Cautious classifier ensembles for set-valued decision-making

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson

▶ **To cite this version:**

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson. Cautious classifier ensembles for set-valued decision-making. International Journal of Approximate Reasoning, 2025, 177, pp.109328. 10.1016/j.ijar.2024.109328 . hal-04885913

HAL Id: hal-04885913
https://hal.science/hal-04885913v1

Submitted on 14 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

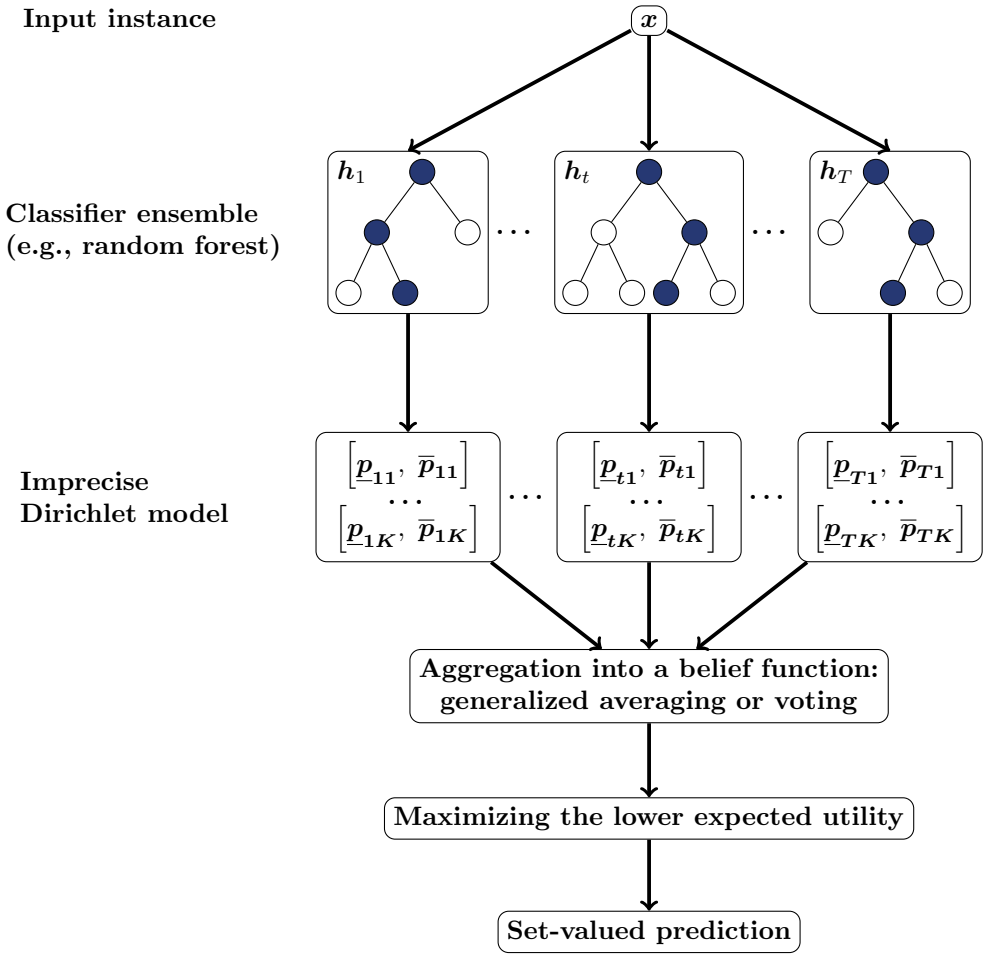
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Graphical Abstract

Cautious classifier ensembles for set-valued decision-making

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson



Highlights

Cautious classifier ensembles for set-valued decision-making

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson

- Averaging and voting in ensemble learning generalized for cautious classification.
- Efficient cautious decision-making strategy by maximizing the lower expected utility.
- Great trade-off between prediction accuracy and imprecision under high uncertainty.

Cautious classifier ensembles for set-valued decision-making

Haifei Zhang[a,*], Benjamin Quost[a], Marie-Hélène Masson[a,b]

[a] *UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, Compiègne 60200, France*
[b] *IUT de l'Oise, Université de Picardie Jules Verne, Beauvais 60000, France*

Abstract

Classifiers now demonstrate impressive performances in many domains. However, in some applications where the cost of an erroneous decision is high, set-valued predictions may be preferable to classical crisp decisions, being less informative but more reliable. Cautious classifiers aim at producing such imprecise predictions so as to reduce the risk of making wrong decisions. In this paper, we describe two cautious classification approaches rooted in the ensemble learning paradigm, which consist in combining probability intervals. These intervals are aggregated within the framework of belief functions, using two proposed strategies that can be regarded as generalizations of classical averaging and voting. Our strategies aim at maximizing the lower expected discounted utility to achieve a good compromise between model accuracy and determinacy. The efficiency and performances of the proposed procedures are illustrated using imprecise decision trees, thus giving birth to cautious variants of the random forest classifier. The performance and properties of these variants are illustrated using 15 datasets.

Keywords: Imprecise classification, Set-valued predictions, Belief functions, Imprecise Dirichlet model, Ensemble learning

1. Introduction

Ensemble learning, which amounts to build a complex model by pooling a set of base models, has emerged as one of the most widely used techniques, in particular for processing tabular data. Arguably its most prominent representative, random forests [1, 2] consist in combining the outputs of classification or regression trees; their efficiency and accuracy have been demonstrated in a wide range of domains. In their classification version, tree outputs are either precise class probability estimates or decisions, obtained from class counts for the training instances falling into the leaf nodes; and decisions are classically made either by averaging the probabilities or by majority voting on the chosen classes. However, trees—and more generally traditional classifiers—may exhibit poor robustness when faced with low-quality data, e.g., in the presence of noise in the training labels, or for instances located in low-density regions of the input space. To overcome this issue, previous works have proposed to use the imprecise Dirichlet model (IDM) to replace precise class probability estimates with convex sets of probability distributions (in the form of probability intervals) whose size depends on the number of training samples [3, 4]: the more samples in a leaf, the shorter the length of the probability intervals.

The joint use of the IDM and decision trees has already been explored in two directions. First, it has been used to improve the training of single trees or tree ensembles. Credal decision trees (CDT) [5, 6] and credal random forests (CRF) [7] use the maximum entropy principle to select split features and values from the probability intervals obtained via the IDM, thus improving their robustness to data noise. To enhance

*Corresponding author, with the current address: Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Etienne 42000, France, and the Email address: haifei.zhang@univ-st-etienne.fr

Email addresses: haifei.zhang@hds.utc.fr (Haifei Zhang), benjamin.quost@hds.utc.fr (Benjamin Quost), mylene.masson@hds.utc.fr (Marie-Hélène Masson)

the generalization performance of tree ensembles trained on small datasets, data sampling and augmentation based on the IDM probability intervals have been proposed to train deep forests [8] and weights associated with each tree in the ensemble can be learned to further optimize their combination [9].

Second, the probability intervals given by the IDM can also be used to make cautious decisions, thereby reducing the risk of prediction error [3, 10]. A cautious decision of a classification task is often in the form of a set-valued prediction: i.e., a cautious classifier may return a set of plausible classes instead of a single class when the uncertainty is too high. An imprecise credal decision tree (ICDT) [11] is a single tree where set-valued predictions are returned by applying the interval dominance principle [12] to the probability intervals obtained via the IDM. In tree ensembles, applying cautious decision-making strategies becomes more complex. One group of approach consists of aggregating the probability intervals given by the trees—for example by conjunction, disjunction, or averaging—before making cautious decisions by computing a partial order between the classes, e.g., using interval dominance [13, 14]. Another group of approaches consists of allowing each tree to make a cautious decision first before pooling them. The Minimum-Vote-Against (MVA) is such an approach, where the set of classes with minimal opposition are retained [15]. It should be noted that MVA generally results in precise predictions, whereas disjunction and averaging often turn out to be inconclusive. Even worse, using conjunction very frequently results in empty predictions due to conflict. In other words, these methods hardly achieve a good compromise between accuracy and cautiousness.

To address this issue, a generalized voting aggregation strategy for binary cautious classification within the theoretical framework of belief functions was proposed in [16, 17]. In this paper, we extend these previous works to the multi-class case, as illustrated in Fig. 1. The contributions in this paper can be summarized as follows:

1. we generalize the averaging and voting aggregation strategies classically used in ensemble learning to cautious classification problems;
2. we propose an efficient implementation of these principles, by maximizing a lower expected utility criterion;
3. through extensive experiments[1] realized using classification trees as base classifiers, we demonstrate that the proposed approaches can achieve a good compromise between the accuracy and the cautiousness of the model, especially in the presence of high uncertainty.

The structure of this paper is as follows. After recalling background material in Section 2, we propose in Section 3 an efficient way for computing the subset of classes maximizing the lower expected utility given a belief function, based on which two cautious decision-making strategies presented in Section 4, which generalize averaging and voting for imprecise tree ensembles, are deduced. The experiments reported in Section 5 demonstrate the ability of our approach to reach a good compromise between accuracy and determinacy, especially when facing high uncertainty, and to remain tractable even in the case of a high number of classes. Finally, a conclusion is drawn in Section 6.

2. Preliminaries

In this section, we provide background knowledge about the imprecise Dirichlet model [4], belief functions [18, 19], and decision-making strategies in the belief functions framework [20].

2.1. Imprecise Dirichlet model

Let $H = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T\}$ be an ensemble of classifiers $\boldsymbol{h}_t$ (e.g., trees in a random forest), trained on a classification problem of $K \geq 2$ classes. Let $\boldsymbol{\mathcal{R}}_t(\boldsymbol{x})$ be the (multidimensional) decision region for classifier $\boldsymbol{h}_t$ in which a given test instance $\boldsymbol{x} \in \mathcal{X}$ is located—typically, for a tree, $\boldsymbol{\mathcal{R}}_t(\boldsymbol{x})$ is defined by the leaf in which $\boldsymbol{x}$ falls; and let n_{tj} denote the number of training samples of class c_j in $\boldsymbol{\mathcal{R}}_t(x)$.

[1] Our code is available on GitHub: https://github.com/Haifei-ZHANG/Cautious-Random-Forest.

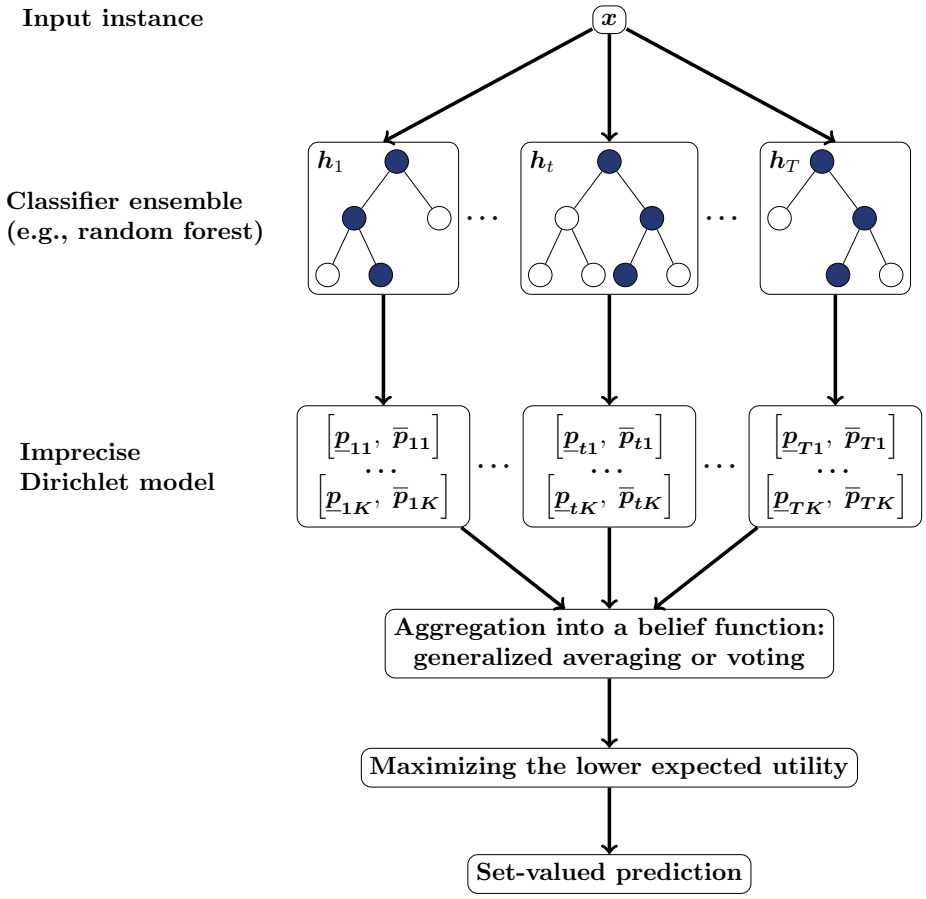


Figure 1: Overview of the proposed aggregation strategy.

The IDM consists in using a family of Dirichlet priors for estimating the class posterior probabilities $\mathbb{P}(c_j|\boldsymbol{x}, \boldsymbol{\mathcal{R}}_t)$, rather than a single Dirichlet prior: as a consequence, each classifier provides a set of probability intervals

$$I_{tj}(\boldsymbol{x}) = \left[\underline{p}_{tj}, \overline{p}_{tj}\right] = \left[\frac{n_{tj}}{N_t + s}, \frac{n_{tj} + s}{N_t + s}\right], \; j = 1, \ldots, K; \tag{1}$$

here, $n_{tj}(\boldsymbol{x})$ and $N_t(\boldsymbol{x}) = \sum_{j=1}^{K} n_{tj}(\boldsymbol{x})$ are the numbers of instances from class c_j and the total number of instances in $\boldsymbol{\mathcal{R}}_t(\boldsymbol{x})$, and s can be interpreted as a number of additional (virtual) samples with unknown class information also located in $\boldsymbol{\mathcal{R}}_t(\boldsymbol{x})$. Although the issue of choosing an appropriate value for s remains open, the parameter is often set in practice to $s = 1$ or $s = 2$, following [4]. The IDM provides a natural local estimate of epistemic uncertainty, i.e., the uncertainty caused by the lack of (training) information when a decision must be made for $\boldsymbol{x}$.

In Section 4, we describe two approaches for pooling these probability intervals into a mass function m, from which a (cautious) decision $\boldsymbol{h}(\boldsymbol{x})$ can then be made.

2.2. Belief functions

The theory of belief functions, also referred to as Dempster-Shafer theory (DST) or the theory of evidence, is a mathematical framework for dealing with uncertainty and reasoning with incomplete or conflicting information [18, 19]. It provides a formal way to combine and reason with uncertain information from multiple sources, allowing for a more robust and cautious approach to decision-making. DST has found

applications in various fields, such as information fusion [21, 22, 23], pattern recognition [24, 25, 26, 27], semantic segmentation [28], fault diagnosis [29, 30, 31], etc.

Let $\Omega = \{c_1, c_2, \ldots, c_K\}$ denote the frame of discernment, i.e., the finite set of mutually exclusive alternatives for our class variable C. A mass function is a mapping $m : 2^\Omega \to [0, 1]$, such that $\sum_{A\subseteq\Omega} m(A) = 1$. Any subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal element of m. The value $m(A)$ measures the degree of evidence supporting the assumption "$C \in A$", but nothing more specific; $m(\Omega)$ represents the degree of total ignorance, i.e., the belief mass that could not be assigned to any specific subset of classes. A mass function is Bayesian if its focal elements are singletons only, and quasi-Bayesian if they are only singletons and Ω.

The belief and plausibility functions can be computed from the mass function m, which are respectively defined as

$$Bel(A) = \sum_{B\subseteq A} m(B), \quad \forall A \in \Omega, \tag{2}$$

$$Pl(A) = \sum_{B\cap A\neq\emptyset} m(B), \quad \forall A \in \Omega. \tag{3}$$

In a nutshell, $Bel(A)$ measures the total degree of support to "$C \in A$", and $Pl(A)$ is the degree of belief that does not contradict "$C \in A$". These two functions are dual since $Bel(A) = 1 - Pl(\overline{A})$, with $\overline{A} = \Omega \setminus A$.

The mass, belief, and plausibility functions are in one-to-one correspondence and can be retrieved from each other. When it comes to making a decision, a mass function can be transformed into a probability distribution through the pignistic transformation [32, 33]:

$$BetP(c_k) = \sum_{A\subseteq\Omega, c_k\in A} \frac{m(A)}{|A|}, \quad \forall c_k \in \Omega, \tag{4}$$

in which the mass of focal sets is equally assigned to their elements.

2.3. Decision-making in the belief function framework

A decision problem can be seen as choosing the most desirable action among a set of alternatives $F = \{f_1, \ldots, f_L\}$, according to a set of states of nature $\Omega = \{c_1, \ldots, c_K\}$ and a corresponding utility matrix U of dimensions $L \times K$. The value of $u_{ij} \in \mathbb{R}$ is the utility or payoff obtained if action $f_i, i = 1, \ldots, L$ is taken and state $c_j, j = 1, \ldots, K$ occurs. In the classification setting, the action is that of assigning a test instance $\boldsymbol{x}$ to a class (or a subset of classes in cautious classification), the states of nature mentioned above being obviously the set of (actual) classes for $\boldsymbol{x}$.

Assume our knowledge of the class of the test instance is represented by a mass function m: then, several criteria to make decisions have been proposed in the framework of belief functions [20]. Should only singleton assignments be considered, a convenient way to build a complete preference order over the set of classes consists in transforming the mass function m into a probability distribution according to Eq. (4), and then calculating the expected utility of each action which results in assigning a singleton class: this latter, called the *pignistic expected utility*, is defined as

$$\mathbb{E}_{BetP}(f_i, U) = \sum_{k=1}^{K} BetP(c_k) u_{ik}. \tag{5}$$

However, when actions f_i are not restricted to assigning a single class but consider subsets $A \subseteq \Omega$ of classes, denoted as f_A, the expected utility criterion may be extended to the *lower and upper expected utilities*, respectively defined as the weighted sums of the minimum and maximum utility within each focal set:

$$\underline{\mathbb{E}}_m(f_A, U) = \sum_{B\subseteq\Omega} m(B) \min_{c_k\in B} u_{Ak}, \tag{6}$$

$$\overline{\mathbb{E}}_m(f_A, U) = \sum_{B\subseteq\Omega} m(B) \max_{c_k\in B} u_{Ak}. \tag{7}$$

It is obvious that $\underline{\mathbb{E}}(m, f_A, U) \leq \overline{\mathbb{E}}(m, f_A, U)$ and only when m is Bayesian, the equality applies, as we retrieve the probabilistic case. The *Hurwicz expected utility* is a convex combination of $\underline{\mathbb{E}}_m(f_A, U)$ and $\overline{\mathbb{E}}_m(f_A, U)$, defined as:

$$\mathbb{E}_{m,\alpha}(f_A, U) = \alpha\, \underline{\mathbb{E}}(m, f_A, U) + (1 - \alpha)\, \overline{\mathbb{E}}(m, f_A, U), \tag{8}$$

where $\alpha \in [0, 1]$ is called the pessimism index. The *minimax regret* criterion can also be extended to belief functions. The regret that action f_A is chosen whereas state c_k occurs is defined as $r_{Ak} = \max_B u_{Bk},\ \forall B \subseteq \Omega$. The expected maximal regret of action f_A is defined as

$$\overline{R}_m(f_A, U) = \sum_{B \subseteq \Omega} m(B) \max_{c_k \in B} r_{Ak}. \tag{9}$$

From now on, for the sake of simplicity, we will identify by abuse of notation the action f_A of choosing a subset $A \subseteq \Omega$ with the subset A itself.

It should be noted that in all of these decision-making strategies, if class assignments are restricted to singletons, i.e., $|A| = 1$, then, all of these four expected measures lead to computing complete preorders among all possible precise assignments, and the one that reaches the highest expected utility or the lowest expected maximal regret will be selected, which results in precise predictions [34]. Otherwise, if all possible partial assignments (i.e., any subset $A \subseteq \Omega$) are considered as legitimate decisions, the lower, upper, Hurwicz expected utilities, and the expected maximal regret establish complete preorders among partial assignments: then, selecting the subset that reaches the highest expected utility or the lowest expected maximal regret leads to make set-valued or cautious predictions [35].

2.4. Evaluation of cautious classifiers

Unlike traditional classifiers, cautious classifiers may return set-valued decisions: in this case, classical evaluation criteria are no longer applicable. Various criteria have been proposed to evaluate the quality of such predictions, which depends both on their ability to avoid making mistakes and on their level of informativeness:

- *determinacy* counts the proportion of samples that are determinately classified (i.e., for which the classifier outputs a single class);
- *single-set accuracy* measures the proportion of correct decisions among the determinate ones;
- *set accuracy* measures the proportion of indeterminate predictions containing the actual class (computed only over indeterminate predictions);
- *set size* gives the average size of indeterminate predictions;
- the *discounted utility* calculates the expected utility of making a correct decision, discounted by the size of the predicted set: two special cases are the classical u_{65} and u_{80} measures described below.

Let $\boldsymbol{h}(\boldsymbol{x}) = A \subseteq \Omega$ be the outcome of the decision procedure for a test sample $\boldsymbol{x}$ with actual class c. Zaffalon et al. [36] proposed to evaluate the quality of this decision using a discounted utility function u_α, which rewards both its cautiousness and accuracy as follows:

$$u_\alpha(A, c) = d_\alpha(|A|)\mathbb{1}(c \in A), \tag{10}$$

where $|A|$ is the cardinality of A, $d_\alpha(.)$ is a discount ratio that adjusts the reward for cautiousness (which is considered preferable to random guessing whenever $d_\alpha(|A|) > 1/|A|$), and $\mathbb{1}(\cdot)$ stands for the indicator function. The u_{65} and u_{80} scores are two notable special cases of u_α, respectively obtained using

$$d_{65}(|A|) = \frac{1.6}{|A|} - \frac{0.6}{|A|^2}, \quad d_{80}(|A|) = \frac{2.2}{|A|} - \frac{1.2}{|A|^2}.$$

3. Lower discounted utility maximization

Let m be a mass function defined on the frame of discernment $\Omega = \{c_1, \ldots, c_K\}$ with $K \geq 2$, representing our knowledge of the actual class of a given instance $\boldsymbol{x}$. Assume we want to compute the lower expected utility $\underline{\mathbb{E}}_m(f_A, U)$ of the action f_A (which consists in assigning $\boldsymbol{x}$ to the subset of classes $A \subseteq \Omega$), according to a utility matrix U, as per Eq. (6). When the utility function has the same form as in Eq. (10), this amounts to calculating the product of the belief degree $Bel(A)$ with the corresponding discounted utility $d_\alpha(|A|)$, as shown in Theorem 3.1.

Theorem 3.1. *Let us consider the utility matrix U of general term $u_{Ak} = d_\alpha(|A|)\mathbb{1}(c_k \in A)$, where c_k refers to the ground truth and $A \subseteq \Omega$ to a set-valued decision; then, the lower expected utility is equal to $\underline{\mathbb{E}}_m(A, U) = d_\alpha(|A|)Bel(A)$.*

Proof 3.1. *Following Eq.* (6)*, and taking any $A \subseteq \Omega$ as action, we have*

$$\begin{aligned}
\underline{\mathbb{E}}_m(A, U) &= \sum_{B \subseteq \Omega} m(B) \min_{c_k \in B} u_{Ak} \\
&= \sum_{B \subseteq \Omega} m(B) \min_{c_k \in B} [d_\alpha(|A|)\mathbb{1}(c_k \in A)] \\
&= d_\alpha(|A|) \sum_{B \subseteq \Omega} m(B) \min_{c_k \in B} \mathbb{1}(c_k \in A) \\
&= d_\alpha(|A|) \sum_{B \subseteq A} m(B) \\
&= d_\alpha(|A|) Bel(A).
\end{aligned}$$

Indeed, for any subset $B \subseteq \Omega$ such that $B \cap A \neq \emptyset$ and $B \nsubseteq A$, there obviously exists $c_k \in B$ such that $c_k \notin A$: thus, $\min_{c_k \in B} \mathbb{1}(c_k \in A) = 1$ if and only if $B \subseteq A$.

Example 1 (Discounted utility matrix and lower discounted utility). *Let us consider a problem defined on $\Omega = \{c_1, c_2, c_3\}$ and an associated mass function m defined as*

$$m(\{c_1\}) = 0.45,\ m(\{c_2\}) = 0.35,\ m(\{c_3\}) = 0.1,\ m(\Omega) = 0.1.$$

In Table 1, the columns associated with u_{Ak} indicate the utility of taking A as the decision and the corresponding state of nature occurs. For example. if we take $\{c_1, c_2\}$ as the decision and c_1 (or c_2) is the true state of nature, then the discounted utility is 0.65. Otherwise, if c_3 is the true state of nature, then the discounted utility is zero.

The column $Bel(A)$ is calculated via Eq. (2)*, which is the belief degree of each decision. The columns associated $\underline{\mathbb{E}}_m(A, U)$ is then the expected lower discounted utility when we take A as the decision and the corresponding state of nature occurs, which is calculated according to Theorem 3.1. It is obvious that $\{c_1, c_2\}$ reaches the maximum of lower expected discounted utility, so it is taken as the decision.*

Following the definition of lower expected utility in Eq. (6)*, $\underline{\mathbb{E}}_m(\{c_1, c_2\}, U)$ can be calculated as follows*

$$\begin{aligned}
\underline{\mathbb{E}}_m(\{c_1, c_2\}, U) &= m(\{c_1\}) \times \min(0.65) + m(\{c_2\}) \times \min(0.65), \\
&\quad + m(\{m_3\}) \times \min(0) + m(\Omega) \times \min(0.65, 0.65, 0), \\
&= 0.45 \times 0.65 + 0.35 \times 0.65 + 0.1 \times 0 + 0.1 \times 0, \\
&= 0.52.
\end{aligned}$$

Remark 3.1. *Obviously, in our setting, the lower expected utility is not the only criterion based on which set-valued decisions can be made.*

Table 1: Example of lower expected discounted utility with three classes

A	u_{Ak}			$d_{65}(\lvert A\rvert)$	$Bel(A)$	$\underline{\mathbb{E}}_m(A,U) = d_{65}(\lvert A\rvert)Bel(A)$
	c_1	c_2	c_3			
$\{c_1\}$	1	0	0	1	0.45	0.45
$\{c_2\}$	0	1	0	1	0.35	0.35
$\{c_3\}$	0	0	1	1	0.1	0.1
$\{c_1, c_2\}$	0.65	0.65	0	0.65	0.8	0.52
$\{c_1, c_3\}$	0.65	0	0.65	0.65	0.55	0.3575
$\{c_2, c_3\}$	0	0.65	0.65	0.65	0.45	0.2925
$\{c_1, c_2, c_3\}$	0.4667	0.4667	0.4667	0.4557	1	0.4667

As we can see in Theorem 3.1, computing $\underline{\mathbb{E}}_m(A,U)$ requires to compute the belief degree $Bel(A)$, based on the mass degrees $m(B)$ of all subsets $B \subseteq A$, $B \neq \emptyset$. In principle, maximizing $\underline{\mathbb{E}}_m(A,U)$ across all subsets requires to check all subsets $A \subseteq \Omega$, the number of which is $2^{|\Omega|}$; however, as it will be seen, several properties make it possible to decrease this complexity, by stopping the search once a given subset cardinality is attained.

Maximizing the upper expected utility $\overline{\mathbb{E}}_m(A,U)$, on the other hand, would require to calculate the plausibility degree $Pl(A)$, based on the mass degrees $m(B)$ of all subsets $B \cap A \neq \emptyset$, which significantly raises the complexity. Since the Hurwicz criterion takes the upper expected utility into account, its complexity is also affected by this issue. The expected maximal regret also faces a similar problem, requiring to scan all subsets $A \subseteq \Omega$.

These complexity considerations constitute yet another incentive to choose the lower expected utility as the criterion for making cautious decisions.

4. Proposed aggregation strategies

In this section, we detail our contributions: first, we expose how the classical averaging and voting strategies in ensemble learning can be generalized to the belief-theoretic case; then, we detail how cautious decisions based on the lower expected utility can be efficiently made, by leveraging the formulation of the criterion provided in Theorem 3.1.

4.1. Generalization of averaging

We start with the generalized averaging strategy. First, we show that the outputs of imprecise classifiers are aggregated into a quasi-Bayesian mass function. Then, we demonstrate the cautious decision-making problem can be solved in a time complexity of $O(K \log K)$, where K is the number of classes in the data.

Classifier averaging

We assume that each classifier output $\boldsymbol{h}_t(\boldsymbol{x})$ is a set of probability intervals

$$I_{tk}(\boldsymbol{x}) = \left[\underline{p}_{tk}(\boldsymbol{x}), \overline{p}_{tk}(\boldsymbol{x})\right], \quad t = 1, \ldots, T, \ k = 1, \ldots, K.$$

In the case of decision trees, these intervals are typically obtained using the imprecise Dirichlet model, as in Eq. (1). According to [37], the corresponding quasi-Bayesian mass function associated with $I_{tk}(\boldsymbol{x})$ is

$$m_t(\{c_k\}) = \underline{p}_{tk}, \ k = 1, \ldots, K; \quad m_t(\Omega) = 1 - \sum_{k=1}^{K} m_t(\{c_k\}). \tag{11}$$

A straightforward classifier combination strategy consists in averaging these mass functions, resulting in the following quasi-Bayesian mass function:

$$m(\{c_k\}) = \frac{\sum_{t=1}^{T} m_t(\{c_k\})}{T}, \ k = 1, \ldots, K; \quad m(\Omega) = \frac{\sum_{t=1}^{T} m_t(\Omega)}{T}. \tag{12}$$

Decision-making

To make a decision based on the mass function defined by Eq. 12, we start by building the sequence of nested subsets $A_{(k)} \subseteq \Omega$, $k = 1, \ldots, K$: defining $A_{(0)} = \emptyset$, we compute

$$A_{(k+1)} = A_{(k)} \cup \arg \max_{c_\ell \notin A_{(k)}} m(\{c_\ell\}), \quad \text{for } k = 1, \ldots, K, \tag{13}$$

thus repeatedly aggregating the class with the highest mass among the remaining ones. Remark that we have $A_{(K)} = \Omega$. As shown by Theorem 4.1, the sequence of nested subsets $A_{(1)}, \ldots, A_{(K)}$ necessarily contains the subset $A^\star$ with the highest lower expected utility, i.e., $A^\star = \arg\max_A \underline{\mathbb{E}}_m(A, U)$. Therefore, once the sequence has been built, we may simply scan it to determine $A^\star$ in linear complexity. For the sake of simplicity, we will use the notation $\underline{\mathbb{E}}(A)$ to refer to the lower expected utility $\underline{\mathbb{E}}_m(A, U)$ as a function of $A \subseteq \Omega$, when both the mass function m and the utility matrix U are fixed.

Theorem 4.1. *Consider a quasi-Bayesian mass function m, where the classes are already sorted by decreasing mass: $m(\{c_{(k)}\}) \geq m(\{c_{(k+1)}\})$, for $k = 1, \ldots, K-1$. The subset $A^\star = \arg\max_A \underline{\mathbb{E}}(A)$ maximizing the lower expected utility can be identified in complexity $O(K)$ by scanning the sequence of nested subsets*

$$A_{(1)} = \{c_{(1)}\} \subset A_{(2)} = \{c_{(1)}, c_{(2)}\} \subset \cdots \subset A_{(K)} = \Omega.$$

Proof 4.1. *Since the masses $m(\{c_{(k)}\})$ are sorted in decreasing order, the focal element with the highest belief among all focal elements of cardinality i is $A_{(i)} = \{c_{(k)}, k = 1, \ldots, i\}$: indeed, for any $B \subseteq \Omega$ such that $|B| = i$, we have*

$$Bel(A_{(i)}) = \sum_{k=1}^{i} m(\{c_{(k)}\}) \geq Bel(B).$$

Since $d_\alpha(|A|)$ only depends on $|A|$, $A_{(k)}$ maximizes the lower expected utility over all subsets of size i.

As a consequence, selecting the subset with maximal lower expected utility in the sequence of nested subsets $A_{(k)}$, $k = 1, \ldots, K$ computes the maximizer $A^\star$ in time complexity $O(K)$.

It should be noted that the total complexity of the decision-making starting from the aggregated mass function provided by Eq. (12) is $O(K \log K)$ due to sorting the classes by decreasing mass. We also remark that the property established in Theorem 4.1 also holds for other kinds of mass functions, and notably for consonant mass functions.

Theorem 4.2. *Consider a consonant mass function m_2, i.e., with nested focal elements $A_{(1)} \subset A_{(2)} \subset \ldots \subset A_{(R)}$.*

The maximizer $A^\star$ of the lower expected utility can be identified in complexity $O(R)$ by scanning the focal elements of m_2.

Proof 4.2. *The proof is trivial and invokes similar arguments to that of Theorem 4.1. Let $B \subseteq \Omega$ be a subset which does not belong to the sequence of focal elements of m_2.*

- *If B does not contain any subset $A_{(i)}$ in this sequence, then B has necessarily a zero belief, and therefore $\underline{\mathbb{E}}(B) = 0$.*
- *If B does contain at least one subset in this sequence, assume $A_{(i)} \subset B$ is the largest of those subsets in the sequence belonging to B: then, we have that $Bel(B) = Bel(A_{(i)})$, but $\underline{\mathbb{E}}(B) < \underline{\mathbb{E}}(A_{(i)})$ since $|A_{(i)}| < |B|$.*

As a consequence, the maximizer of the lower expected utility necessarily belongs to the sequence of nested focal elements $A_{(r)}$, $r = 1, \ldots, R$ of the mass function m_2, and can be identified by scanning this sequence.

As a matter of fact, finding the maximizer $A^\star$ of the lower expected utility does not require to scan the whole sequence of nested subsets as suggested in Theorem 4.1 or Theorem 4.2. Indeed, we can proceed by scanning subsets of increasing cardinality and stop whenever a subset does not improve the lower expected utility, as shown by Proposition 4.1.

Algorithm 1: Cautious Decision Making by Averaging

Input: Classifier outputs $I_{tk}(\boldsymbol{x}) = [\underline{p}_{tk}(\boldsymbol{x}); \overline{p}_{tk}(\boldsymbol{x})]_{t=1,\ldots,T}^{k=1,\ldots,K}$; discount ratio d_α
Output: Decision $A \subseteq \Omega$

1 **for** $k = 1, \ldots, K$ **do**
2 $m(\{c_k\}) = 1/T \times \sum_{t=1}^{T} \underline{p}_{tk}$
3 $m(\Omega) = 1 - \sum_{k=1}^{K} m(\{c_k\})$
4 Sort classes by decreasing mass: $m(\{c_{(1)}\}) \geq m(\{c_{(2)}\}) \geq \cdots \geq m(\{c_{(K)}\})$
5 $A = \emptyset$
6 $bel = 0$
7 $mleu = 0$ // maximum lower expected utility
8 **for** $i = 1, \ldots, K$ **do**
9 $bel = bel + m(\{c_{(i)}\})$
10 $leu = d_\alpha(i) \times bel$ // lower expected utility
11 **if** $leu > mleu$ **then**
12 $mleu = leu$
13 $A = A \cup \{c_{(i)}\}$
14 **if** $leu > d_\alpha(i+1)$ **then**
15 break;
16 **return** A

Proposition 4.1. *Consider a subset $A \subseteq \Omega$ (typically, the current maximizer of the lower expected utility in the search procedure) such that we have $\underline{\mathbb{E}}(A) \geq d_\alpha(i)$, for some $i > |A|$, $i \in \{1, \ldots, K\}$; then, $\underline{\mathbb{E}}(A) \geq \underline{\mathbb{E}}(B)$ for any $B \subseteq \Omega$ with cardinality $|B| \geq i$.*[2]

Proof 4.3. *Let $A \subseteq \Omega$ be a subset of classes. Assume that $\underline{\mathbb{E}}(A) > d_\alpha(i)$ for some $i > |A|$. Since $Bel(B) \leq 1$ for any subset $B \subseteq \Omega$, then $d_\alpha(i)$ is an upper bound for the lower expected utility of any subset B such that $|B| = i$, and therefore $\underline{\mathbb{E}}(A) > \underline{\mathbb{E}}(B)$. The generalization to all subsets B such that $|B| > i$ comes from $d_\alpha(i)$ being monotone decreasing in i.*

Overall procedure

Our overall procedure for averaging imprecise classifier outputs, hereafter referred to as CDM/Ave, extends classical averaging for precise probabilities. It is summarized in Algorithm 1. Note that the stopping condition used in lines 14–15 proceeds from Proposition 4.1 above.

Note that a theorem similar to Theorem 4.1 was independently proven in [38], which addressed set-valued prediction in a probabilistic framework for a wide range of utility functions. Since the masses obtained by averaging the interval-valued classifier outputs are quasi-Bayesian, the procedure described in Algorithm 1 is close to that described in [38]. The overall complexity of Algorithm 1 is $O(K \log K)$ due to sorting the classes by decreasing mass.

Example 2 (Cautious decision-making via generalised averaging). *Assume the averaged mass function on $\Omega = \{c_1, c_2, c_3, c_4\}$ is given as follows:*

$$m(\{c_1\}) = 0.32,\ m(\{c_2\}) = 0.48,\ m(\{c_3\}) = 0.04,\ m(\{c_4\}) = 0.06,$$
$$m(\Omega) = 0.05.$$

The classes ordered by decreasing mass are thus $c_2 \succ c_1 \succ c_4 \succ c_3$. They are repeatedly aggregated to the candidate maximizer of the lower expected utility, which is re-computed (using d_{65}) each time a new class is added. The results are displayed in Table 2.

[2] Note that $\underline{\mathbb{E}}(A) > d_\alpha(i)$ for some $i > |A|$ implies $\underline{\mathbb{E}}(A) > \underline{\mathbb{E}}(B)$ for any $B \subseteq \Omega$.

Table 2: Intermediate results obtained using Alg. 1 on the mass function in Example 2.

$\|A\|$	A	$d_{65}(\|A\|)$	$Bel(A)$	$\underline{\mathbb{E}}_m(A,U)$	$> d_{65}(\|A\|+1)$?	Status
1	$\{c_2\}$	1	0.48	0.48	No (< 0.65)	Continue
2	$\{c_2, c_1\}$	0.65	0.8	0.52	Yes (> 0.467)	Stop

The cautious prediction made is $A^\star = \{c_2, c_1\}$, which reaches the maximum expected lower discounted utility since $\underline{\mathbb{E}}(\{c_1, c_2\}) > d_{65}(3) = 0.467$.

4.2. Generalization of voting

We now present our generalized voting strategy, which aggregates the imprecise outputs of trees into a single mass function, which is usually not quasi-Bayesian, via the interval dominance criterion. We then propose a decision-making procedure with a reasonable time complexity to select the best subset of classes as cautious predictions.

Classifier output aggregation

We now address the combination of probability intervals via voting. As above, we assume that each classifier output is a set of probability intervals $I_{tk}(\boldsymbol{x})$, for $k = 1, \ldots, K$.

The first step of our approach consists in identifying, for each classifier, the set A_t of non-dominated classes, using for instance the interval dominance criterion. The classifier outputs are then aggregated by computing the frequencies of all subsets $B \subseteq \Omega$ across the sets of non-dominated classes A_t, $t = 1, \ldots, T$:

$$m(B) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(A_t = B\right), \tag{14}$$

which is equivalent to letting each classifier vote for its set A_t of non-dominated classes. Algorithm 2 describes this approach to combining classifier outputs into a single mass function m, in time complexity $O(TK^2)$.

Algorithm 2: Tree aggregation via interval dominance

Input: Classifier outputs $I_{tk}(\boldsymbol{x}) = [\underline{p}_{tk}(\boldsymbol{x}); \overline{p}_{tk}(\boldsymbol{x})]_{t=1,\ldots,T}^{k=1,\ldots,K}$
Output: Mass function m

```
m(A) = 0, ∀A ⊆ Ω
for t = 1, ..., T do
    DC = ∅                                          // set of dominated classes
    for k = 1, ..., K do
        for j = 1, ..., K and j ≠ k do
            if p̄_tk < p_tj then
                DC = DC ∪ c_k
                break
    NDC = Ω \ DC                                    // non-dominated classes
    m(NDC) = m(NDC) + 1/T
return m
```

Decision-making

Our decision-making strategy consists again to compute the maximizer $A^\star$ of the lower expected discounted utility $\underline{\mathbb{E}}(A)$ over all subsets $A \subseteq \Omega$. However, the mass function m obtained by aggregating the classifier outputs does not have any specific property; in particular, it is neither quasi-Bayesian nor consonant. As a consequence, computing the maximizer $A^\star$ of the lower expected discounted utility requires in principle to check all candidate subsets $A \subseteq \Omega$, the worst-case complexity of which is exponential ($O(2^K)$) and therefore prohibitive for datasets with large numbers of classes. In order to alleviate this complexity, we leverage the following properties:

(i) we may arbitrarily restrict the decision to subsets $A \subseteq \Omega$ with cardinality $|A| \leq Q < K$, which reduces the complexity to $O(\sum_{k=1}^{Q} \binom{K}{k})$.

(ii) we scan subsets $A \subseteq \Omega$ of increasing cardinality: given a current maximizer A, we stop the procedure when larger subsets B such that $|B| > |A|$ are known not to increase the lower expected discounted utility as per Proposition 4.1;

(iii) in addition, during the search, for a given cardinality i, only subsets A composed of classes appearing in focal elements B such that $|B| \leq i$ need to be considered (see Proposition 4.2);

Proposition 4.2 shows that for a given cardinality i, we do not need to consider all subsets $A : |A| = i$ as candidate maximizers for the lower expected discounted utility, but only those composed of classes appearing in focal elements $B : |B| \leq i$.

Proposition 4.2. *Let Ω_i be the set of classes appearing in focal elements of cardinality less than or equal to i, for some $i \in \{1, \ldots, K\}$.*

The subset $A_i^\star \subseteq \Omega$ maximizing the lower expected utility among all A such that $|A| = i$ is a subset of Ω_i composed of classes appearing in focal elements B such that $|B| \leq i$.

Proof 4.4. *Assume a subset A of cardinality $|A| = i$ is such that $A = A_1 \cup A_2$, with $A_1 = A \cap \Omega_i$: then, $Bel(A) = Bel(A_1)$.*

If $A_2 \neq \emptyset$, then $\underline{\mathbb{E}}(A) < \underline{\mathbb{E}}(A_1)$ since $Bel(A) = Bel(A_1)$ and $|A| > |A_1|$: classes $c_k \notin \Omega_i$ necessarily decrease $\underline{\mathbb{E}}(A)$, and focal elements B such that $|B| > i$ do not contribute to $Bel(A)$.

Example 3. *Let $\Omega = \{c_1, c_2, c_3, c_4\}$ and m be a mass function defined by*

$$m(\{c_1\}) = 0.3, \qquad m(\{c_2\}) = 0.2, \qquad m(\{c_1, c_3\}) = 0.15,$$
$$m(\{c_2, c_3, c_4\}) = 0.25, \qquad m(\Omega) = 0.1;$$

the subset of classes appearing in focal elements B such that $|B| \leq 2$ is $\Omega_2 = \{c_1, c_2, c_3\}$: therefore, we get the following belief degrees for focal elements of cardinality $i = 2$ which are subsets of Ω_2:

$$Bel(\{c_1, c_2\}) = 0.5, \qquad Bel(\{c_1, c_3\}) = 0.45, \qquad Bel(\{c_2, c_3\}) = 0.2.$$

The maximizer of the lower expected discounted utility among subsets of cardinality $|A| = 2$ is thus $A_2^\star = \{c_1, c_2\}$.

The procedure described in Algorithm 3, to which we refer in the following as CDM/Vote, extends voting when votes are expressed as subsets of classes and returns the subset $A^\star = \arg\max \underline{\mathbb{E}}(A)$ among all subsets $A \subseteq \Omega$ such that $|A| \leq Q \leq K$. It generalizes the method proposed in the previous paper [17, 16] for binary cautious classification, which amounts to maximizing the discounted accuracy dr_{acc} when $\Omega = \{c_1, c_2\}$. Note that CDM/Vote is computationally less efficient than CDM/Ave by design, even if the time complexity can be controlled. However, as it is shown in the experimental part, this approach remains able to address cautious classification problems with large numbers of classes.

Algorithm 3: Cautious Decision Making by Voting

Input: Mass function m (typically obtained by Algorithm 2); cardinality bound Q; discount ratio d_α

Output: Decision A

$FE = \emptyset$ // focal elements
$\Omega_i = \emptyset$ // considering classes
$A = \emptyset$
$mleu = 0$ // maximum lower expected utility
for $i = 1, \ldots, Q$ // trick 1
do
  $dr = d_\alpha(i)$
  if $mleu > dr$ **then**
    Return A // trick 2 (Prop. 4.1)
  else
    $FE = FE \cup \{B : m(B) > 0, |B| = i, B \subseteq \Omega\}$
    $\Omega_i = \Omega_i \cup \{c : c \in B, B \in FE\}$ // trick 3 (Prop. 4.2)
    for *all* $B \subseteq \Omega_i$ *and* $|B| = i$ **do**
      $bel = \sum_{C \in FE,\ C \subseteq B} m(C)$
      $leu = dr \times bel$ // lower expected utility for B
      if $leu > mleu$ **then**
        $mleu = leu$
        $A = B$
return A

Example 4 (Cautious decision-making by generalised voting). *Let $\Omega = \{c_1, c_2, c_3, c_4\}$, and let the mass function m obtained via Algorithm 2 be defined by*

$$m(\{c_1\}) = 0.15, \quad m(\{c_2\}) = 0.25, \quad m(\{c_1, c_2\}) = 0.35,$$
$$m(\{c_1, c_3\}) = 0.05, \quad m(\{c_2, c_3\}) = 0.1, \quad m(\{c_2, c_3, c_4\}) = 0.05,$$
$$m(\Omega) = 0.05.$$

Let us apply Algorithm 3 to make a decision, using d_{65} as the discount ratio. The first iteration considers candidate maximizers of the lower expected discounted utility of cardinality $i = 1$. We have $A_1^\star = \{c_2\}$ as current maximizer for the lower expected utility: since $\underline{\mathbb{E}}(\{c_2\}) < d_\alpha(2)$, the process continues. At iteration $i = 2$, the process determines $A_2^\star = \{c_1, c_2\}$ as current maximizer for the lower expected utility, with $\underline{\mathbb{E}}(\{c_1, c_2\}) > d_\alpha(3)$: the search process is therefore stopped.
The final set-valued prediction is then $A^\star = \{c_1, c_2\}$ since any subset $B \subseteq \Omega$ with $|B| > 2$ has a smaller lower expected utility than $A^\star$. Remark that class c_4 has never been considered in the process since it would have appeared in iteration $i = 3$ as a component of the focal element $\{c_2, c_3, c_4\}$: however, focal elements with cardinality $i = 3$ are not considered, since they are detected as unable to ameliorate the lower expected utility.

5. Experiments and results

In this section, we detail the experiments conducted using random forests as an ensemble classification method. We report the results obtained on 15 datasets of various sizes and numbers of classes. These datasets, which are described in Table 4, were collected from the UCI repository [39] and the Kaggle website [40]. In all the following experiments, we used the scikit-learn implementation of random forests [41] as the base ensemble learning model.

Table 3: Intermediate results obtained using Alg. 3 on the mass function in Example 4.

Iteration $i = 1$: subset of considered classes $\Omega_1 = \{c_1, c_2\}$					
A	$d_{65}(\lvert A\rvert)$	$Bel(A)$	$\underline{\mathbb{E}}_m(A, U)$	$> d_{65}(\lvert A\rvert + 1)$?	Status
$\{c_1\}$	1	0.15	0.15	No (< 0.65)	
$\{c_2\}$	1	0.25	0.25	No (< 0.65)	Continue
Iteration $i = 2$: subset of considered classes $\Omega_2 = \{c_1, c_2, c_3\}$					
A	$d_{65}(\lvert A\rvert)$	$Bel(A)$	$\underline{\mathbb{E}}_m(A, U)$	$> d_{65}(\lvert A\rvert + 1)$?	Status
$\{c_1, c_3\}$	0.65	0.20	0.13	No (< 0.4667)	
$\{c_2, c_3\}$	0.65	0.35	0.2275	No (< 0.4667)	
$\{c_1, c_2\}$	0.65	0.75	0.4875	Yes (> 0.4667)	Stop

Table 4: Brief description of the datasets used: numbers of instances, features, and classes.

	Datasets	Nb. instances	Nb. features	Nb. classes
1	Balance-scale	625	4	3
2	Ecoli	336	7	8
3	Forest	523	27	4
4	Glass	214	9	6
5	Letter	20000	16	26
6	Libras	360	90	15
7	Optdigits	5620	64	10
8	Page-blocks	5473	10	5
9	Seeds	210	7	3
10	Spectrometer	531	101	48
11	Vehicle	846	18	4
12	Vowel	990	10	11
13	Waveform	5000	40	3
14	Wine-quality	1599	11	6
15	Yeast	1484	8	10

5.1. Generalized voting efficiency

As mentioned above, a limitation of our generalized voting strategy is its computational complexity. Therefore, the first experiment studies the time complexity of the CDM/Vote approach as a function of the number of labels.

Experimental setting

For a given number of labels i, we first picked i labels at random and extracted the corresponding samples from the original dataset to construct a dataset with i labels. We then divided this dataset into a training set (containing 80% of the instances) and a test set (20% of the instances). A random forest consisting of 100 trees was trained with the parameter `min_samples_leaf` set to one. The IDM parameter was set to $s = 1$ in this experiment.

During the test phase, we applied the CDM/Vote approach using the d_{65} discounted ratio to make predictions for the test data. For each sample, we recorded the elapsed time of the entire decision-making procedure (aggregation of classifier outputs via interval dominance, followed by the lower expected discounted utility maximization step), as well as the elapsed time for the second step only (lower expected utility maximization), respectively referred to as ID+MLEDU and MLEDU. In particular, since for high values of i, the decision-making would be intractable without any control of the complexity, we always set the allowed

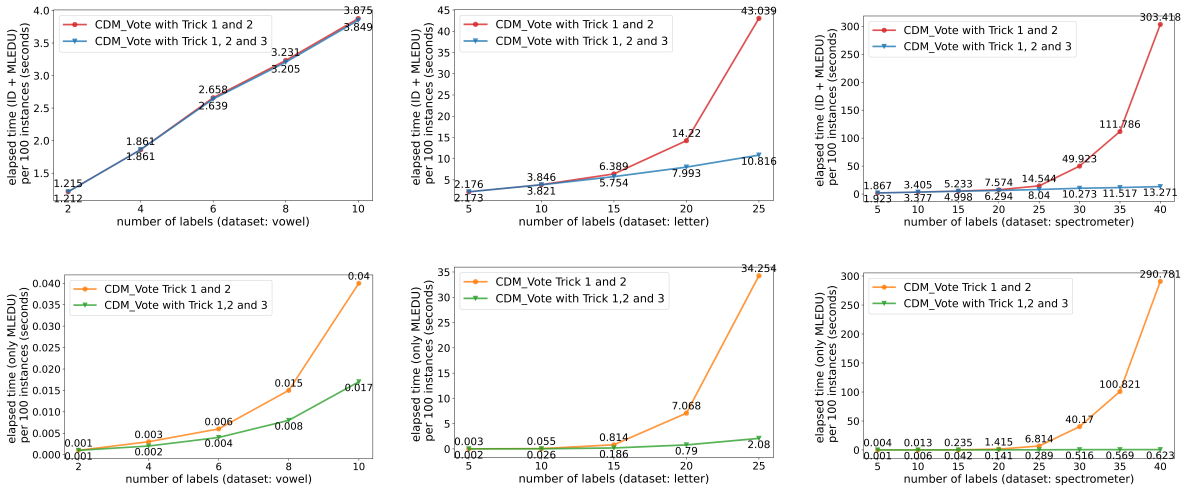


Figure 2: Decision-making time complexity of CDM/Vote according to the number of labels (for 100 samples), with (up) and without (down) filtering as per Proposition 4.2. Left to right: *vowel*, *letter*, and *spectrometer*.

maximum labels in the prediction to $Q = 5$, and stopped the procedure when a larger prediction cardinality was known not to increase the utility: thus, we compared the efficiency with and without filtering the focal elements based on Proposition 4.2.

For each given number of labels i, we repeated the process of dataset construction, data division, model training, and decision-making 10 times. Finally, the average elapsed time per 100 inferences was reported.

Results and dicussion

Fig. 2 shows that for a small number of labels (less than 15), filtering out subsets $A \not\subseteq \Omega_i$ does not significantly improve the efficiency, as applying interval dominance prevails over computing the subset maximizing the lower expected discounted utility in terms of computational time. However, for a larger number of labels, this latter step becomes prominent: filtering out subsets $A \not\subseteq \Omega_i$ substantially accelerates the procedure, regardless of the number of labels, as shown in the right column of Fig. 2. This experiment demonstrates that CDM/Vote remains applicable with a large number of labels.

5.2. Reducing the risk of making wrong decisions

Cautious classification aims at producing set-valued predictions for instances whose actual class is difficult to accurately identify, so as to decrease the risk of missing it. This experiment studies the behavior of our approach when facing difficult test instances. In precise classification, the level of difficulty can be measured by the margin between the highest class posterior probability and the second highest [42]:

$$\mu(x) = \mathbb{P}(c_{i_1}|\boldsymbol{x}) - \mathbb{P}(c_{i_2}|\boldsymbol{x}), \tag{15}$$

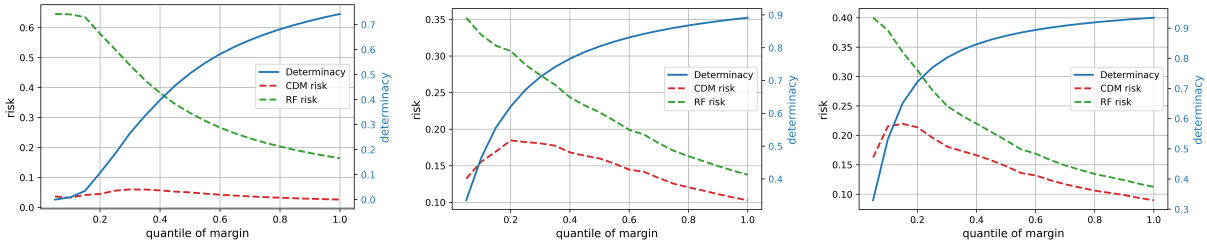
where

$$i_1 = \arg\max_j \mathbb{P}(c_j|\boldsymbol{x}), \quad i_2 = \arg\max_{j \neq i_1} \mathbb{P}(c_j|\boldsymbol{x}).$$

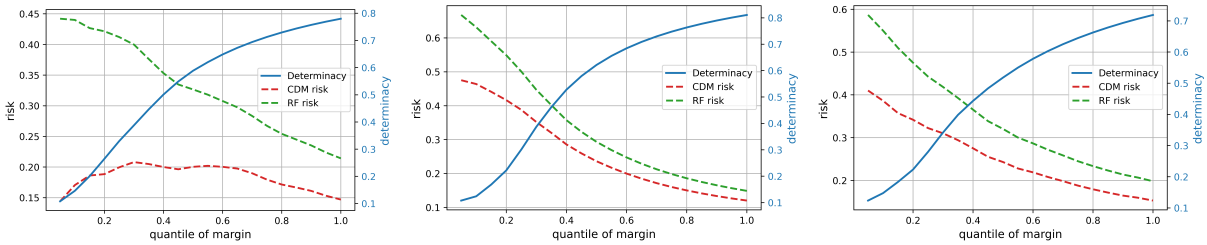
This margin can be considered as an indicator of the aleatoric uncertainty pertaining to $\boldsymbol{x}$: the smaller $\mu(x)$, the harder it is to correctly classify $\boldsymbol{x}$.

Experimental setting

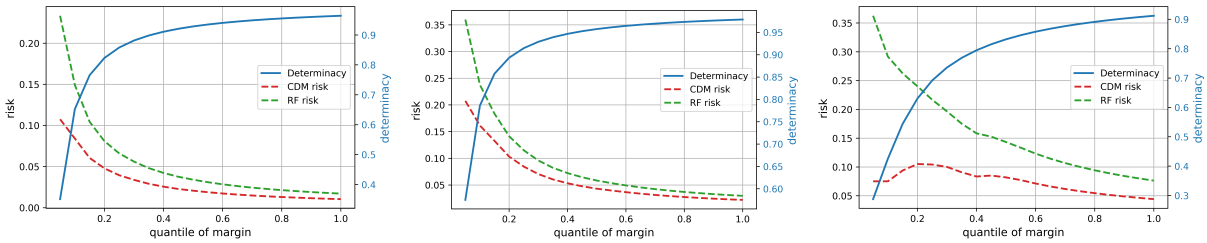
In this experiment, we repeated the decision-making procedure via two times 5-fold cross-validation, and the average results over all folds are reported.



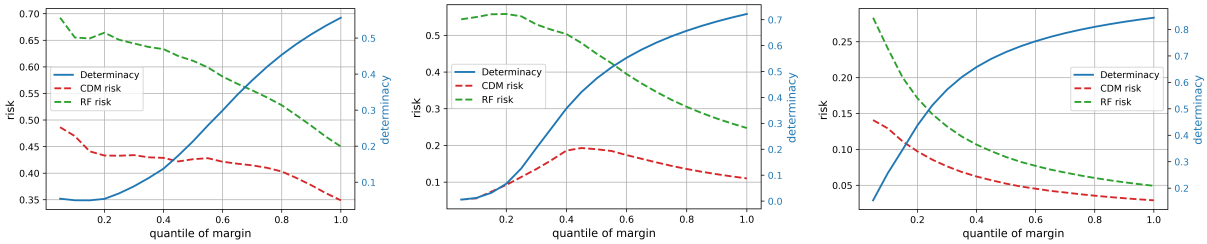
(a) balance scale (b) ecoli (c) forest



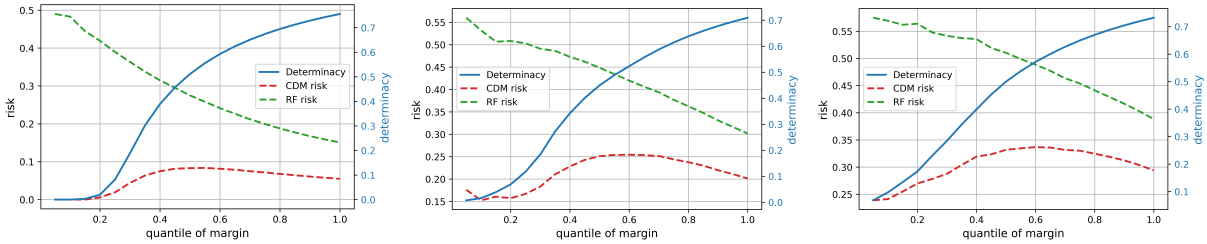
(d) glass (e) letter (f) libras



(g) optdigits (h) page blocks (i) seeds



(j) spectrometer (k) vehicle (l) vowel



(m) waveform (n) wine quality (o) yeast

Figure 3: Performances of the cautious and classical random forests as a function of instance difficulty.

In each fold, a random forest consisting of 100 trees was trained on training data, with the parameter min_samples_leaf set to one, and the IDM parameter set to $s = 2$. Then, in each fold, we computed the quantile of level α of obtained margins over the validation samples, noted as μ_α, and we selected the instances with a margin $\mu(x) < \mu_\alpha$ as test samples. The value of α directly quantifies the level of difficulty of these selected instances, in terms of (accurate) classification. For these selected instances, we evaluated the determinacy and the risk of the proposed cautious decision-making strategy, and compared them to those of a conventional random forest. Note that the risk of our cautious decision-making approach (CDM/Vote using the d_{65} discounted ratio) is the proportion of predictions that do not contain true labels, i.e., it is the set accuracy computed over all test instances. The risk of the precise random forest used as a benchmark is its error rate.

Results and discussion

In Fig. 3, for all datasets, we can first observe what was expected: as α increases, the proportion of difficult instances decreases and determinacy consequently increases. The second observation is that the risk of the cautious classifier is significantly lower than the one of the conventional random forest. The superiority of our method in terms of risk is even more notable when the proportion of difficult samples is larger (when α is low).

It is important to note that in the above experiments, we investigated the performance of our approach in the presence of aleatoric uncertainty. However, we also found that our approach does not present the same behavior when facing epistemic uncertainty, i.e., there is no guarantee that an increase in epistemic uncertainty (via an adjustment of the IDM parameter s) will result in an increase in terms of cautiousness.

5.3. Performance comparison of cautious classifiers

In this experiment, we compared our two proposed approaches (CDM/Ave, presented in Section 4.1, which generalizes averaging, and CDM/Vote, presented in Section 4.2, which generalizes voting) with the following cautious classification approaches:

- Constant Risk (CR) consists in fixing an acceptable risk threshold r (usually by cross-validation or by expert users) and selecting the smallest number of best classes with cumulative probability exceeding $1 - r$;
- the NonDeterministic Classifier (NDC) aims to maximize the discounted utility $F_\beta = \frac{1+\beta^2}{\beta^2+|A|} \cdot \sum_{c_k \in A} p(c_k|\boldsymbol{x})$, where the value of β is often taken as one;
- AVEraging (AVE) computes the average probability intervals provided by the trees and applies interval dominance to make cautious predictions, following [43] and [14];
- Minimum Vote Against (MVA) counts the number of classifiers that predict a class as dominated (vote against), and outputs the set of classes with the lowest amount of votes against as final decision [15].

All of these methods can be divided into two groups. While CR and NDC are two cautious classifiers based on precise estimation of the class posterior probabilities (after aggregation of precise decision trees), the four others are based on the aggregation of imprecise classifier outputs.

Experimental setting

In this experiment, the number of trees in the forest was also set to 100, and the parameter min_samples_leaf was set to one. The CDM/Vote and CDM/Ave procedures used the d_{65} discounted ratio to make decisions.

We first compared these models with the original data. For each dataset and each compared model, we conducted a procedure of two times 5-fold cross-validation to evaluate the determinacy, single-set accuracy, set accuracy computed over indeterminate predictions only, set size, u_{65} score and u_{80} score. In each fold, we set the parameter value $\beta = 1$ for NDC, and for the other five methods, a nested 5-fold cross-validation was conducted to choose the best value for the parameter s within the set $\{0.5, 1, 1.5, 2, 2.5\}$ by maximizing the u_{65} score. It should be noted that the same decision trees were used in all of these six approaches.

Table 5: Average results across all datasets

Results on data without label noise						
Criteria	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Determinacy	0.767	0.775	0.833	**0.987**	0.787	0.796
Single-set accuracy	0.891	**0.894**	0.875	0.822	0.890	0.885
Set accuracy	0.873	0.873	**0.904**	0.802	0.881	0.878
Set size	2.115	2.054	**3.544**	2.042	2.135	2.116
u_{65} score	0.813	**0.824**	0.815	0.818	0.822	0.821
u_{80} score	0.842	**0.852**	0.834	0.819	0.849	0.847
Results on data with 30% label noise						
Criteria	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Determinacy	0.524	0.632	0.655	**0.976**	0.658	0.680
Single-set accuracy	**0.874**	0.847	0.855	0.745	0.850	0.847
Set accuracy	0.849	0.838	**0.934**	0.779	0.863	0.841
Set size	2.239	2.100	**5.463**	2.058	2.417	2.192
u_{65} score	0.705	0.735	0.710	0.740	0.738	**0.748**
u_{80} score	0.763	0.780	0.746	0.742	0.780	**0.786**

Then, we also studied the behavior of the classifiers by introducing label noise. We followed the same procedure mentioned above in each cross-evaluation fold: 30% of training samples were randomly selected and their labels were replaced with a randomly selected label different from the actual one. The labels of the test instances were left unchanged.

Results and discussion

Table 5 shows the evaluation of each metric averaged across all datasets to show the overall performance of the different models. The detailed evaluation of each dataset can be found in Appendix A, from Table A.6 to Table A.11. Below, we discuss the results obtained on data without and with label noise.

1. Noise-free data:
   - *Determinacy and single-set accuracy*: On data without label noise, the MVA method almost always produces determinate predictions, but this confidence comes at the cost of a significant drop in single-set accuracy. Conversely, the CR approach is too cautious, resulting in a low determinacy. The NDC and our proposed CDM methods have a lower determinacy and comparable accuracy.
   - *Set accuracy and set Size*: Regarding indeterminate predictions, the AVE method frequently achieves a higher set accuracy but at the expense of a much larger number of labels in predictions, especially for several datasets such as "letters", "libras", and "spectrometer". In contrast, other cautious classifiers, including the NDC and CDM models, manage to keep the average size of indeterminate predictions reasonable (less than three labels in predictions).
   - *Utility Scores (u_{65} and u_{80})*: On the original datasets, NDC shows the best performance, but CDM/Ave and CDM/Vote achieve very close results. This observation demonstrates that these three approaches can achieve a better balance between accuracy and cautiousness.
2. Data with 30% label noise:
   - *Determinacy and single-set accuracy*: When facing label noise in training data, all classifiers except MVA show a reduced determinacy, which indicates more cautious decisions. The drop of CR in terms of determinacy is significant, due to lower gaps between the class posterior probabilities. Compared to NDC and AVE, our proposed CDM models can return more determinate predictions while keeping a very similar accuracy, which means our models are more efficient in capturing samples that are difficult to classify.

- *Set accuracy and set size*: In the presence of label noise, the AVE method still performs best in set accuracy but with very imprecise predictions. Other cautious classifiers, such as NDC and CDM, maintain reasonable set sizes and good set accuracy despite the noise. Meanwhile, we find that with the introduction of noise, the models tend to provide indeterminate predictions of larger size, except for MVA, which keeps two labels in the indeterminate predictions.
- *Utility Scores (u_{65} and u_{80})*: On noisy datasets, CDM/Ave and CDM/Vote show higher discounted utility values. CDM/Vote, in particular, achieves the best u_{65} and u_{80} scores on 9 and 10 out of 15 datasets, respectively, outperforming all other methods. This can be explained by its ability to deal with uncertainty in the decision-making process and to keep a better balance between cautiousness and accuracy.

In summary, our cautious decision-making approaches, especially CDM/Vote, exhibit a sensitivity to label noise in training data, and are able to achieve a good compromise between model accuracy and cautiousness, in particular when the data are pervaded with noise.

6. Conclusion

In this paper, we have proposed two aggregation strategies to make cautious decisions in the case of multi-class classification problems. In this setting, we consider ensembles of classifiers that provide intervals of posterior probabilities as outputs, such as those provided by the imprecise Dirichlet model for classifiers based on sample counts (like, e.g., decision trees).

Our two strategies respectively generalize averaging and voting for classical tree ensembles. In both cases, they aim at making decisions by computing the subset of classes which maximizes the lower expected utility over all possible subsets of classes. Our generalized averaging approach is computationally more efficient than our generalized voting strategy, the complexity of which can nevertheless be controlled—by leveraging two theoretically supported tricks that avoid scanning all candidate subsets of classes, and by restricting the cardinality of the set-valued predictions.

The experiments conducted on different datasets illustrate the interest of our proposals in order to achieve a good compromise between model accuracy and determinacy, especially for difficult instances. This is especially the case when the data are pervaded with label noise, in which case the performances of our proposals compare very favorably to that of the other strategies used as benchmarks. The experiments also confirm that our cautious decision-making procedure is able to process datasets with a large number of classes in a limited computational complexity.

In the future, we may further investigate how to efficiently calculate the upper expected utility for the CDM methods. Thus, the Hurwicz expected utility can be applied to make the model more flexible to adjust the cautiousness. Moreover, for a better compromise between the model accuracy and cautiousness, the weight assignment for imprecise trees in multi-class cases is also an interesting study direction. Finally, we may explore tree-based ensembles that can adapt to epistemic uncertainty in data space (our current approach primarily captures aleatory uncertainty), for example using generative tree models.

Declaration of special issue

This paper extends an earlier conference version published in ECSQARU 2023, where the aggregation approaches were first introduced and tested on three datasets [44]. The current version includes more detailed theorems, propositions, and proofs, highlights how our model reduces the risk of wrong decisions, and provides a more comprehensive evaluation of the proposed approaches, comparing them with more cautious classifiers on 15 datasets.

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 datasets used in experiments are accessible on https://github.com/Haifei-ZHANG/Cautious-Random-Forest.

Acknowledgements

This work was mainly conducted while the first author was a research and teaching assistant at the Université de Technologie de Compiègne, and we thank the Heudiasyc laboratory for supporting this work. We also thank the Laboratory of Hubert Curien at the Université Jean Monnet for supporting the first author during the revision of the article.

Appendix A. Comparison results of cautious classifiers

Table A.6: Comparison of determinacy

Results on data without label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.871±0.059	0.807±0.028	0.844±0.024	**0.980±0.009**	0.778±0.030	0.787±0.022
Ecoli	0.844±0.081	0.833±0.052	0.912±0.023	**0.994±0.007**	0.844±0.043	0.850±0.048
Forest	0.930±0.042	0.906±0.032	0.954±0.027	**0.999±0.003**	0.912±0.030	0.920±0.030
Glass	0.771±0.069	0.745±0.059	0.829±0.029	**0.993±0.011**	0.762±0.065	0.771±0.063
Letter	0.662±0.018	0.792±0.019	0.761±0.025	**0.978±0.010**	0.798±0.018	0.799±0.019
Libras	0.628±0.045	0.703±0.060	0.764±0.029	**0.976±0.020**	0.711±0.053	0.731±0.049
Optdigits	0.941±0.009	0.956±0.008	0.976±0.005	**0.998±0.001**	0.959±0.007	0.959±0.007
Page-blocks	0.982±0.013	0.975±0.010	0.984±0.007	**1.000±0.001**	0.978±0.009	0.979±0.007
Seeds	0.952±0.054	0.919±0.047	0.962±0.036	**0.998±0.007**	0.917±0.049	0.924±0.041
Spectrometer	0.370±0.032	0.476±0.049	0.541±0.036	**0.954±0.020**	0.502±0.043	0.519±0.039
Vehicle	0.636±0.083	0.683±0.027	0.733±0.023	**0.987±0.008**	0.703±0.028	0.710±0.028
Vowel	0.804±0.025	0.834±0.028	0.888±0.027	**0.992±0.007**	0.831±0.024	0.834±0.023
Waveform	0.845±0.037	0.709±0.010	0.853±0.027	**0.990±0.004**	0.741±0.014	0.742±0.016
Wine-quality	0.664±0.105	0.664±0.027	0.800±0.048	**0.989±0.003**	0.691±0.023	0.702±0.017
Yeast	0.610±0.078	0.623±0.032	0.687±0.026	**0.980±0.007**	0.682±0.037	0.720±0.047
Average	0.767	0.775	0.833	**0.987**	0.787	0.796
#Best	0	0	0	**15**	0	0
Results on data with 30% label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.540±0.140	0.632±0.045	0.646±0.062	**0.982±0.010**	0.563±0.054	0.595±0.043
Ecoli	0.656±0.073	0.726±0.053	0.744±0.039	**0.979±0.019**	0.781±0.046	0.845±0.040
Forest	0.737±0.052	0.722±0.054	0.823±0.043	**0.988±0.012**	0.753±0.048	0.746±0.049
Glass	0.448±0.097	0.540±0.066	0.593±0.059	**0.979±0.025**	0.584±0.080	0.635±0.088
Letter	0.416±0.032	0.625±0.019	0.514±0.029	**0.957±0.012**	0.680±0.022	0.732±0.020
Libras	0.376±0.066	0.622±0.049	0.582±0.039	**0.976±0.015**	0.635±0.037	0.625±0.044
Optdigits	0.675±0.011	0.895±0.009	0.876±0.011	**0.995±0.002**	0.922±0.010	0.933±0.009
Page-blocks	0.846±0.020	0.821±0.022	0.878±0.015	**0.995±0.002**	0.873±0.016	0.887±0.022
Seeds	0.586±0.190	0.702±0.074	0.712±0.138	**0.990±0.012**	0.664±0.126	0.702±0.074
Spectrometer	0.155±0.028	0.436±0.036	0.326±0.037	**0.934±0.027**	0.476±0.046	0.481±0.057
Vehicle	0.505±0.117	0.516±0.044	0.638±0.068	**0.971±0.015**	0.547±0.047	0.526±0.053
Vowel	0.470±0.035	0.628±0.023	0.617±0.034	**0.976±0.009**	0.666±0.030	0.694±0.033
Waveform	0.695±0.009	0.509±0.013	0.784±0.010	**0.980±0.004**	0.524±0.011	0.509±0.013
Wine-quality	0.425±0.102	0.582±0.028	0.641±0.031	**0.973±0.009**	0.606±0.031	0.630±0.042
Yeast	0.327±0.028	0.529±0.027	0.455±0.026	**0.971±0.007**	0.591±0.026	0.661±0.025
Average	0.524	0.632	0.655	**0.976**	0.658	0.680
#Best	0	0	0	**15**	0	0

Table A.7: Comparison of single-set accuracy

Results on data without label noise

Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.881±0.047	0.930±0.025	0.930±0.020	0.863±0.024	**0.960±0.017**	0.956±0.019
Ecoli	0.896±0.050	0.914±0.036	0.891±0.031	0.867±0.036	**0.915±0.034**	0.911±0.039
Forest	0.909±0.018	**0.918±0.017**	0.902±0.043	0.889±0.020	0.916±0.015	0.915±0.015
Glass	0.836±0.083	**0.847±0.081**	0.817±0.045	0.783±0.067	0.844±0.084	0.833±0.090
Letter	**0.984±0.008**	0.948±0.015	0.959±0.012	0.852±0.019	0.940±0.018	0.937±0.015
Libras	**0.936±0.017**	0.935±0.038	0.885±0.070	0.797±0.032	0.922±0.031	0.916±0.027
Optdigits	**0.994±0.002**	0.993±0.002	0.990±0.003	0.980±0.005	0.992±0.002	0.991±0.003
Page-blocks	0.976±0.007	**0.978±0.009**	0.976±0.008	0.969±0.010	0.977±0.009	0.977±0.008
Seeds	0.943±0.029	**0.956±0.038**	0.938±0.030	0.931±0.047	0.956±0.038	0.956±0.037
Spectrometer	**0.741±0.040**	0.686±0.076	0.673±0.040	0.544±0.054	0.677±0.063	0.664±0.066
Vehicle	**0.890±0.048**	0.878±0.027	0.849±0.027	0.748±0.030	0.876±0.026	0.864±0.022
Vowel	0.990±0.006	**0.991±0.007**	0.977±0.014	0.941±0.017	0.987±0.008	0.984±0.008
Waveform	0.895±0.013	**0.939±0.008**	0.897±0.013	0.852±0.010	0.930±0.009	0.929±0.008
Wine-quality	0.786±0.041	**0.788±0.024**	0.743±0.025	0.691±0.018	0.777±0.026	0.763±0.028
Yeast	**0.712±0.038**	0.707±0.029	0.692±0.021	0.621±0.029	0.689±0.025	0.681±0.026
Average	0.891	**0.894**	0.875	0.822	0.891	0.885
#Best	**6**	5	0	0	3	1

Results on data with 30% label noise

Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.871±0.060	0.873±0.045	0.876±0.052	0.751±0.034	**0.908±0.055**	0.887±0.046
Ecoli	**0.905±0.034**	0.896±0.031	0.896±0.034	0.811±0.045	0.897±0.030	0.877±0.034
Forest	0.908±0.029	**0.910±0.022**	0.900±0.024	0.841±0.028	0.909±0.026	0.906±0.025
Glass	**0.828±0.120**	0.801±0.085	0.787±0.076	0.666±0.071	0.796±0.086	0.787±0.074
Letter	0.945±0.018	0.889±0.025	**0.954±0.020**	0.752±0.020	0.894±0.027	0.890±0.019
Libras	**0.816±0.096**	0.768±0.067	0.797±0.069	0.654±0.067	0.766±0.074	0.771±0.055
Optdigits	**0.999±0.001**	0.992±0.003	0.996±0.002	0.972±0.004	0.992±0.002	0.991±0.003
Page-blocks	0.957±0.008	0.967±0.010	0.967±0.012	0.916±0.012	0.971±0.012	**0.972±0.012**
Seeds	**0.923±0.055**	0.892±0.077	0.900±0.081	0.810±0.071	0.907±0.072	0.895±0.077
Spectrometer	**0.849±0.118**	0.680±0.072	0.754±0.070	0.523±0.031	0.662±0.067	0.675±0.055
Vehicle	0.835±0.067	0.844±0.040	0.807±0.048	0.699±0.042	0.844±0.044	**0.845±0.047**
Vowel	**0.949±0.018**	0.923±0.014	0.935±0.018	0.797±0.028	0.927±0.010	0.923±0.015
Waveform	0.898±0.009	0.943±0.008	0.890±0.008	0.831±0.006	**0.946±0.008**	0.943±0.008
Wine-quality	**0.707±0.045**	0.665±0.029	0.671±0.032	0.594±0.017	0.669±0.035	0.688±0.028
Yeast	**0.727±0.056**	0.659±0.045	0.696±0.041	0.562±0.027	0.667±0.045	0.657±0.034
Average	**0.874**	0.847	0.855	0.745	0.850	0.847
#Best	**10**	1	0	0	2	2

Table A.8: Comparison of set accuracy

Results on data without label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.727±0.077	0.676±0.086	0.713±0.079	0.520±0.409	**0.867±0.085**	0.838±0.101
Ecoli	0.924±0.128	0.920±0.094	0.932±0.087	**1.000±0.000**	0.905±0.090	0.921±0.078
Forest	0.945±0.088	0.939±0.053	0.937±0.076	**1.000±0.000**	0.946±0.048	0.952±0.056
Glass	0.877±0.079	0.899±0.083	**0.932±0.090**	0.331±0.471	0.882±0.096	0.898±0.094
Letter	0.809±0.035	0.735±0.052	**0.933±0.020**	0.591±0.186	0.729±0.054	0.697±0.041
Libras	0.832±0.074	0.818±0.093	0.865±0.103	**0.882±0.165**	0.798±0.115	0.763±0.118
Optdigits	0.944±0.027	0.943±0.026	0.938±0.044	0.926±0.167	0.940±0.036	**0.954±0.025**
Page-blocks	0.901±0.165	0.941±0.081	0.871±0.126	**1.000±0.000**	0.932±0.147	0.953±0.147
Seeds	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
Spectrometer	0.683±0.049	0.689±0.071	**0.795±0.060**	0.469±0.188	0.654±0.075	0.661±0.083
Vehicle	0.890±0.046	0.907±0.045	**0.956±0.028**	0.818±0.214	0.924±0.041	0.916±0.036
Vowel	0.961±0.020	0.963±0.026	0.964±0.039	0.934±0.165	**0.967±0.033**	0.954±0.043
Waveform	0.999±0.003	0.999±0.001	0.999±0.002	**1.000±0.000**	**1.000±0.000**	**1.000±0.001**
Wine-quality	0.836±0.035	0.866±0.031	**0.886±0.050**	0.777±0.278	0.868±0.031	0.875±0.028
Yeast	0.768±0.050	0.795±0.034	**0.841±0.030**	0.776±0.204	0.801±0.041	0.785±0.030
Average	0.873	0.873	**0.904**	0.802	0.881	0.878
#Best	1	1	**8**	5	4	2

Results on data with 30% label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.835±0.076	0.815±0.047	0.898±0.048	0.826±0.173	**0.965±0.029**	0.875±0.094
Ecoli	0.818±0.058	0.853±0.068	**0.913±0.047**	0.643±0.416	0.844±0.071	0.875±0.097
Forest	0.887±0.072	0.931±0.047	0.951±0.039	0.847±0.346	**0.965±0.023**	0.951±0.040
Glass	0.818±0.073	0.823±0.089	0.885±0.058	**0.890±0.400**	0.860±0.099	0.796±0.113
Letter	0.823±0.020	0.720±0.021	**0.977±0.013**	0.541±0.113	0.704±0.026	0.666±0.028
Libras	0.775±0.047	0.713±0.084	**0.894±0.057**	0.471±0.412	0.700±0.087	0.707±0.087
Optdigits	0.979±0.009	0.947±0.028	**0.990±0.007**	0.885±0.111	0.951±0.023	0.931±0.032
Page-blocks	0.895±0.033	0.920±0.025	0.963±0.023	0.950±0.150	**0.969±0.021**	0.967±0.022
Seeds	0.914±0.097	0.928±0.080	0.967±0.051	**1.000±0.000**	0.993±0.043	0.936±0.083
Spectrometer	0.739±0.053	0.638±0.058	**0.913±0.036**	0.586±0.275	0.633±0.057	0.610±0.072
Vehicle	0.854±0.065	0.890±0.045	**0.938±0.020**	0.776±0.172	0.923±0.025	0.911±0.046
Vowel	0.861±0.041	0.841±0.054	**0.949±0.026**	0.792±0.215	0.847±0.050	0.842±0.049
Waveform	0.983±0.007	0.992±0.003	0.998±0.002	0.971±0.044	**1.000±0.000**	0.992±0.003
Wine-quality	0.805±0.027	0.823±0.028	0.878±0.024	**0.886±0.125**	0.833±0.030	0.813±0.030
Yeast	0.747±0.028	0.739±0.028	**0.896±0.026**	0.621±0.154	0.754±0.040	0.744±0.031
Average	0.849	0.838	**0.934**	0.779	0.863	0.841
#Best	1	0	**7**	3	4	0

Table A.9: Comparison of set size

Results on data without label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	2.000±0.000	2.037±0.037	2.292±0.135	2.000±0.000	**2.809±0.116**	2.752±0.130
Ecoli	2.067±0.067	2.045±0.069	**2.339±0.261**	2.000±0.000	2.019±0.029	2.010±0.033
Forest	2.000±0.000	2.000±0.000	2.063±0.072	2.000±0.000	**2.153±0.186**	2.132±0.182
Glass	2.020±0.039	2.037±0.070	**2.260±0.198**	2.000±0.000	2.049±0.132	2.051±0.088
Letter	2.828±0.090	2.202±0.053	**12.493±1.261**	2.182±0.106	2.176±0.053	2.181±0.068
Libras	2.205±0.064	2.145±0.102	**3.382±0.526**	2.294±0.544	2.135±0.091	2.088±0.101
Optdigits	2.057±0.029	2.047±0.025	**2.542±0.142**	2.037±0.105	2.041±0.028	2.050±0.055
Page-blocks	2.000±0.000	2.000±0.000	**2.129±0.100**	2.000±0.000	2.034±0.082	2.082±0.147
Seeds	**2.000±0.000**	**2.000±0.000**	**2.000±0.000**	**2.000±0.000**	**2.000±0.000**	**2.000±0.000**
Spectrometer	2.410±0.120	2.140±0.043	**9.995±2.307**	2.102±0.194	2.134±0.058	2.112±0.040
Vehicle	2.024±0.023	2.030±0.018	**2.348±0.136**	2.000±0.000	2.223±0.119	2.163±0.075
Vowel	2.031±0.021	2.040±0.037	**2.421±0.185**	2.000±0.000	2.042±0.031	2.040±0.029
Waveform	2.000±0.000	2.000±0.001	2.009±0.005	2.000±0.000	**2.032±0.015**	2.023±0.021
Wine-quality	2.004±0.005	2.021±0.013	**2.139±0.054**	2.000±0.000	2.015±0.011	2.016±0.014
Yeast	2.074±0.036	2.070±0.015	**2.747±0.183**	2.017±0.030	2.165±0.241	2.042±0.013
Average	2.115	2.054	3.544	**2.042**	2.135	2.116
#Best	5	3	1	**12**	3	2

Results on data with 30% label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	2.012±0.024	2.030±0.024	2.339±0.073	2.000±0.000	**2.784±0.147**	2.318±0.412
Ecoli	2.073±0.044	2.103±0.079	**3.541±0.350**	2.000±0.000	2.401±0.364	2.067±0.086
Forest	2.000±0.000	2.089±0.074	**2.432±0.178**	2.000±0.000	2.426±0.274	**2.432±0.392**
Glass	2.080±0.092	2.096±0.081	**3.069±0.286**	2.112±0.133	2.667±0.784	2.096±0.072
Letter	3.107±0.127	2.219±0.051	**15.273±1.008**	2.163±0.136	2.331±0.163	2.240±0.053
Libras	2.461±0.071	2.165±0.101	**5.990±0.779**	2.176±0.624	2.262±0.410	2.163±0.109
Optdigits	2.247±0.019	2.112±0.042	**4.755±0.195**	2.033±0.064	2.434±0.209	2.116±0.041
Page-blocks	2.000±0.000	2.061±0.018	2.522±0.086	2.050±0.150	2.684±0.415	**2.735±0.531**
Seeds	2.011±0.015	2.032±0.034	2.397±0.232	2.000±0.000	**2.674±0.297**	2.064±0.097
Spectrometer	3.139±0.092	2.204±0.059	**23.621±3.143**	2.171±0.176	2.182±0.024	2.176±0.051
Vehicle	2.008±0.012	2.088±0.027	**2.553±0.199**	2.082±0.109	2.401±0.116	2.238±0.256
Vowel	2.188±0.044	2.098±0.041	**4.120±0.408**	2.021±0.075	2.210±0.206	2.086±0.025
Waveform	2.000±0.000	2.022±0.006	2.082±0.023	2.000±0.000	**2.321±0.022**	2.022±0.006
Wine-quality	2.107±0.080	2.074±0.020	**2.719±0.102**	2.023±0.041	2.152±0.112	2.052±0.016
Yeast	2.154±0.027	2.100±0.018	**4.528±0.233**	2.034±0.065	2.333±0.215	2.079±0.029
Average	2.239	2.100	5.463	**2.058**	2.418	2.192
#Best	5	0	0	**11**	0	1

Table A.10: Comparison of u_{65} scores

Results on data without label noise

Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.828±0.027	0.834±0.023	0.849±0.017	**0.852±0.027**	0.839±0.023	0.839±0.023
Ecoli	0.848±0.033	0.860±0.033	0.861±0.032	**0.865±0.037**	0.864±0.032	0.863±0.034
Forest	0.888±0.014	**0.889±0.014**	0.888±0.043	**0.889±0.020**	0.888±0.013	**0.889±0.015**
Glass	0.775±0.048	0.778±0.048	0.773±0.040	**0.779±0.066**	0.777±0.051	0.774±0.057
Letter	0.800±0.016	**0.846±0.014**	0.793±0.019	0.842±0.018	0.842±0.015	0.837±0.015
Libras	0.779±0.031	**0.808±0.030**	0.777±0.070	0.790±0.035	0.799±0.027	0.800±0.029
Optdigits	0.971±0.004	0.976±0.003	**0.979±0.003**	**0.979±0.005**	0.976±0.003	0.976±0.004
Page-blocks	**0.969±0.008**	**0.969±0.009**	**0.969±0.006**	**0.969±0.010**	**0.969±0.009**	**0.969±0.009**
Seeds	0.929±0.021	0.931±0.035	0.927±0.032	0.930±0.046	0.930±0.035	**0.933±0.034**
Spectrometer	0.533±0.027	**0.553±0.027**	0.515±0.035	0.533±0.050	0.545±0.028	0.546±0.023
Vehicle	0.775±0.011	**0.785±0.014**	0.772±0.030	0.746±0.030	0.784±0.013	0.779±0.016
Vowel	0.917±0.009	0.929±0.010	0.931±0.016	**0.939±0.015**	0.925±0.008	0.922±0.010
Waveform	0.857±0.006	0.854±0.004	**0.860±0.007**	0.850±0.010	0.856±0.004	0.856±0.002
Wine-quality	0.704±0.013	**0.711±0.018**	0.705±0.021	0.689±0.019	0.710±0.020	0.704±0.022
Yeast	0.626±0.023	**0.632±0.022**	0.621±0.012	0.619±0.028	0.631±0.021	**0.632±0.021**
Average	0.813	**0.824**	0.815	0.818	0.822	0.821
#Best	0	**6**	3	4	0	2

Results on data with 30% label noise

Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.719±0.025	0.745±0.027	**0.751±0.023**	0.747±0.033	0.722±0.019	0.735±0.029
Ecoli	0.773±0.042	0.798±0.032	0.779±0.035	0.803±0.046	0.813±0.032	**0.827±0.036**
Forest	0.821±0.033	0.821±0.028	**0.838±0.027**	**0.838±0.029**	0.824±0.031	0.817±0.027
Glass	0.658±0.036	0.672±0.036	0.646±0.036	0.664±0.069	0.669±0.041	**0.683±0.043**
Letter	0.641±0.020	0.723±0.020	0.608±0.022	0.735±0.018	0.747±0.020	**0.762±0.017**
Libras	0.587±0.055	0.645±0.066	0.601±0.052	0.646±0.063	0.645±0.065	**0.648±0.053**
Optdigits	0.868±0.005	0.951±0.003	0.923±0.006	**0.969±0.004**	0.960±0.004	0.964±0.004
Page-blocks	0.899±0.011	0.899±0.011	0.916±0.012	0.914±0.011	0.917±0.011	**0.923±0.011**
Seeds	0.786±0.049	0.804±0.045	0.801±0.053	**0.809±0.070**	0.778±0.057	0.806±0.045
Spectrometer	0.443±0.030	0.519±0.028	0.390±0.038	0.513±0.030	**0.522±0.032**	**0.522±0.029**
Vehicle	0.696±0.035	**0.708±0.032**	0.704±0.040	0.693±0.041	0.707±0.032	**0.708±0.032**
Vowel	0.728±0.027	0.778±0.017	0.741±0.021	0.790±0.027	0.795±0.019	**0.804±0.017**
Waveform	0.819±0.007	0.795±0.005	**0.835±0.007**	0.827±0.006	0.777±0.006	0.795±0.005
Wine-quality	0.592±0.019	0.606±0.015	0.601±0.017	0.594±0.017	0.611±0.017	**0.626±0.012**
Yeast	0.551±0.017	0.569±0.022	0.518±0.017	0.557±0.026	0.584±0.026	**0.595±0.020**
Average	0.705	0.736	0.710	0.740	0.738	**0.748**
#Best	0	1	3	2	0	**9**

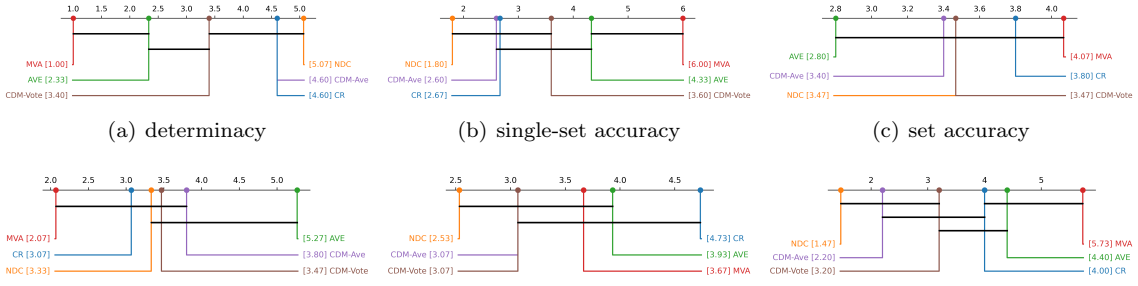
Table A.11: Comparison of u_{80} scores

Results on data without label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance scale	0.8421	0.8533	**0.8646**	0.8539	**0.8646**	0.8635
Balance-scale	0.842±0.029	0.853±0.023	**0.865±0.016**	0.854±0.027	**0.865±0.022**	0.864±0.023
Ecoli	0.870±0.035	0.883±0.029	0.873±0.032	0.866±0.036	**0.885±0.028**	0.884±0.032
Forest	0.898±0.015	**0.902±0.013**	0.894±0.042	0.889±0.020	0.900±0.012	0.901±0.014
Glass	0.805±0.049	**0.812±0.049**	0.796±0.039	0.780±0.065	0.808±0.054	0.805±0.059
Letter	0.837±0.016	**0.869±0.014**	0.811±0.018	0.844±0.017	0.864±0.015	0.858±0.015
Libras	0.825±0.030	**0.844±0.029**	0.803±0.071	0.793±0.034	0.833±0.029	0.830±0.029
Optdigits	0.980±0.003	**0.983±0.003**	**0.983±0.002**	0.980±0.005	0.982±0.003	0.982±0.004
Page-blocks	0.972±0.007	**0.973±0.009**	0.971±0.006	0.969±0.010	0.972±0.009	0.972±0.008
Seeds	0.936±0.023	0.943±0.034	0.933±0.030	0.930±0.047	0.943±0.034	**0.944±0.034**
Spectrometer	0.595±0.030	**0.606±0.030**	0.555±0.038	0.536±0.050	0.593±0.033	0.593±0.025
Vehicle	0.824±0.019	**0.828±0.014**	0.809±0.030	0.747±0.030	0.824±0.013	0.818±0.017
Vowel	0.945±0.006	**0.953±0.007**	0.946±0.015	0.940±0.015	0.949±0.006	0.946±0.009
Waveform	0.880±0.008	**0.898±0.005**	0.882±0.009	0.851±0.010	0.895±0.005	0.895±0.004
Wine-quality	0.746±0.020	**0.754±0.019**	0.731±0.021	0.690±0.019	0.750±0.020	0.743±0.022
Yeast	0.671±0.029	**0.676±0.022**	0.658±0.012	0.621±0.028	0.668±0.020	0.664±0.020
Average	0.842	**0.852**	0.834	0.820	0.849	0.847
#Best	0	**11**	3	0	2	1

Results on data with 30% label noise						
Data	CR	NDC	AVE	MVA	CDM/Ave	CDM/Vote
Balance-scale	0.776±0.028	0.790±0.030	**0.796±0.026**	0.749±0.034	0.780±0.021	0.786±0.030
Ecoli	0.815±0.038	0.833±0.031	0.808±0.035	0.805±0.047	0.840±0.032	**0.847±0.035**
Forest	0.856±0.030	0.859±0.025	**0.862±0.025**	0.839±0.029	0.858±0.028	0.851±0.026
Glass	0.725±0.041	**0.728±0.037**	0.693±0.033	0.667±0.070	0.718±0.039	0.726±0.038
Letter	0.705±0.018	0.762±0.018	0.640±0.021	0.738±0.018	0.780±0.019	**0.788±0.017**
Libras	0.656±0.056	0.685±0.067	0.639±0.051	0.647±0.064	0.683±0.067	**0.687±0.055**
Optdigits	0.914±0.004	0.965±0.003	0.936±0.005	0.970±0.004	0.971±0.003	**0.973±0.004**
Page-blocks	0.920±0.010	0.924±0.010	0.932±0.012	0.915±0.011	0.934±0.012	**0.938±0.011**
Seeds	0.842±0.030	0.845±0.047	0.840±0.048	0.810±0.070	0.824±0.048	**0.848±0.047**
Spectrometer	0.525±0.032	**0.572±0.027**	0.429±0.038	0.519±0.031	0.570±0.034	0.569±0.027
Vehicle	0.759±0.042	**0.772±0.034**	0.751±0.038	0.696±0.041	0.767±0.033	0.771±0.035
Vowel	0.795±0.026	0.824±0.017	0.784±0.019	0.793±0.028	0.836±0.019	**0.842±0.015**
Waveform	0.864±0.006	**0.868±0.005**	0.867±0.007	0.830±0.006	0.846±0.005	**0.868±0.005**
Wine-quality	0.661±0.015	0.657±0.018	0.644±0.020	0.597±0.017	0.660±0.019	**0.671±0.013**
Yeast	0.625±0.020	0.620±0.023	0.572±0.017	0.560±0.026	0.629±0.027	**0.632±0.022**
Average	0.763	0.780	0.746	0.742	0.780	**0.786**
#Best	0	3	2	0	0	**10**

Appendix B. Friedman and Nemenyi tests for cautious classifier comparison

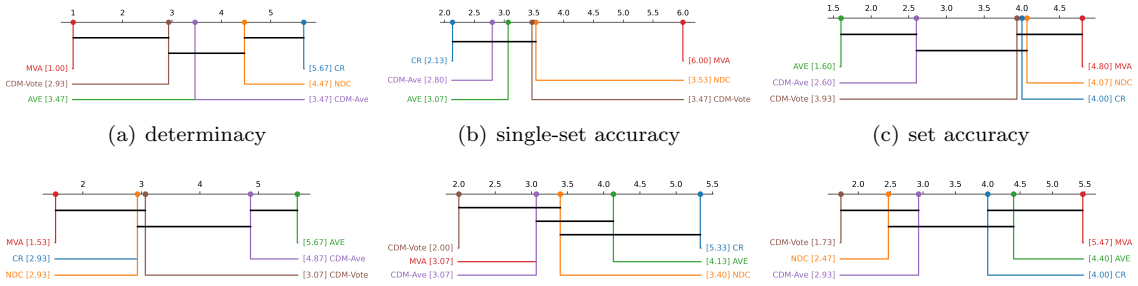
For further comparison, we conducted Friedman and Nemenyi tests following the recommendation in [45]. The results of the Friedman tests showed that there are significant differences between the compared models in terms of all evaluation measures except for the set accuracy on the original data sets. Thus, for each measure, we conducted the Nemenyi test to determine which pairs of models are significantly different: we display the results in the form of critical difference diagrams, in Figures B.4 and B.5. In these diagrams, for each measure, we show the average rank of each model across all datasets: the best-performing model receives a rank of 1, the second-best model a rank of 2, and so on. Models that could not be significantly deemed as different (p-value $\widehat{\alpha} > 0.05$) are connected by a horizontal black crossbar.



(a) determinacy (b) single-set accuracy (c) set accuracy

(d) set size (e) u_{65} score (f) u_{80} score

Figure B.4: Critical difference diagrams of the Nemenyi test for cautious classifier evaluations on original data sets.



(a) determinacy (b) single-set accuracy (c) set accuracy

(d) set size (e) u_{65} score (f) u_{80} score

Figure B.5: Critical difference diagrams of the Nemenyi test for cautious classifier evaluations on data sets with 30% label noise.

References

[1] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
URL https://doi.org/10.1023/a:1010933404324

[2] I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Computer Science 2 (3) (2021) 1–21.
URL https://doi.org/10.1007/s42979-021-00592-x

[3] J.-M. Bernard, An introduction to the imprecise dirichlet model for multinomial data, International Journal of Approximate Reasoning 39 (2-3) (2005) 123–150.
URL https://doi.org/10.1016/j.ijar.2004.10.002

[4] P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 3–34.
URL https://doi.org/10.1111/j.2517-6161.1996.tb02065.x

[5] J. Abellán, S. Moral, Building classification trees using the total uncertainty criterion, International Journal of Intelligent Systems 18 (12) (2003) 1215–1225.
URL https://doi.org/10.1002/int.10143
[6] C. J. Mantas, J. Abellán, Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data, Expert Systems with Applications 41 (5) (2014) 2514–2525.
URL https://doi.org/10.1016/j.eswa.2013.09.050
[7] J. Abellán, C. J. Mantas, J. G. Castellano, A random forest approach using imprecise probabilities, Knowledge-Based Systems 134 (2017) 72–84.
URL https://doi.org/10.1016/j.knosys.2017.07.019
[8] L. V. Utkin, An imprecise deep forest for classification, Expert Systems with Applications 141 (2020) 112978.
URL https://doi.org/10.1016/j.eswa.2019.112978
[9] L. V. Utkin, M. S. Kovalev, F. P. Coolen, Imprecise weighted extensions of random forests for classification and regression, Applied Soft Computing 92 (2020) 106324.
URL https://doi.org/10.1016/j.asoc.2020.106324
[10] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (3) (2001) 203–231.
URL https://doi.org/10.1023/a:1007601015854
[11] J. Abellan, A. R. Masegosa, Imprecise classification with credal decision trees, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20 (05) (2012) 763–787.
URL https://doi.org/10.1142/S0218488512500353
[12] M. C. Troffaes, Decision making under uncertainty using imprecise probabilities, International Journal of Approximate Reasoning 45 (1) (2007) 17–29.
URL https://doi.org/10.1016/j.ijar.2006.06.001
[13] L. M. De Campos, J. F. Huete, S. Moral, Probability intervals: a tool for uncertain reasoning, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2 (02) (1994) 167–196.
URL https://doi.org/10.1142/S0218488594000146
[14] P. Fink, Ensemble methods for classification trees under imprecise probabilities, Master's thesis, Ludwig Maximilian University of Munich (2012).
URL https://doi.org/10.5282/ubm/epub.25521
[15] S. Moral-García, C. J. Mantas, J. G. Castellano, M. D. Benítez, J. Abellan, Bagging of credal decision trees for imprecise classification, Expert Systems with Applications 141 (2020) 112944.
URL https://doi.org/10.1016/j.eswa.2019.112944
[16] H. Zhang, B. Quost, M.-H. Masson, Cautious random forests: a new decision strategy and some experiments, in: International Symposium on Imprecise Probability: Theories and Applications, PMLR, 2021, pp. 369–372.
URL https://proceedings.mlr.press/v147/zhang21a.html
[17] H. Zhang, B. Quost, M.-H. Masson, Cautious weighted random forests, Expert Systems with Applications 213 (2023) 118883.
URL https://doi.org/10.1016/j.eswa.2022.118883
[18] A. P. Dempster, Upper and Lower Probabilities Induced by a Multivalued Mapping, The Annals of Mathematical Statistics 38 (1967) 325–339.
URL https://doi.org/10.1214/aoms/1177698950
[19] G. Shafer, A mathematical theory of evidence, Princeton university press, 1976.
URL https://doi.org/10.2307/j.ctv10vm1qb
[20] T. Denoeux, Decision-making with belief functions: a review, International Journal of Approximate Reasoning 109 (2019) 87–110.
URL https://doi.org/10.1016/j.ijar.2019.03.009
[21] D. Dubois, H. Prade, On the use of aggregation operations in information fusion processes, Fuzzy sets and systems 142 (1) (2004) 143–161.
URL https://doi.org/10.1016/j.fss.2003.10.038
[22] N. Li, A. Martin, R. Estival, Heterogeneous information fusion: Combination of multiple supervised and unsupervised classification methods based on belief functions, Information Sciences 544 (2021) 238–265.
URL https://doi.org/10.1016/j.ins.2020.07.039
[23] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, T. Denœux, Multimodal information fusion for urban scene understanding, Machine Vision and Applications 27 (3) (2016) 331–349.
URL https://doi.org/10.1007/s00138-014-0649-7
[24] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE transactions on systems, man, and cybernetics 25 (5) (1995) 804–813.
URL https://doi.org/10.1109/21.376493
[25] T. Denoeux, A neural network classifier based on Dempster-Shafer theory, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 30 (2) (2000) 131–150.
URL https://doi.org/10.1109/3468.833094
[26] L. Huang, S. Ruan, P. Decazes, T. Denœux, Lymphoma segmentation from 3d pet-ct images using a deep evidential network, International Journal of Approximate Reasoning 149 (2022) 39–60.
URL https://doi.org/10.1016/j.ijar.2022.06.007
[27] Z. Tong, P. Xu, T. Denoeux, An evidential classifier based on Dempster-Shafer theory and deep learning, Neurocomputing 450 (2021) 275–293.

URL https://doi.org/10.1016/j.neucom.2021.03.066
[28] Z. Tong, P. Xu, T. Denoeux, Evidential fully convolutional network for semantic segmentation, Applied Intelligence 51 (2021) 6376–6399.
URL https://doi.org/10.1007/s10489-021-02327-0
[29] F. Xiao, Z. Cao, A. Jolfaei, A novel conflict measurement in decision-making and its application in fault diagnosis, IEEE Transactions on Fuzzy Systems 29 (1) (2020) 186–197.
URL https://doi.org/10.1109/TFUZZ.2020.3002431
[30] Y. Xu, Y. Li, Y. Wang, D. Zhong, G. Zhang, Improved few-shot learning method for transformer fault diagnosis based on approximation space and belief functions, Expert Systems with Applications 167 (2021) 114105.
URL https://doi.org/10.1016/j.eswa.2020.114105
[31] H. Zhang, Y. Deng, Weighted belief function of sensor data fusion in engine fault diagnosis, Soft computing 24 (2020) 2329–2339.
URL https://doi.org/10.1007/s00500-019-04063-7
[32] P. Smets, Constructing the pignistic probability function in a context of uncertainty, in: UAI '89: Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, Vol. 89, 1989, pp. 29–40.
URL https://doi.org/10.1016/B978-0-444-88738-2.50010-5
[33] P. Smets, Decision making in the tbm: the necessity of the pignistic transformation, International journal of approximate reasoning 38 (2) (2005) 133–147.
URL https://doi.org/10.1016/j.ijar.2004.05.003
[34] T. Denoeux, Analysis of evidence-theoretic decision rules for pattern classification, Pattern recognition 30 (7) (1997) 1095–1107.
URL https://doi.org/10.1016/S0031-3203(96)00137-9
[35] L. Ma, T. Denoeux, Partial classification in the belief function framework, Knowledge-Based Systems 214 (2021) 106742.
URL https://doi.org/10.1016/j.knosys.2021.106742
[36] M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, in: International Journal of Approximate Reasoning, Vol. 53, Elsevier, 2012, pp. 1282–1301.
URL https://doi.org/10.1016/j.ijar.2012.06.022
[37] T. Denœux, Constructing belief functions from sample data using multinomial confidence regions, International Journal of Approximate Reasoning 42 (3) (2006) 228–252.
URL https://doi.org/10.1016/j.ijar.2006.01.001
[38] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, W. Waegeman, Efficient set-valued prediction in multi-class classification, Data Mining and Knowledge Discovery 35 (4) (2021) 1435–1469.
URL https://doi.org/10.1007/s10618-021-00751-x
[39] M. Kelly, R. Longjohn, K. Nottingham, Uci machine learning repository.
URL https://archive.ics.uci.edu
[40] Kaggle, Kaggle.
URL https://www.kaggle.com
[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.
URL https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
[42] V.-L. Nguyen, M. H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Machine Learning 111 (1) (2022) 89–122.
URL https://doi.org/10.1007/s10994-021-06003-9
[43] C. K. Murphy, Combining belief functions when evidence conflicts, Decision support systems 29 (1) (2000) 1–9.
URL https://doi.org/10.1016/S0167-9236(99)00084-6
[44] H. Zhang, B. Quost, M.-H. Masson, Cautious decision-making for tree ensembles, in: 17th European Conference on Symbolic and Quantitative Approaches with Uncertainty, Springer, 2023, pp. 3–14.
URL https://doi.org/10.1007/978-3-031-45608-4_1
[45] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine learning research 7 (2006) 1–30.
URL https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf