Cautious Decision-Making for Tree Ensembles

Haifei Zhang[1,2], Benjamin Quost[1,2], and Marie-Hélène Masson[1,3]

[1] UMR CNRS 7253 Heudiasyc, 60200 Compiègne, France
{haifei.zhang, benjamin.quost, mylene.masson}@hds.utc.fr
[2] Université de Technologie de Compiègne, 60200 Compiègne, France
[3] IUT de l'Oise, Université de Picardie Jules Verne, 60000 Beauvais, France

Abstract. Cautious classifiers are designed to make indeterminate decisions when the uncertainty on the input data or the model output is too high, so as to reduce the risk of making wrong decisions. In this paper, we propose two cautious decision-making procedures, by aggregating trees providing probability intervals via the imprecise Dirichlet model. The trees are aggregated in the belief functions framework, by maximizing the lower expected discounted utility, so as to achieve a good compromise between model accuracy and determinacy. They can be regarded as generalizations of the two classical aggregation strategies for tree ensembles, i.e. averaging and voting. The efficiency and performance of the proposed procedures are tested on random forests and illustrated on three UCI datasets.

Keywords: Cautious decision making · Belief functions · Lower expected utility · Ensemble learning.

1 Introduction

Tree ensembles like random forests are highly efficient and accurate machine-learning models widely applied in various domains [5,16]. Tree outputs consist of precise class probability estimates based on counts of training instances falling in the leaf nodes. Decisions are classically made either by averaging the probabilities or by majority voting. However, trees may lack robustness when confronted with low-quality data, for instance for noisy samples, or samples located in low-density regions of the input space. To overcome this issue, previous works have proposed to use the imprecise Dirichlet model (IDM) so as to replace precise class probability estimates with a convex set of probability distributions (in the form of probability intervals) whose size depends on the number of training samples [4,21].

The joint use of the IDM and decision trees is not new, it has been explored in two directions. First, it has been used to improve the training of single trees or tree ensembles. Credal decision trees (CDT) [3,11] and credal random forests (CRF) [1] use a maximum entropy principle to select split features and values from the probability intervals obtained via the IDM, thus improving robustness to data noise. To enhance 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 [19] and weights associated with each tree in the ensemble can be learned to further optimize their combination [20]. Second, the probability intervals given by the IDM can also be used to make cautious decisions, thereby reducing the risk of prediction error [4,15]. A cautious decision is a set-valued decision, i.e. a cautious classifier may return a set of classes instead of a single one when the uncertainty is too high. An imprecise credal decision tree (ICDT) [2] is a single tree where set-valued predictions are returned by applying the interval dominance principle [18] to the probability intervals obtained via the IDM.

In tree ensembles, applying cautious decision-making strategies becomes more complex. One approach consists in 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 [6,9]. Another approach consists in allowing each tree to make a cautious decision first, before pooling them. The Minimum-Vote-Against (MVA) is such an approach, where the classes with minimal opposition are retained [12]. 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 [23,24], we have proposed a generalized voting aggregation strategy for binary cautious classification within the belief function framework. In the present paper, we generalize these previous works in the multi-class case. After recalling background material in Section 2, we propose in Section 3 two cautious decision-making strategies in the belief function framework, which generalize averaging and voting for imprecise tree ensembles. These strategies are axiomatically principled: they amount to maximizing the lower expected discounted utility, rather than the expected utility as done in the conventional case. Our approach can be applied to any kind of classifier ensemble where classifier outputs are probability intervals; however, it is particularly well-suited to tree ensembles. The experiments reported in Section 4 show that a good compromise between accuracy and determinacy can be achieved and that our algorithms remain tractable even in the case of a high number of classes.

2 Preliminaries

2.1 Imprecise Dirichlet Model and trees

Let $F = \{h_1, \ldots, h_T\}$ be a random forest with trees h_t trained on a classification problem with $K \geq 2$ classes. Let $h_t(x)$ be the leaf in which a given test instance $x \in \mathcal{X}$ falls for tree h_t, and let n_{tj} denote the number of training samples of class c_j in $h_t(x)$.

The IDM consists in using a family of Dirichlet priors for estimating the class posterior probabilities $\mathbb{P}(c_j|x, h_t)$, resulting in interval estimates:

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

where $N_t = \sum_{j=1}^{K} n_{tj}$ is the total number of instances in $h_t(x)$, and s can be interpreted as a number of additional virtual samples with unknown actual classes also falling in $h_t(x)$. In the case of trees, the IDM therefore provides a natural local estimate of epistemic uncertainty, i.e. due to the lack of training data.

2.2 Belief Functions

The theory of belief functions [7,17] provides a general framework for modeling and reasoning with uncertainty. Let the frame of discernment $\Omega = \{c_1, c_2, \ldots, c_K\}$ denote the finite set that contains all values for our class variable C of interest.

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 $C \in A$ only; $m(\Omega)$ 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 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: they are respectively defined as

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{2}$$

for all $A \subseteq \Omega$. In a nutshell, $Bel(A)$ measures the total degree of support to A, and $Pl(A)$ the degree of belief not contradicting A. These two functions are dual since $Bel(A) = Pl(\Omega) - Pl(\overline{A})$, with $\overline{A} = \Omega \backslash A$. The mass, belief, and plausibility functions are in one-to-one correspondence and can be retrieved from each other.

2.3 Decision Making with Belief Functions

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.

Assume our knowledge of the class of the test instance is represented by a mass function m: the expected utility criterion under probability setting may be extended using lower and upper expected utilities, respectively defined as the weighted averages of the minimum and maximum utility within each focal set:

$$\underline{EU}(m, f_i, U) = \sum_{B \subseteq \Omega} m(B) \min_{c_j \in B} u_{ij}, \quad \overline{EU}(m, f_i, U) = \sum_{B \subseteq \Omega} m(B) \max_{c_j \in B} u_{ij}. \tag{3}$$

We obviously have $\underline{EU}(m, f_i, U) \leq \overline{EU}(m, f_i, U)$, the equality applies when m is Bayesian. Note that actions f_i are not restricted to choosing a single class. Based on Eq. (3), we may choose the action with the highest lower expected utility (pessimistic attitude), or with the highest upper expected utility (optimistic attitude). More details on decision-making principles in the belief functions framework can be found in [8].

2.4 Evaluation of Cautious Classifiers

Unlike traditional classifiers, cautious classifiers may return indeterminate decisions so that classical evaluation criteria are no longer applicable. We mention here several evaluation criteria to evaluate the quality of such set-valued predictions: the *determinacy* counts the proportion of samples that are determinately classified; the *single-set accuracy* measures the proportion of correct determinate decisions; the *set accuracy* measures the proportion of indeterminate predictions containing the actual class; the *set size* gives the average size of indeterminate predictions; finally, the *discounted utility* calculates the expected utility of predictions, discounted by the size of the predicted set as explained below.

Let A be a decision made for a test sample with actual class c. Zaffalon et al. [22] proposed to evaluate this decision using a discounted utility function u_α which rewards cautiousness and reliability as follows:

$$u_\alpha(A, c) = d_\alpha(|A|)\mathbb{1}(c \in A), \tag{4}$$

where $|A|$ is the cardinality of A and $d_\alpha(.)$ is a discount ratio that adjusts the reward for cautiousness: cautiousness is considered preferable to random guessing whenever $d_\alpha(|A|) > 1/|A|$. The u_{65} and u_{80} scores are two notable special cases:

$$d_{65}(|A|) = \frac{1.6}{|A|} - \frac{0.6}{|A|^2}, \quad d_{80}(|A|) = \frac{2.2}{|A|} - \frac{1.2}{|A|^2}\ . \tag{5}$$

Theorem 1. *Given the utility matrix U of general term $u_{Aj} = u_\alpha(A, c_j)$ with $c_j \in \Omega$ and $A \subseteq \Omega$ an imprecise decision, the lower expected utility $\underline{EU}(m, A, U)$ is equal to $d_\alpha(|A|)Bel(A)$.*

Proof. Following Eq. (3), and taking any $A \subseteq \Omega$ as action, we have

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

Indeed, for any $B \cap A \neq \emptyset$ such that $B \nsubseteq A$, there obviously exists $c_j \in B$ such that $c_j \notin A$: thus, $\min_{c_j \in B} \mathbb{1}(c_j \in A) = 1$ iff $B \subseteq A$.

3 Cautious Decision-Making for Tree Ensembles

Classical belief-theoretic combination approaches such as the conjunctive rule, which assumes independence and is sensitive to conflict, are in general not well-suited to combining tree outputs. This calls for specific aggregation strategies, such as those proposed below.

Algorithm 1: Cautious Decision Making by Averaging

Input: Tree outputs $\left\{(\underline{p}_{tj},\ \overline{p}_{tj}), t = 1, \ldots, T,\ j = 1, \ldots, K\right\}$, discount ratio d_α
Output: Decision A

```
for j = 1, ..., K do
    m({c_j}) = 1/T × Σ_{t=1}^{T} p_tj
m(Ω) = 1 − Σ_{j=1}^{K} m({c_j})
Sort classes by decreasing mass: m({c_(1)}) ≥ m({c_(2)}) ≥ ··· ≥ m({c_(K)})
A = ∅
bel = 0
mleu = 0                                          // Maximum lower EU
for i = 1, ..., K do
    bel = bel + m({c_(i)})
    leu = d_α(i) × bel                            // Lower EU
    if leu > mleu then
        mleu = leu
        A = A ∪ {c_(i)}
Return A
```

3.1 Generalization of Averaging

We assume that the output of each decision tree h_t is no longer a precise probability distribution, but a set of probability intervals as defined by Eq. (1). As mentioned above, the corresponding quasi-Bayesian mass function is

$$m_t(\{c_j\}) = \underline{p}_{tj},\ j = 1, \ldots, K; \quad m_t(\Omega) = 1 - \sum_{j=1}^{K} m_t(\{c_j\}). \tag{6}$$

These masses can then be averaged across all trees:

$$m(\{c_j\}) = \frac{\sum_{t=1}^{T} m_t(\{c_j\})}{T},\ j = 1, \ldots, K; \quad m(\Omega) = \frac{\sum_{t=1}^{T} m_t(\Omega)}{T}. \tag{7}$$

To make a decision based on this mass function, we build a sequence of nested subsets $A \subseteq \Omega$ by repeatedly aggregating the class with the highest mass, and we choose the subset $A^\star$ which maximizes $\underline{EU}(A) := \underline{EU}(m, A, U)$ over all $A \subseteq \Omega$. Note that there exists several kinds of decision-making strategies resulting in imprecise predictions [10]; maximizing the lower EDU is a conservative strategy, and can be done efficiently using gthe algorithms presented below.

Theorem 2. *Consider the mass function in Eq.* (7) *with classes sorted by decreasing mass:* $m(\{c_{(j)}\}) \geq m(\{c_{(j+1)}\})$*, for* $j = 1, \ldots, K-1$*. Scanning the sequence of nested subsets* $\{c_{(1)}\} \subset \{c_{(1)}, c_{(2)}\} \subset \cdots \subset \Omega$ *makes it possible to identify the subset* $A^\star = \arg\max \underline{EU}(A)$ *in complexity* $O(K)$*.*

Proof. Since the masses $m(\{c_{(j)}\})$ are sorted in a decreasing order, the focal element with the highest belief among those of cardinality i is $A_i^\star = \{c_{(j)}, j =$

Algorithm 2: Tree aggregation via interval dominance

Input: Tree outputs $\left\{(\underline{p}_{tj}, \overline{p}_{tj}), t = 1, \dots T, j = 1, \dots, K\right\}$
Output: Mass function m

$m(A) = 0, \forall A \subseteq \Omega$
for $t = 1, \dots, T$ **do**
  $DC = \emptyset$ `// set of dominated classes`
  for $j = 1, \dots, K$ **do**
    for $j' = 1, \dots, K$ *and* $j' \neq j$ **do**
      if $\overline{p}_{tj} < \underline{p}_{tj'}$ **then**
        $DC = DC \cup c_j$
        break
  $NDC = \Omega \setminus DC$ `// non-nominated classes`
  $m(NDC) = m(NDC) + \frac{1}{T}$
Return m

$1, \dots, i\}$, i.e. $Bel(A_i^\star) = \sum_{j=1}^{i} m(\{c_{(j)}\}) \geq Bel(B)$, for all $B \subseteq \Omega : |B| = i$. Since $d_\alpha(|A|)$ only depends on $|A|$, $A_i^\star$ maximizes the lower EU over all subsets of size i. As a consequence, keeping the subset with maximal lower EU in the sequence of nested subsets defined above gives the maximizer $A^\star$ in time complexity $O(K)$.

The overall procedure, hereafter referred to as CDM_Ave (standing for "cautious decision-making via averaging"), extends classical averaging for precise probabilities to averaging mass functions across imprecise trees, is summarized in Alg. 1. Note that a theorem similar to Theorem 2 was proven in [13], which addressed set-valued prediction in a probabilistic framework for a wide range of utility functions. Since the masses considered here are quasi-Bayesian, the procedure described in Alg. 1 is close to that described in [13]. The overall complexity of Alg. 1 is $O(K \log K)$—due to sorting the classes by decreasing mass.

3.2 Generalization of Voting

We now address the combination of probability intervals via voting. Our approach consists to identify first, for each tree, the set of non-dominated classes as per interval dominance, i.e. trees vote for the corresponding subset of classes. Then, we again compute the subset $A^\star$ maximizing $\underline{EU}(A)$ over all $A \subseteq \Omega$.

Alg. 2 describes how interval dominance can be used to aggregate all tree outputs into a single mass function m, in time complexity $O(TK^2)$. In this approach, the focal elements of m can be any subset of Ω. Since m is not quasi-Bayesian anymore, maximizing the lower EU requires in principle to check all subsets of Ω in the decision step: the worst-case complexity of $O(2^K)$ prohibits using this strategy for datasets with large numbers of classes.

In order to reduce the complexity, we exploit three tricks: (i) we arbitrarily restrict the decision to subsets $A \subseteq \Omega$ with cardinality $|A| \leq M$, which reduces

Algorithm 3: Cautious Decision Making by Voting

Input: Mass function m from Alg 2, cardinality bound M, discount ratio d_α
Output: Decision A

$m =$ Alg 2 (I_{tj}, $t = 1, \dots T$, and $j = 1, \dots, K$)
$FE = \emptyset$ // Focal Elements
$\Omega' = \emptyset$ // Considering Classes
$A = \emptyset$
$mleu = 0$ // Maximum lower EU
for $i = 1, \dots, M$ **do**
  $d = d_\alpha(i)$
  if $mleu > d$ **then**
    Return A // Early Stopping
  else
    $FE = FE \cup \{B : \ m(B) > 0, \ |B| = i, \ B \subseteq \Omega\}$
    $\Omega' = \Omega' \cup \{c : \ c \in B, \ B \in FE\}$
    for *all* $B \subseteq \Omega'$ *and* $|B| = i$ **do**
      $bel = \sum_{C \in FE, C \subseteq B} m(C)$
      $leu = d \times bel$ // Lower EU for B
      if $leu > mleu$ **then**
        $mleu = leu$
        $A = B$
Return A

the complexity to $O(\sum_{k=1}^{M} \binom{K}{k})$; then, we can show that (ii) when searching for a maximizer of the lower EU by scanning subsets of classes of increasing cardinality, we can stop the procedure when larger subsets are known not to further improve the lower EU (see Proposition 1); and (iii) during this 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.

Proposition 1. *If the lower EU of a subset $A \subseteq \Omega$ is (strictly) greater than $d_\alpha(i)$ for some $i > |A|$, then it is (strictly) greater than that of any subset $B \subseteq \Omega$ with cardinality $|B| \geq i$.*

Proof. Let $A \subset \Omega$ be a subset of classes (typically, the current maximizer of the lower EU in the procedure described in Alg 3). Assume that $\underline{EU}(A) > d_\alpha(i)$ for some $i > |A|$. Since $Bel(B) \leq 1$ for all $B \subseteq \Omega$, then $\underline{EU}(A) > \underline{EU}(B)$ for all subsets B such that $|B| = i$. The generalization to all subsets B such that $|B| \geq i$ comes from $d_\alpha(i)$ being monotone decreasing in i.

Proposition 2. *The subset $A_i^\star \subseteq \Omega$ maximizing the lower EU among all A such that $|A| = i$ is a subset of Ω_i which is the set of classes appearing in focal elements B such that $|B| \leq i$.*

Proof. Let Ω_i be the set of classes appearing in focal elements of cardinality less or equal to i, for some $i \in \{1, \dots, K\}$. Assume a subset A of cardinality i is such

that $A = A_1 \cup A_2$, with $A \cap \Omega_i = A_1$, then, $Bel(A) = Bel(A_1)$. If $A_2 \neq \emptyset$, then $\underline{EU}(A) < \underline{EU}(A_1)$ since $|A_1| < |A|$: classes $c_j \notin \Omega_i$ necessarily decrease $\underline{EU}(A)$. Moreover, since $Bel(A)$ sums masses $m(B)$ of subsets $B \subseteq A$, any focal element B such that $|B| > i$ does not contribute to $Bel(A)$.

The procedure described in Alg. 3, hereafter referred to as CDM_Vote, extends voting when votes are expressed as subsets of classes and returns the subset $A^\star = \arg\max \underline{EU}(A)$ among all subsets $A \subseteq \Omega$: $|A| \leq M \leq K$. It generalizes the method proposed in [23,24] for binary cautious classification. It is computationally less efficient than CDM_Ave, even if time complexity can be controlled, as it will be shown in the experimental part.

4 Experiments and Results

We report here two experiments. First, we study the effectiveness of controlling the complexity of CDM_Vote. Then we compare the performances of both versions of CDM with two other imprecise tree aggregation strategies (MVA and Averaging). In both experiments, we used three datasets from the UCI: *letter*, *spectrometer*, and *vowel*, with a diversity in size (2000, 531, and 990 samples), number of classes (26, 48, and 11), and number of features (16, 100, and 10). We applied the scikit-learn implementation of random forests with default parameter setting: n_estimators=100, criterion='gini', and min_samples_leaf=1 [14]. We have set the parameter M to 5 in Alg. 3.

4.1 Decision-Making Efficiency

First, we studied the time complexity as a function of the number of labels. For a given integer i, we first picked i labels at random and extracted the corresponding samples. Then, we trained a random forest with the parameter s of the IDM set to 1, and processed the test data using CDM_Vote. During the test phase, we recorded for each sample the elapsed time of the entire process (interval dominance plus maximizing lower expected discounted utility), and the elapsed time needed to maximize the lower EU after having applied interval dominance, respectively referred to as ID+MLEDU and MLEDU. For each i, we report average times per 100 inferences, computed over 10 repetitions of the above process. Since for high values of i, decision-making would be intractable without any control of the complexity, we compared the efficiency when using all tricks in Section 3.2 with that when using only the two first ones.

Fig. 1 shows that for a small number of labels (e.g., less than 15), trick 3 (filtering out subsets $A \not\subseteq \Omega_i$) does not significantly improve the efficiency, as the time required for interval dominance dominates. However, for a large number of labels, the time required for maximizing the lower EU dominates, and filtering out subsets $A \not\subseteq \Omega_i$ accelerates the procedure. Apart from interval dominance, this filtering step accelerates the decision-making process regardless of the number of labels, as shown in the right column of Fig. 1. This experiment demonstrates that CDM_Vote remains applicable with a large number of labels.

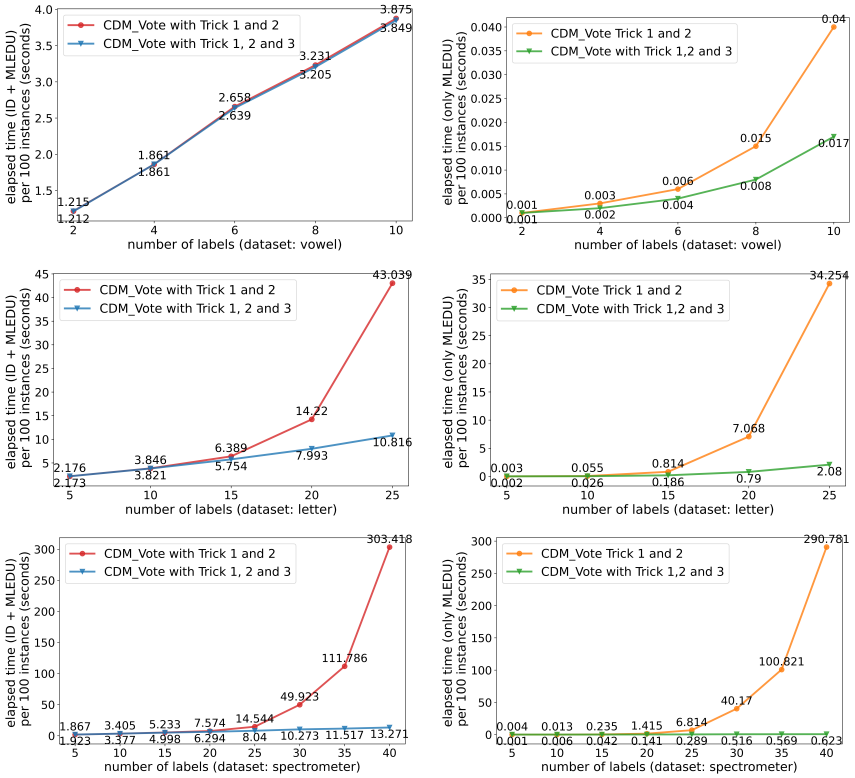

Fig. 1. Decision-making time complexity of CDM_Vote according to the number of labels (for 100 samples). Left: ID+MLEDU, right: MLEDU only.

4.2 Cautious Decision-Making Performance Comparison

We compared CDM_Ave and CDM_Vote with Minimum-Vote-Against (MVA) and Averaging (AVE) according to the metrics listed in Section 2.4. For each metric, each dataset, and each aggregation approach, we used 10-fold cross-validation: the results (mean and standard deviation) are reported in Tables 1(a) to 1(c), with the best results printed in bold. In each CV fold, the optimal value of s for each model is determined by a separate validation set using the u_{65} score. CDM_Vote and CDM_Ave also make decisions using the d_{65} discount ratio.

The results show that MVA often tends to be determinate, while AVE and CDM tend to be more cautious, without a clear difference between both latter. The same can be observed for the single-set accuracy which is negatively correlated to determinacy. AVE always achieves the highest set accuracy, due to a high average set size of indeterminate predictions, in contrast to MVA. Our approach turns out to be in-between. According to the u_{65} and u_{80} scores, CDM turns out to provide a better compromise between accuracy (single-set accuracy and set accuracy) and cautiousness (determinacy and set size) than MVA and AVE.

Table 1. Cautious decision-making performance comparisons.

(a) Dataset: vowel (11 labels)

Criteria	MVA	AVE	CDM_Vote	CDM_AVE
Determinacy	**0.995±0.007**	0.918±0.032	0.874±0.036	0.867±0.038
Single-set accuracy	0.952±0.024	0.982±0.015	0.991±0.013	**0.994±0.011**
Set accuracy	0.944±0.168	**0.974±0.063**	0.967±0.056	0.962±0.053
Set size	**2.0±0.0**	2.418±0.275	2.054±0.064	2.056±0.064
u_{65} score	**0.950±0.025**	0.948±0.019	0.944±0.016	0.941±0.017
u_{80} score	0.950±0.024	0.960±0.017	**0.963±0.013**	0.960±0.013

(b) Dataset: letter (26 labels)

Criteria	MVA	AVE	CDM_Vote	CDM_AVE
Determinacy	**0.988±0.008**	0.772±0.026	0.816±0.026	0.811±0.026
Single set accuracy	0.861±0.026	**0.964±0.016**	0.943±0.018	0.949±0.016
Set-accuracy	0.717±0.259	**0.949±0.030**	0.710±0.078	0.728±0.071
Set size	**2.077±0.208**	12.197±1.390	2.139±0.058	2.163±0.062
u_{65} score	0.855±0.026	0.809±0.023	0.852±0.021	**0.856±0.020**
u_{80} score	0.856±0.026	0.826±0.022	0.871±0.020	**0.876±0.019**

(c) Dataset: spectrometer (48 labels)

Criteria	MVA	AVE	CDM_Vote	CDM_AVE
Determinacy	**0.978±0.023**	0.544±0.071	0.480±0.063	0.499±0.064
Single-set accuracy	0.550±0.068	0.694±0.074	**0.700±0.076**	0.690±0.077
Set accuracy	0.741±0.280	**0.817±0.080**	0.722±0.097	0.712±0.099
Set size	**2.067±0.222**	9.582±3.213	2.132±0.072	2.121±0.065
u_{65} score	0.545±0.066	0.538±0.050	**0.571±0.051**	0.568±0.052
u_{80} score	0.546±0.066	0.580±0.052	**0.626±0.055**	0.621±0.055

However, there is no significant difference between CDM_Vote and CDM_Ave. Moreover, since the average cardinality of predictions is around 2, setting $M = 5$ has no influence on the performances. In summary, our approaches seem to be appropriate for applications requiring highly reliable determinate predictions and indeterminate predictions containing as few labels as possible.

5 Conclusions and Perspectives

In this paper, we proposed two aggregation strategies to make cautious decisions from trees providing probability intervals as outputs, which are typically obtained by using the imprecise Dirichlet model. The two strategies respectively generalize averaging and voting for tree ensembles. In both cases, they aim at making decisions by maximizing the lower expected discounted utility, thus providing set-valued predictions. The experiments conducted on different datasets confirm the interest of our proposals in order to achieve a good compromise between model accuracy and determinacy, especially for difficult datasets, with a limited computational complexity.

In the future, we may further investigate how to make our cautious decision-making strategy via voting more efficient and tractable for classification problems with a high number of classes. We may also compare both our cautious decision-making strategies with other cautious classifiers beyond tree-based models.

References

1. Abellán, J., Mantas, C.J., Castellano, J.G.: A random forest approach using imprecise probabilities. Knowledge-Based Systems **134**, 72–84 (2017)
2. Abellan, J., Masegosa, A.R.: Imprecise classification with credal decision trees. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **20**(05), 763–787 (2012)
3. Abellán, J., Moral, S.: Building classification trees using the total uncertainty criterion. International Journal of Intelligent Systems **18**(12), 1215–1225 (2003)
4. Bernard, J.M.: An introduction to the imprecise dirichlet model for multinomial data. International Journal of Approximate Reasoning **39**(2-3), 123–150 (2005)
5. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
6. De Campos, L.M., Huete, J.F., Moral, S.: Probability intervals: a tool for uncertain reasoning. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **2**(02), 167–196 (1994)
7. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. The Annals of Mathematical Statistics **38**, 325–339 (1967)
8. Denoeux, T.: Decision-making with belief functions: a review. International Journal of Approximate Reasoning **109**, 87–110 (2019)
9. Fink, P.: Ensemble methods for classification trees under imprecise probabilities. Master's thesis, Ludwig Maximilian University of Munich (2012)
10. Ma, L., Denoeux, T.: Making set-valued predictions in evidential classification: A comparison of different approaches. In: International Symposium on Imprecise Probabilities: Theories and Applications. pp. 276–285. PMLR (2019)
11. Mantas, C.J., Abellán, J.: Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data. Expert Systems with Applications **41**(5), 2514–2525 (2014)
12. Moral-García, S., Mantas, C.J., Castellano, J.G., Benítez, M.D., Abellan, J.: Bagging of credal decision trees for imprecise classification. Expert Systems with Applications **141**, 112944 (2020)
13. 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)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(Oct), 2825–2830 (2011)
15. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning **42**(3), 203–231 (2001)
16. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN Computer Science **2**(3), 1–21 (2021)
17. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
18. Troffaes, M.C.: Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning **45**(1), 17–29 (2007)
19. Utkin, L.V.: An imprecise deep forest for classification. Expert Systems with Applications **141**, 112978 (2020)
20. Utkin, L.V., Kovalev, M.S., Coolen, F.P.: Imprecise weighted extensions of random forests for classification and regression. Applied Soft Computing **92**, 106324 (2020)
21. Walley, P.: Inferences from multinomial data: learning about a bag of marbles. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 3–34 (1996)

22. Zaffalon, M., Corani, G., Mauá, D.: Evaluating credal classifiers by utility-discounted predictive accuracy. In: International Journal of Approximate Reasoning. vol. 53, pp. 1282–1301. Elsevier (2012)
23. Zhang, H., Quost, B., Masson, M.H.: Cautious random forests: a new decision strategy and some experiments. In: International Symposium on Imprecise Probability: Theories and Applications. pp. 369–372. PMLR (2021)
24. Zhang, H., Quost, B., Masson, M.H.: Cautious weighted random forests. Expert Systems with Applications **213**, 118883 (2023)