Except for* $K = 2$*, the reference point* $\boldsymbol{p}^*$ (5) *under* f_1 (11) *may not be the minimizer of the relaxed optimization problem*

$$\bar{\boldsymbol{p}} \in \operatorname*{argmin}_{\boldsymbol{p}} \sum_{m=1}^{M} L_1(\boldsymbol{p}, \boldsymbol{p}^m) = \operatorname*{argmin}_{\boldsymbol{p}} \sum_{k=1}^{K} \left(\sum_{m=1}^{M} |p_k - p_k^m| \right) \,. \tag{16}$$

Proof. Without enforcing the probability axioms (i.e., $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$, $k \in [K]$), a minimizer $\bar{\boldsymbol{p}}$ of the relaxed optimization problem (16) is defined as

$$\bar{p}_k := \text{median}(p_k^1, \ldots, p_k^M) \,, k \in [K] \,. \tag{17}$$

This can be verified by showing that, for any $\boldsymbol{p} \neq \bar{\boldsymbol{p}}$, we have

$$f_1(p_k) := \sum_{m=1}^{M} |p_k - p_k^m| \geq \sum_{m=1}^{M} |\bar{p}_k - p_k^m| := f_1(p'_k)\,, k \in [K]\,, \tag{18}$$

which implies the relation $f_1(\boldsymbol{p}) \geq f_1(\bar{\boldsymbol{p}})$.

Let L_k (S_k) be the number of p_k^m which is larger (smaller) than $\bar{p}_k$.

- $p_k > \bar{p}_k$: Let S'_k be the number of p_k^m s.t. $\bar{p}_k \leq p_k^m \leq p_k$. We have

$$\begin{aligned} f_1(p_k) &\geq \sum_{m=1}^{M} |\bar{p}_k - p_k^m| + (p_k - \bar{p}_k)S_k - |p_k - \bar{p}_k|(L_k - S'_k) - |p_k - \bar{p}_k|S'_k \\ &= \sum_{m=1}^{M} |\bar{p}_k - p_k^m| + |p_k - \bar{p}_k|(S_k - L_k + S'_k - S'_k) = f_1(p'_k)\,. \end{aligned}$$

- $p_k < \bar{p}_k$: Let L'_k be the number of p_k^m s.t. $\bar{p}_k \geq p_k^m \geq p_k$. We have $f_1(p_k) \geq f_1(p'_k)$ by replacing S_k, L_K and S'_K respectively by L_k, S_K and L'_K

For $K > 2$, $\bar{\boldsymbol{p}}$ may not satisfy the probability axioms (see next Table).

	$K=3$
$\boldsymbol{p}^1$	0.8 0.1 0.1
$\boldsymbol{p}^2$	0.2 0.5 0.3
$\boldsymbol{p}^3$	0.1 0.4 0.5
$\bar{\boldsymbol{p}}$	0.2 0.4 0.3

	$K>3$
$\boldsymbol{p}^1$	0.4 0.2 0.4/(K-3) ... 0.4/(K-3)
$\boldsymbol{p}^2$	0.2 0.7 0.1/(K-3) ... 0.1/(K-3)
$\boldsymbol{p}^3$	0.1 0.6 0.3/(K-3) ... 0.3/(K-3)
$\bar{\boldsymbol{p}}$	0.2 0.6 0.3/(K-3) ... 0.3/(K-3)

When $K = 2$, the probability axioms of $\bar{\boldsymbol{p}}$ are ensured by the fact that the total rank of each distribution $\boldsymbol{p}^m$, $m \in [M]$, on the first and the second classes is always $M+1$ (as the masses should sum up to 1). Thus, $\bar{\boldsymbol{p}}$ is either one element of $\mathbf{H}(\boldsymbol{x})$ or the average of two elements of $\mathbf{H}(\boldsymbol{x})$. □

In the next section, we recall the inference problem with credal sets [19,32].

4 Inference Problem

As said, when our uncertainty is described by a credal set $\mathcal{P}(\mathcal{Y}\,|\,\boldsymbol{x})$, instead of a single probability $\boldsymbol{p}(\mathcal{Y}\,|\,\boldsymbol{x})$, it is necessary to make predictions using some theoretically founded decision rule extending classical expectation [19,32]. For any $\boldsymbol{p} \in \mathcal{P}(\mathcal{Y}\,|\,\boldsymbol{x})$ and any loss function ℓ, we shall denote by

$$\hat{y}_\ell^{\boldsymbol{p}} \in \operatorname*{argmin}_{\bar{\boldsymbol{y}} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \ell(y, \bar{y})\, \boldsymbol{p}(y\,|\,\boldsymbol{x})\,. \tag{19}$$

Definition 3. *An optimal set-valued prediction under the E-admissibility rule is*

$$\hat{\mathbf{Y}}_{\ell,\mathcal{P}}^{E} = \{y \in \mathcal{Y}\,|\,\exists\, \boldsymbol{p} \in \mathcal{P} \text{ s.t. } y = \hat{y}_\ell^{\boldsymbol{p}}\}\,. \tag{20}$$

Definition 4. *An optimal set-valued prediction* $\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}$ *under the Maximality rule is the set of the maximal, non-dominated elements of the partial order* $\pi^{\mathcal{P}}_\ell$ *such that* $\bar{y} \succ_{\ell,\mathcal{P}} \bar{y}'$ *if*

$$\inf_{\boldsymbol{p}\in\mathcal{P}} \mathbf{E}\boldsymbol{p}\left(\ell(y,\bar{y}') - \ell(y,\bar{y})\right) > 0\,. \tag{21}$$

In other words, we have

$$\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}} = \{\bar{y}\in\mathcal{Y} \mid \nexists\, \bar{y}' \text{ s.t. } \bar{y}' \succ_{\ell,\mathcal{P}} \bar{y}\}\,. \tag{22}$$

It is known that the set-valued prediction given by the E-admissibility rule is a subset of the one given by the Maximality rule [32].

In the following, we discuss the computational complexity of the inference problem when ℓ is the 0/1 loss, i.e., $\ell(y,\bar{y}) = [\![y \neq \bar{y}]\!]$, where $[\![A]\!] = 1$ if the predicate A is true and equals 0 otherwise

Let us start with the case of Maximality rule. For any $\boldsymbol{p} \in \mathbf{CH}_\alpha(\boldsymbol{x})$, we have

$$\mathbf{E}\boldsymbol{p}\left(\ell(y,\bar{y}') - \ell(y,\bar{y})\right) = \boldsymbol{p}(\bar{y}\,|\,\boldsymbol{x}) - \boldsymbol{p}(\bar{y}'\,|\,\boldsymbol{x})\,. \tag{23}$$

Thus, the relation $\bar{y} \succ_{\ell,\mathcal{P}} \bar{y}'$ holds if the maximum of the linear program

$$\underset{\boldsymbol{p}}{\text{maximize}} \quad f(\boldsymbol{p}) := \boldsymbol{p}(\bar{y}'\,|\,\boldsymbol{x}) - \boldsymbol{p}(\bar{y}\,|\,\boldsymbol{x}) \tag{24}$$

$$\text{subject to} \quad \boldsymbol{p} - \sum_{m=1}^{M_\alpha} \alpha_m\, \boldsymbol{p}^m = 0, \alpha_m \geq 0, \sum_{m=1}^{M_\alpha} \alpha_m = 1\,, \tag{25}$$

is negative. Note that if $f(\boldsymbol{p})$ has a maximum value on the feasible region, then it has this value on (at least) one of the extreme points, i.e., elements of $\mathbf{H}_\alpha(\boldsymbol{x})$ [25][Theorem 3.3]. Thus, a naive algorithmic solution is to compute $f(\boldsymbol{p})$ for the extreme $\boldsymbol{p}$ and compare it with 0. This requires time $O(K^2 M_\alpha)$ because in the worst case, one needs to check all the $K(K-1)$ relation $\bar{y} \succ_{\ell,\mathcal{P}} \bar{y}'$, $\bar{y} \neq \bar{y}' \in \mathcal{Y}$.

We now tackle the case of the E-admissibility rule. Reminding that, $\forall\, y \in \hat{\mathbf{Y}}^E_{\ell,\mathcal{P}}$, there must exist at least one $\boldsymbol{p} \in \mathbf{CH}_\alpha(\boldsymbol{x})$ such that $y = \hat{y}^{\boldsymbol{p}}_\ell$. This is equivalent to having at least one $\boldsymbol{p} \in \mathbf{CH}_\alpha(\boldsymbol{x})$ such that $\boldsymbol{p}(y\,|\,\boldsymbol{x}) \geq \boldsymbol{p}(y'\,|\,\boldsymbol{x}), y' \neq y$. Thus, given any outer approximation $\mathcal{Y}^O_{\ell,\mathcal{P}}$ of $\hat{\mathbf{Y}}^E_{\ell,\mathcal{P}}$ we can follow the suggestion of [19] and formulate the problem of checking whether a given $y \in \mathcal{Y}^O_{\ell,\mathcal{P}}$ satisfies the relation $y \in \hat{\mathbf{Y}}^E_{\ell,\mathcal{P}}$ as finding the maximum value of a linear program

$$\underset{\boldsymbol{p}}{\text{maximize}} \quad f(\boldsymbol{p}) := \boldsymbol{p}(y\,|\,\boldsymbol{x}) - \boldsymbol{p}(y'\,|\,\boldsymbol{x}) \tag{26}$$

$$\text{subject to} \quad \boldsymbol{p} - \sum_{m=1}^{M_\alpha} \alpha_m\, \boldsymbol{p}^m = 0, \alpha_m \geq 0, \sum_{m=1}^{M_\alpha} \alpha_m = 1\,, \tag{27}$$

$$\boldsymbol{p}(y\,|\,\boldsymbol{x}) - \boldsymbol{p}(y''\,|\,\boldsymbol{x}) \geq 0\,, y'' \in \mathcal{Y} \setminus \{y, y'\}\,, \tag{28}$$

where $y' \neq y$, and comparing it with 0. Hence, finding $\hat{\mathbf{Y}}^E_{\ell,\mathcal{P}}$ requires solving $|\mathcal{Y}^O_{\ell,\mathcal{P}}|$ linear programs, one per $y \in |\mathcal{Y}^O_{\ell,\mathcal{P}}|$. The naive algorithmic solution, i.e., iterating over all the extreme points, can not be applied here because a class y may be optimal only for probabilities in the interior of $\mathbf{CH}_\alpha(\boldsymbol{x})$.

5 Experiment

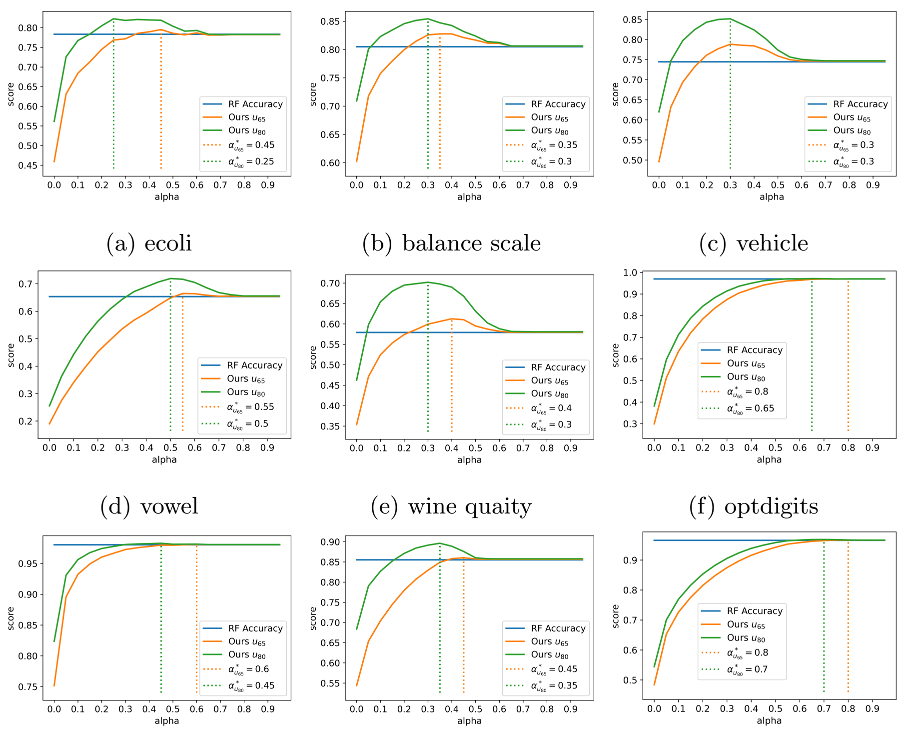
To motivate the potential use of the proposed framework, we perform some experiments on 9 tabular datasets from the UCI repository (cf. the left part of Table 1), following a 10-fold cross-validation procedure. We employ random forests (RFs) [18] (with default setting of scikit-learn) as the base learner. RFs are compared to an instantiation of our framework, where $\mathbf{H}_\alpha(\boldsymbol{x})$ is constructed under the f_{sqe} (13) and used to produce $\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}$. For each train test split, we follow a 10-fold nested cross-validation procedure to choose α optimizing u_{65}. The RF is then retrained using the entire training dataset and the chosen α is used to construct $\mathbf{H}_\alpha(\boldsymbol{x})$ during the inference phase. Source code is given in this link.

Table 1. Statistics of data sets (P is the number of features) and experimental results.

Statistics of data sets					Overall results				Cases of abstention	
					RF	Ours			RF	Ours
#	Name	N	P	K	Acc. (%)	u_{65} (%)	u_{80} (%)	$\lvert\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}\rvert$	Acc. (%)	Corr. (%)
1	ecoli	336	7	8	**78.35**	77.77	79.38	2.05	69.84	93.59
2	balance scale	625	4	3	80.50	**82.17**	83.15	2.02	26.75	67.75
3	vehicle	846	18	4	74.46	**78.16**	82.63	2.04	47.31	90.24
4	vowel	990	10	11	65.35	**65.89**	68.71	2.05	41.05	71.80
5	wine quality	1599	11	6	57.91	**61.67**	68.54	2.02	49.69	86.73
6	optdigits	1797	64	10	96.95	**97.03**	97.19	2.03	50.74	80.19
7	segment	2300	19	7	**98.05**	98.02	98.22	2.09	50.12	78.93
8	waveform	5000	21	3	85.52	**85.81**	88.33	2	62.06	99.91
9	letter	20000	16	26	96.57	**96.58**	96.64	2.03	34.33	81.71

Overall results (accuracy, u_{65} and u_{80} scores [38] and cardinality $|\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}|$) show that our proposal may provide a promising correctness-precision trade-off, compared to RFs itself. Ideally, a reliable classifier should be more cautious on difficult cases, on which the conventional classifier is likely to fail [27,34]. To verify this ability of our proposal, for each dataset, we report the correctness (i.e., the percentage of times the true class is in $\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}$), given the prediction was imprecise, versus the accuracy of RF on those instances. The results (in the right part of Table 1) strongly support our proposal. This also suggests that the use of the E-admissibility rule (listed as future work) may improve the overall results because, under the f_{sqe}, predictions of RFs should belong to $\hat{\mathbf{Y}}^E_{\ell,\mathcal{P}} \subset \hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}$ [32].

To gain more insights about the influence of α, we consider u_{65} and u_{80} scores on the test set as functions of the value of α. The results in Figure 1 are indeed in agreement with our expectations. The $\mathcal{P}(\mathcal{Y}\,|\,\boldsymbol{x})$ induced by $\mathbf{H}_\alpha(\boldsymbol{x})$ with small α may contain extreme/noisy distributions and produce large $\hat{\mathbf{Y}}^M_{\ell,\mathcal{P}}$ (resulting in low u_{65} and u_{80} scores). Moderate α may provide a nice correctness-precision trade-off (reflected via promising u_{65} and u_{80} scores). For large α, $\mathcal{P}(\mathcal{Y}\,|\,\boldsymbol{x})$ is shrunken as (small) neighborhood of the $\boldsymbol{p}^*$ (5) and our proposal (under f_{sqe}) becomes similar to RFs, which use the $\boldsymbol{p}^*$ to make predictions. The results also suggest that, in practice, nested cross-validation procedure can help us to find some good value of α (even if the ideal gain provided by the optimal α is small).



(a) ecoli (b) balance scale (c) vehicle

(d) vowel (e) wine quaity (f) optdigits

(g) segment (h) waveform (i) letter

Fig. 1. u_{65} and u_{80} scores on the test set as functions of the value of α

6 Conclusion

We propose a simple, easy to use quantile-based framework for estimating credal sets using output of ensemble methods, that can also cope with complex types of data. Preliminary experiments suggest that our proposal may provide a promising correctness-precision trade-off, compared to ensemble methods. To seek for a complete picture on the usefulness of our proposal, we envision the following works: (1) implement our proposal with other distances and the E-admissibility rule and analyze (dis)advantages provided by different combinations of distance and decision rule, (2) include threshold-based classifiers as competitors, and (3) include complex types of data (such as images) into our empirical studies.

References

1. Augustin, T., Coolen, F.P., De Cooman, G., Troffaes, M.C.: Introduction to imprecise probabilities. John Wiley & Sons (2014)
2. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. Journal of Machine Learning Research **9**(Aug), 1823–1840 (2008)
3. Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
4. Cha, S.H.: Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical models and Methods in Applied Sciences **1**(4), 300–307 (2007)
5. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on information theory **16**(1), 41–46 (1970)
6. Corani, G., Zaffalon, M.: Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. Journal of Machine Learning Research **9**(4) (2008)
7. Cortes, C., DeSalvo, G., Mohri, M.: Learning with rejection. In: Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT). pp. 67–82. Springer Verlag (2016)
8. Cozman, F.G.: Credal networks. Artificial intelligence **120**(2), 199–233 (2000)
9. Datta, B.N.: Numerical linear algebra and applications, vol. 116. Siam (2010)
10. De Leon, A., Soo, A., Williamson, T.: Classification with discrete and continuous variables via general mixed-data models. Journal of Applied Statistics **38**(5), 1021–1032 (2011)
11. Del Coz, J.J., Díez, J., Bahamonde, A.: Learning nondeterministic classifiers. Journal of Machine Learning Research **10**(10) (2009)
12. Dietterich, T.G.: Ensemble methods in machine learning. In: Proceedings of the First International workshop on multiple classifier systems (MCS). pp. 1–15. Springer (2000)
13. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the 17th International Conference on Artificial Intelligence (IJCAI). pp. 973–978 (2001)
14. Franc, V., Prusa, D.: On discriminative learning of prediction uncertainty. In: Proceedings of the 36th International Conference on Machine Learning (ICML). pp. 1963–1971 (2019)
15. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. International statistical review **70**(3), 419–435 (2002)
16. Grandvalet, Y., Rakotomamonjy, A., Keshet, J., Canu, S.: Support vector machines with a reject option. In: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS) (2008)
17. Hellman, M.E.: The nearest neighbor classification rule with a reject option. IEEE Transactions on Systems Science and Cybernetics **6**(3), 179–185 (1970)
18. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition (ICDAR). vol. 1, pp. 278–282. IEEE (1995)
19. Jansen, C., Schollmeyer, G., Augustin, T.: Quantifying degrees of e-admissibility in decision making with imprecise probabilities. In: Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld, pp. 319–346. Springer (2022)
20. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL). pp. 25–32 (1999)

21. Levi, I.: The enterprise of knowledge: An essay on knowledge, credal probability, and chance. MIT press (1983)
22. Mantas, C.J., Abellan, J.: Credal-c4. 5: Decision tree based on imprecise probabilities to classify noisy data. Expert Systems with Applications **41**(10), 4625–4637 (2014)
23. Montes, I., Miranda, E., Destercke, S.: Unifying neighbourhood and distortion models: part i–new results on old models. International Journal of General Systems **49**(6), 602–635 (2020)
24. Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., Waegeman, W.: Efficient set-valued prediction in multi-class classification. Data Mining and Knowledge Discovery **35**(4), 1435–1469 (2021)
25. Murty, K.G.: Linear programming. Springer (1983)
26. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML). pp. 689–696 (2011)
27. Nguyen, V.L., Destercke, S., Masson, M.H., Hüllermeier, E.: Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). pp. 5089–5095 (2018)
28. Pugh, C.C.: Real mathematical analysis. Undergraduate Texts in Mathematics (2015)
29. Rahimian, H., Mehrotra, S.: Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659 (2019)
30. Rockafellar, R.T.: Lagrange multipliers and optimality. SIAM review **35**(2), 183–238 (1993)
31. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research **11**, 1517–1561 (2010)
32. Troffaes, M.C.: Decision making under uncertainty using imprecise probabilities. International journal of approximate reasoning **45**(1), 17–29 (2007)
33. Xu, Z., So, D.R., Dai, A.M.: Mufasa: Multimodal fusion architecture search for electronic health records. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI). vol. 35, pp. 10532–10540 (2021)
34. Yang, G., Destercke, S., Masson, M.H.: Nested dichotomies with probability sets for multi-class classification. In: Proceedings of the Twenty-first European Conference on Artificial Intelligence (ECAI). pp. 363–368 (2014)
35. Yang, Y., Krompass, D., Tresp, V.: Tensor-train recurrent neural networks for video classification. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 3891–3900 (2017)
36. Yin, M., Sui, Y., Liao, S., Yuan, B.: Towards efficient tensor decomposition-based dnn model compression with optimization framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10683 (2021)
37. Zaffalon, M.: The naive credal classifier. Journal of statistical planning and inference **105**(1), 5–21 (2002)
38. Zaffalon, M., Corani, G., Mauá, D.: Evaluating credal classifiers by utility-discounted predictive accuracy. International Journal of Approximate Reasoning **53**(8), 1282–1301 (2012)