Cautious weighted random forests

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson

Cautious Weighted Random Forests

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

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

ARTICLE INFO

Keywords:
Cautious classification
Imprecise classification
Imprecise Dirichlet Model
Belief functions

ABSTRACT

Random forest is an efficient and accurate classification model, which makes decisions by aggregating a set of trees, either by voting or by averaging class posterior probability estimates. However, tree outputs may be unreliable in presence of scarce data. The imprecise Dirichlet model (IDM) provides workaround, by replacing point probability estimates with interval-valued ones. This paper investigates a new tree aggregation method based on the theory of belief functions to combine such probability intervals, resulting in a cautious random forest classifier. In particular, we propose a strategy for computing tree weights based on the minimization of a convex cost function, which takes both determinacy and accuracy into account and makes it possible to adjust the level of cautiousness of the model. The proposed model is evaluated on 25 UCI datasets and is demonstrated to be more adaptive to the noise in training data and to achieve a better compromise between informativeness and cautiousness.

1. Introduction

Setting

Nowadays, machine learning algorithms have been applied to various fields with remarkable success, such as e.g. loan approval (Baesens et al., 2003; Ambika & Biradar, 2021), medical diagnosis (Foster et al., 2014), recommendation systems (Isinkaye et al., 2015), and autonomous driving (Maurer et al., 2016). Among the numerous machine learning approaches, ensemble learning is prominent because of its ability to combine numerous learners so as to improve classification accuracy. Ensemble learning can be divided into two categories, based on the classifiers being trained independently from each other or not. *Independent approaches* notably include bagging (Breiman, 1996) and random forests (Breiman, 2001). *Dependent methods* include stacking (Wolpert, 1992) and a variety of boosting algorithms (Freund et al., 1999), such as AdaBoost (Freund & Schapire, 1997), XGBoost (Chen & Guestrin, 2016), etc.

Random forest, as a variant of bagging, consists in training a large number of unpruned decision trees and aggregating them to make a decision. Training diverse individual classifiers, and combining them in an appropriate fashion, makes it possible to limit the influence of outliers and thus to achieve a high classification accuracy (Grandvalet, 2004). In addition to their very good classification performances, random forests inherit a number of properties of decision trees, and in particular their versatility (i.e., ability to handle quantitative as well as qualitative predictors, outside of any distributional assumption), and their compatibility with explanation strategies, as illustrated by recent works (Haddouchi & Berrado, 2019).

When training a random forest, one of the keys is to choose the appropriate combination strategy. There are two main aggregation approaches: voting schemes, such as majority voting or weighted voting; and averaging class probabilities across trees and picking the most probable class (Shaik & Srinivasan, 2019). The difference between these two families of approaches has been discussed in (Sage et al., 2020). Note that both approaches can be made more elaborate by assigning weights to trees. In (Li et al., 2010), the normalized accuracy of each tree estimated on out-of-bag instances is used as a weight. In (Kim et al., 2011), weights are assigned by an iterative approach that takes both the capacity of the classifiers and the difficulty of the examples into account. Forward step-wise model selection is considered in (Caruana et al., 2004) as an implicit weight assignment strategy. Finally, in (Utkin et al., 2020, 2019), the weights are determined by optimizing a criterion based on the accuracy of the forest.

Imprecise classification

Traditionally, classification models make precise decisions, in the form of a single class (or a point prediction in regression). However, enforcing the assignment of the instance to a single class is questionable when the available information from which the decision is made is scarce. As well, in ensemble learning, a large conflict between the outputs of individual learners should lead to avoiding reaching a definitive conclusion. Therefore, in some critical systems where wrong decisions may have serious consequences, one alternative is to produce imprecise predictions such as subsets of plausible classes (or intervals in regression), to ensure that the model will avoid taking chances when excessive uncertainty occurs. Following Provost & Fawcett (2001), when imprecise predictions are allowed to be made, we will refer to the corresponding model as a *cautious classifier*.

*Corresponding author
haifei.zhang@hds.utc.fr (H. Zhang); benjamin.quost@hds.utc.fr (B. Quost); mylene.masson@hds.utc.fr (M. Masson)
https://www.hds.utc.fr/~zhanghai/ (H. Zhang); www.hds.utc.fr/~quostben/ (B. Quost); www.hds.utc.fr/~massomar/ (M. Masson)
ORCID(s): 0000-0003-4488-1631 (H. Zhang)

Walley's imprecise Dirichlet model (IDM) (Walley, 1996) is a simple yet powerful approach to propagate epistemic uncertainty, i.e. arising from a data sample being small. Assuming that we have a set of instances falling into the same leaf of a decision tree, classical inference is based on the estimated (multinomial) posterior probability distribution over the classes. In a Bayesian setting, a prior may be considered — a typical choice would be the Dirichlet distribution, being conjugate to the multinomial. The IDM rather makes use of a set of Dirichlet distributions as a prior, thus resulting in a set of posterior Dirichlet distributions after updating (Bernard, 2005). These class posterior probability intervals are as large as the amount of available data in the leaf is small.

Extensive previous research has shown the interest of applying the IDM to decision trees, in order to increase robustness or to make cautious decisions by applying an appropriate decision strategy (Troffaes, 2007). For instance, in Mantas & Abellán (2014), a minimax approach is used to determine robust splits by minimizing the highest entropy obtained over the distributions compatible with the IDM; the tree outputs are however single probability distributions, and the decision is therefore precise (it results in a single class). In Abellán & Masegosa (2012), probability intervals are obtained for each leaf node, and then used to compute the set of non-dominated classes according to a dominance criterion: the decision is therefore imprecise, as it may result in a subset of plausible classes.

Combining imprecise trees

Several works have considered combining imprecise decision trees, in order to take advantage of both the accuracy of tree ensembles and the robustness of cautious classification. For some approaches, imprecision is only considered during tree growth, as a way of increasing robustness to noise or missing data. For instance, Abellán & Masegosa (2010a) used the IDM in order to define a new split criterion; the trees are pooled using simple and weighted majority voting, resulting in precise predictions. In Abellán & Masegosa (2010b); Abellán (2013), a stacking procedure is used to select a set of trees specifically for each test instance to be classified, and class frequencies are computed over these selected trees: therefore, the tree outputs, as well as the final decision, are also precise.

Other works propose to exploit the imprecision, either by propagating it using a suitable aggregation operator, resulting in (possibly indeterminate) decisions or in imprecise probabilistic aggregates from which such decisions should then be made; or exploiting it in the aggregation procedure, for instance in order to compute tree weights (e.g. so as to alleviate the weight of uncertain trees in the combination step).

The voting strategy can be directly adapted to combining probability intervals, by first obtaining a cautious prediction for each tree (for example using interval dominance (Troffaes, 2007)) and then making a final decision by simple or weighted majority voting or minimum-against-voting (Moral-García et al., 2020). Another possibility consists in directly merging all associated probability intervals, either using disjunction or conjunction (De Campos et al., 1994), or by averaging (Murphy, 2000; Fink, 2012); a decision can then be made based on the resulting probability interval. Note that using a classical or weighted voting approach 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.

More recently, Utkin et al. (2019) proposed to compute sets of probability distributions for each tree using the imprecise pari mutuel model, and to use the resulting uncertainty in order to compute tree weights using an optimization procedure. In Utkin et al. (2020), sets of distributions are obtained using the IDM, and tree weights are learned via a maximin strategy so as to make the random forest estimates more robust. We stress out that both of these tree aggregation techniques provide precise predictions, i.e. instances are classified into a single class.

Focus of our work

We address the problem of constructing a cautious random forest by combining imprecise-probabilistic trees trained on a binary classification problem. In order to improve the effectiveness of the forest, we address both the aggregation and weight assignment strategies. We adopt the theoretical framework of belief functions (Dempster, 1967; Shafer, 1976). When evaluating a test instance, the trees are assumed to provide pieces of evidence about its actual class in the form of closed random intervals defined on $[0, 1]$. These intervals of posterior probabilities can be aggregated into belief and plausibility degrees that one of the two classes is strictly preferable to the other, degrees that can then be used in a cautious decision-making process.

Contributions of this paper

This paper builds upon preliminary work (Zhang et al., 2021), in which we described how these degrees can be calculated, a resulting tree combination strategy akin to voting was proposed, and several simple and non-adaptive weighting strategies were investigated. In the present paper, we refine this combination strategy, by proposing a specific cost function so as to automatically learn tree weights. This leads to a better compromise between the proneness of the classification system to make precise decisions and its ability to avoid making wrong decisions. To this extent, our approach can be be seen as a way of directly optimizing an utility-discounted accuracy measure (Zaffalon et al., 2012) such as u_{65}. Our strategy is also adaptive, in that its level of cautiousness can be adjusted by tuning a hyper-parameter (which corresponds to the utility of making indeterminate predictions). The cost function being difficult to optimize, we propose an upper bound, for which we provide the gradient and Hessian, and which we consequently prove to be convex. This paper finally includes a thorough experimental study which establishes the validity of our whole

approach, from the tree aggregation procedure to the weight assignment strategy. The numerous experiments realized on 25 datasets show the interest of aggregating cautious trees with our approach, particularly when the data at hand are pervaded with noise.

The paper is structured as follows. Section 2 recalls the setting and provides basic knowledge of the theoretical frameworks and the models used throughout the paper. We describe our new aggregation strategy for an ensemble of imprecise decision trees, which can be regarded as an extension of weighted voting in the imprecise classification case, in Section 3. Section 4 details our weighting strategy which aims at optimizing a compromise between determinacy and accuracy. Sections 5 presents the experiments realized on several classical datasets, noisy labeled data (aleatory uncertainty), and data with a limited training set size (epistemic uncertainty), and then discusses the results, which illustrate that our model is competitive and more adaptive to these two kinds of uncertainty compared to the baselines. Finally, a conclusion is drawn in Section 6.

2. Preliminaries

2.1. Random forests

A random forest is an ensemble learning technique based on combining decision trees; this approach is very popular due to its capability to reach excellent generalization performances and avoid overfitting issues, compared to a single decision tree. Each decision tree in a random forest is trained without pruning on a bootstrap replicate of the original training set. Training samples that are not selected for training a specific tree are called out-of-bag samples for that tree. Trees are classically grown, i.e. by determining the split which achieves the highest homogeneity (using, e.g., information gain for ID3 (Quinlan, 1986), information gain ratio for C4.5 (Quinlan, 1993), or the Gini index for CART (Breiman et al., 1984)). The main difference of a tree in a random forest, with respect to a classical tree, is that the candidate features for each split are randomly selected among all features. If a node cannot be split (i.e. homogeneity cannot be improved, the maximum depth has been reached, or the minimum node size is attained), it will be regarded as a terminal node or a leaf and used to classify test samples.

The number of candidate features for each split thus directly impacts the diversity in the tree ensemble. Besides, the minimal size of terminal nodes, or alternatively the maximal depth of the tree, make it possible to control the tree complexity and therefore its ability to fit training data (low bias). In a random forest, trees are constructed so as to have very low bias, and are consequently generally not pruned. The total number T of trees in the forest influences the variance of predictions (the larger the forest, the more stable the predictions). Combining a large number of decision trees makes it possible to exploit the diversity granted by both feature and sample randomness, and helps limiting the detrimental influence of outliers (Grandvalet, 2004), ultimately improving generalization performances.

As we mentioned above, the aggregation of the tree predictions is key in ensuring good performances of the ensemble. Given a test instance x, $\mathrm{P}(Y = y_j | X = x)$, should be estimated, for $y_j \in \Omega = \{y_1, y_2, \dots, y_K\}$, by

$$\hat{\mathrm{P}}(Y = y_j | X = x) = \frac{\sum_{t=1}^{T} w_t h_t(x, y_j)}{\sum_{t=1}^{T} w_t}, \tag{1}$$

where $h_t(x, y_j)$ is either the decision regarding class y_j, or a probability estimate for class y_j, provided by the tth tree, and w_t is the weight assigned to the tree: in simple voting or averaging, w_t is set to $1/T$ for all $t = 1, \dots, T$. Let $n_t(x, y_j)$ denote the number of training samples from class y_j that fall into the same leaf as x for tree t: the probability estimates provided by the tree are

$$h_t(x, y_j) = \frac{n_t(x, y_j)}{\sum_{j'=1}^{K} n_t(x, y_{j'})}. \tag{2}$$

The decisions used for voting can be seen as a crude, rounded version of these probability estimates:

$$h_t(x, y_j) = \mathbb{1}(n_t(x, y_j) > n_t(x, y_{j'}), \forall y_{j'} \neq y_j). \tag{3}$$

In averaging, once the averaged probability has been computed using Equation (1), the decision can be made by picking the class with highest estimated posterior probability.

2.2. Imprecise Dirichlet model

Let $\Omega = \{y_1, y_2, \dots, y_K\}$ be the aforementioned set of $K \geq 2$ classes, or more generally mutually exclusive categories, and let $\pi_j = \mathrm{P}(y_j)$, with $\pi_j \geq 0$ and $\sum_{j=1}^{K} \pi_j = 1$, for $j = 1, \dots, K$. Assume that N iid observations have been sampled from a unknown multinomial distribution $\mathcal{M}(N; \pi_1, \dots, \pi_K)$: let n_j denote the corresponding number of occurrences of y_j, with $\sum_{j=1}^{K} n_j = N$. For the sake of simplicity, we write $\boldsymbol{n} = \{n_1, \dots, n_K\}$ and $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$. The likelihood of the parameter vector writes as

$$L(\boldsymbol{\pi} | \boldsymbol{n}) \propto \prod_{j=1}^{K} \pi_j^{n_j}. \tag{4}$$

In a standard Bayesian setting, prior knowledge over the probabilities π_j can be specified using the conjugate Dirichlet distribution $Dir(s, \boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ and $\sum_{j=1}^{K} \alpha_j = s$:

$$\Pr(\boldsymbol{\pi} | \boldsymbol{\alpha}) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j - 1}; \tag{5}$$

Note that each parameter can be decomposed into $\alpha_j = s\, t_j$, with $s \geq 0$, $0 \leq t_j \leq 1$, and $\sum_j t_j = 1$: then, the parameters t_js are the prior frequencies, with $\mathbb{E}(\pi_j) = t_j$; whereas s corresponds to the prior's global strength. The posterior distribution then writes as

$$\Pr(\boldsymbol{\pi} | \boldsymbol{n}, \boldsymbol{\alpha}) \propto \prod_{j=1}^{K} \pi_j^{n_j + s t_j - 1}, \tag{6}$$

which is a Dirichlet distribution due to conjugacy.

In standard Bayesian inference, the parameters s and $\boldsymbol{t}$ are determined in advance, which results in point estimates for the π_j. However, in the imprecise Dirichlet model (IDM) (Walley, 1996), a set of Dirichlet distributions is defined by considering all vectors $\boldsymbol{t}$ satisfying the constraints $0 \leq t_j \leq 1$ and $\sum_{j=1}^{K} t_j = 1$. Taking this set as a prior amounts to making as few assumptions as possible regarding $\boldsymbol{\pi}$, i.e. the prior is near-ignorance (Mangili & Benavoli, 2015). As a result, the posterior information is no longer a single distribution, but a set of distributions, from which we can now deduce lower and upper bounds on the probabilities $(\boldsymbol{\pi}|\boldsymbol{n}, s)$, reached respectively for $t_j \to 0$ and $t_j \to 1$:

$$\underline{\mathrm{E}}(\pi_j|n) = \frac{n_j}{N+s}, \quad \overline{\mathrm{E}}(\pi_j|n) = \frac{n_j+s}{N+s}. \tag{7}$$

Note that the parameter s remains to be chosen in advance: it can be interpreted as the number of virtual instances with unknown class information. Although several studies have been conducted with regard to choosing an appropriate value (Abellán et al., 2006), this problem remains open. In practice, values of $s = 1$ or $s = 2$ are often picked, following Walley (1996).

2.3. Theory of belief functions

The theory of belief functions also referred to as the theory of evidence or Dempster–Shafer theory (Dempster, 1967; Shafer, 1976), provides a general framework for modeling and reasoning with uncertainty. Let $\Omega = \{y_1, y_2, \ldots, y_K\}$ be a finite set that contains all the possible, mutually exclusive values for a variable Y of interest, referred to as the frame of discernment. A mass function is a mapping $m : 2^{\Omega} \to [0,1]$, such that $\sum_{A\subseteq\Omega} m(A) = 1$. The value $m(A)$ measures the degree of evidence supporting $Y \in A$ only, but nothing else. The constraint $m(\emptyset) = 0$ is also often required. A subset A of Ω is called a focal element if $m(A) > 0$. If there is only one such subset $A \subseteq \Omega$, then m is said to be logical; and if furthermore $A = \Omega$, m is vacuous (it represents total ignorance). A mass function is Bayesian if $|A| = 1$ for all A such that $m(A) > 0$. This framework can therefore be seen as an extension of both sets theory and classical probability theory.

Belief and plausibility functions can be computed from the mass function m: they are respectively defined as

$$Bel(A) = \sum_{B\subseteq A} m(B), \quad Pl(A) = \sum_{B\cap A\neq\emptyset} m(B), \tag{8}$$

for all $A \subseteq \Omega$. Then, $Bel(A)$ measures the total degree of support to A, and $Pl(A)$ measures the degree of support that could be transferred to A, provided that further evidence supporting this transfer became available. These functions are dual since $Bel(A) = 1 - Pl(\overline{A})$, with $\overline{A}$ the complement of A. It should be noted that mass, belief, and plausibility functions are equivalent as they can be retrieved from each other.

The theory of belief functions can also be extended to infinite frames of discernment (Denœux, 2009). Let U and V be two random variables such that $U \leq V$; they may be viewed as determining a random interval $[U, V]$ defining a belief and plausibility function on $\mathbb{R}$:

$$\begin{aligned} Bel(A) &= \Pr([U,V] \subseteq A), \\ Pl(A) &= \Pr([U,V] \cap A \neq \emptyset), \end{aligned} \tag{9}$$

for any element A of the Borel sigma-algebra $\mathcal{B}(\mathbb{R})$ of the real line. Let $I_i = [u_i, v_i]$, $i = 1, \ldots, n$ be a collection of N intervals, and let $m : \mathcal{I} \to [0,1]$ be a mass function defined on the set $\mathcal{I}$ of closed real intervals of $[0,1]$ such that $m(I_i) = m_i$ with $i = 1, \ldots, n$ and $\sum_{i=1}^{n} m_i = 1$. Under this setting, the belief and plausibility functions of an event A are

$$Bel(A) = \sum_{I_i\subseteq A} m_i, \quad Pl(A) = \sum_{I_i\cap A\neq\emptyset} m_i. \tag{10}$$

The intervals I_i are called focal intervals of m (Denœux, 2009). In the case of a random forest, this definition provides a basis for pooling pieces of information regarding the class posterior probabilities provided by the trees.

3. Cautious Random Forests

We now present our strategy for aggregating imprecise probabilistic tree outputs. Our approach is akin to that proposed by Abellán & Masegosa (2012), except that out aggregation operator can be seen as a kind of voting. In this paper, we focus on binary classification problems. We consider a training set composed of N pairs of examples (x_i, y_i) with class labels $y_i \in \{0, 1\}$, where $i = 1, \ldots, N$. The probability that instance x_i belongs to category 1 (respectively, 0) is written $p^1(x_i)$ (resp., $p^0(x_i)$).

The cautious random forest is composed of T decision trees $f_1, \ldots, f_t, \ldots, f_T$, trained here using the CART algorithm without pruning. Each tree divides the feature space into regions associated with its leaves. A sample x is thus associated with a set of regions $R_1(x), \ldots, R_t(x), \ldots, R_T(x)$, with $R_t(x)$ the region into which x falls for tree f_t. This region contains $n_t^0(x)$ and $n_t^1(x)$ training instances from classes 0 and 1, respectively: in a classical setting, these numbers are used to estimate the class posterior probabilities defined by Equation (1), using either (2) or (3), before a decision is made. Note that this approach degenerates into simple averaging or majority voting whenever $w_t = 1/T$ for all $t = 1, \ldots, T$. Obviously, the reliability of an individual estimate (or decision) provided by a tree strongly depends on the sample size $N_t(x) = n_t^0(x) + n_t^1(x)$ in the leaf attained by x, and might therefore differ from the actual probability for some small leaves (e.g. with only one or two samples).

In order to reflect epistemic uncertainty (i.e., the lack or information at the tree leaf level), the IDM can be used to produce interval-valued probability estimates, the size of which will decrease according to the amount $N_t(x)$ of training instances in $R_t(x)$: for the positive class,

$$I_t(x) = \left[\underline{p}_t^1(x), \overline{p}_t^1(x)\right] = \left[\frac{n_t^1(x)}{N_t(x)+s}, \frac{n_t^1(x)+s}{N_t(x)+s}\right], \tag{11}$$

with $\underline{p}_t^1(x)$ and $\overline{p}_t^1(x)$ the lower and upper bounds of $p_t^1(x)$ (their counterparts for the alternative class can be retrieved by duality). In order to aggregate these probability intervals (i.e., at the forest level), we propose to calculate the belief and plausibility of the event "$p^1(x) \in [0.5, 1]$", which quantify the available evidence regarding the proposition that instance x belongs to class 1. Using Equation (10),

$$\begin{aligned} bel^1(x) &= bel(p^1(x) \in [0.5, 1]) \\ &= \sum_{t=1}^{T} w_t \mathbb{1}(\underline{p}_t^1(x) \geq 0.5), \end{aligned} \tag{12}$$

and

$$\begin{aligned} pl^1(x) &= pl(p^1(x) \in]0.5, 1]) \\ &= \sum_{t=1}^{T} w_t \mathbb{1}(\overline{p}_t^1(x) > 0.5); \end{aligned} \tag{13}$$

here, the weight of the tree in the aggregation process (following the notation in Equation (1)) actually corresponds to the degree of support $m(I_t(x))$ to each interval $I_t(x)$ provided by the tree (which is regarded as a focal element on the unit interval $[0, 1]$).

The proposed tree aggregation approach can thus be seen as a generalized voting mechanism: instead of voting for point probabilities, each tree votes for probability intervals and also produces interval-valued probability estimates, which can in turn be used to make imprecise predictions. A natural choice is $m(I_t(x)) = 1/T$, for all $t = 1, \ldots, T$. Several other mass assignment methods on the leaf level were studied in (Zhang et al., 2021), for instance based on the level of epistemic uncertainty in leaves.

We remark that by duality, $bel^0(x) = 1 - pl^1(x)$ and $pl^0(x) = 1 - bel^1(x)$. Based on the final interval $[bel^1(x), pl^1(x)]$, the interval dominance decision rule can be applied to make a decision:

$$\hat{y} = \begin{cases} 1, \text{ if } bel^1(x) \geq 0.5, \\ 0, \text{ if } pl^1(x) < 0.5; \\ \{0, 1\}, \text{ otherwise.} \end{cases} \tag{14}$$

Algorithm 1 describes the inference process of our cautious random forest strategy.

Algorithm 1: Cautious random forest predictions

Input: random forest RF, tree weights w_t, IDM parameter s, set of test instances X
Output: predictions $\hat{Y}$ for test instances

```
Ŷ ← {}
for x_i ∈ X do
    for f_t ∈ RF do
        Compute I_t(x_i) via Eq. (11)
    Calculate bel_i^1 via Eq. (12)
    Calculate pl_i^1 via Eq. (13)
    if bel_i^1 ≥ 0.5 then
        ŷ_i ← 1
    else if pl_i^1 < 0.5 then
        ŷ_i ← 0
    else
        ŷ_i ← {0, 1}
    Ŷ ← Ŷ ∪ ŷ_i
```

4. Learning tree weights

In this section, we investigate assigning weights to trees in our combination scheme. As in Utkin et al. (2019, 2020), we propose to automatically learn the tree weights w_t so as to optimize the tree ensemble performances. However, to our knowledge, all existing approaches (Li et al., 2010; Kim et al., 2011; Caruana et al., 2004; Utkin et al., 2019, 2020) are based on tree accuracy, and are therefore not well-suited to our imprecise classification setting, since they would amount to give indeterminate predictions the same status as faults. We propose here to make use of a cautious criterion, which rewards both the cautiousness (associated with indeterminate predictions) and the accuracy (associated with accurate determinate predictions) of the ensemble. In spirit, optimizing this criterion so as to determine tree weights amounts to replacing the classically optimized accuracy measure with a utility-discounted accuracy metric.

Let us define

$$\begin{aligned} \boldsymbol{w} &= \left(w_1 \ldots w_T\right)^\top, \\ \underline{\boldsymbol{\delta}}(x) &= \left(\mathbb{1}(\underline{p}_1^1(x) \geq 0.5) \ldots \mathbb{1}(\underline{p}_T^1(x) \geq 0.5)\right)^\top, \\ \overline{\boldsymbol{\delta}}(x) &= \left(\mathbb{1}(\overline{p}_1^1(x) > 0.5) \ldots \mathbb{1}(\overline{p}_T^1(x) > 0.5)\right)^\top. \end{aligned}$$

Here, $\boldsymbol{w}$, $\underline{\boldsymbol{\delta}}(x)$ and $\overline{\boldsymbol{\delta}}(x)$ are all column vectors of T elements, with $\boldsymbol{w}$ the vector of variables to be identified. Using these notations, Equations (12) and (13) can be rewritten as

$$bel^1(x) = \boldsymbol{w}^\top \underline{\boldsymbol{\delta}}(x), \quad pl^1(x) = \boldsymbol{w}^\top \overline{\boldsymbol{\delta}}(x). \tag{15}$$

Note that the vectors $\underline{\boldsymbol{\delta}}(x)$ and $\overline{\boldsymbol{\delta}}(x)$ of binary values are constant once the random forest has been trained. Remark also that the duality property holds: $bel^0(x) = 1 - pl^1(x)$, and $pl^0(x) = 1 - bel^1(x)$. In the following, for the sake of simplicity, we will write $bel_i^1 = bel^1(x_i)$, $pl_i^1 = pl^1(x_i)$, $\underline{\boldsymbol{\delta}}_i = \underline{\boldsymbol{\delta}}(x_i)$ and $\overline{\boldsymbol{\delta}}_i = \overline{\boldsymbol{\delta}}(x_i)$, for any training instance x_i.

We may naturally define an optimization criterion based on the log-loss:

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{w}) = &- \frac{1}{N} \sum_{i=1}^{N} \left\{ y_i \ln(bel_i^1) + (1 - y_i) \ln(bel_i^0) \right\} \\ &+ \lambda \parallel \boldsymbol{w} \parallel_2^2, \\ &s.t. \sum_{t=1}^{T} w_t = 1, \quad w_t \geq 0, \quad \forall t = 1, \ldots, T. \end{aligned} \tag{16}$$

Note that a similar cost function was introduced in (Utkin et al., 2019) as

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{w}) = &- \frac{1}{N}\sum_{i=1}^{N}\left\{y_i\, bel_i^1 + (1-y_i)\, bel_i^0\right\} + \lambda \parallel \boldsymbol{w} \parallel_2^2 \\ &s.t.\ \sum_{t=1}^{T} w_t = 1,\ \frac{1-\epsilon}{T} \le w_t \le \frac{1-\epsilon}{T} + \epsilon,\ \forall t. \end{aligned} \tag{17}$$

While (16) is akin to a cross-entropy loss, (17) can be regarded as a kind of hinge loss; both are convex. However, both methods tend to produce determinate predictions, since indeterminate predictions are penalized as errors. In a cautious setting, the cost of an indeterminate prediction should be lower than that of a determinate, erroneous one.

We therefore propose to optimize a cost function which considers both determinate and indeterminate predictions:

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{w}) = &\frac{1}{N}\sum_{i=1}^{N}\left\{y_i H(0.5 - bel_i^1) + (1-y_i)H(pl_i^1 - 0.5)\right. \\ &\left. -\gamma H((0.5 - bl_i^1)(pl_i^1 - 0.5))\right\}, \end{aligned} \tag{18}$$

where $H(\cdot)$ is the Heaviside function. Using this cost function, determinate predictions cost nothing if they are correct, and are penalized (cost 1) if they are wrong; all indeterminate predictions cost $1-\gamma$. Optimizing this cost function amounts to look for a compromise between making precise predictions and avoiding mistakes. To this extent, the criterion in Eq. (18) can be seen as a utility-discounted accuracy measure (Zaffalon et al., 2012). The parameter γ can be considered as the utility of being indeterminate, which can be tuned to adjust the cautiousness of the model (the larger the value of gamma, the more cautious the model). For example, consider an instance x_i with actual label $y_i = 1$: should the model return $bel_i^1 = 0.1$ and $pl_i^1 = 0.2$, the prediction would be $\hat{y}_i = 0$ (wrong), with a cost equal to 1; conversely, with $bel_i^1 = 0.8$ and $pl_i^1 = 0.9$, the prediction would be 1 (correct) and cost 0. Eventually, with $bel_i^1 = 0.4$ and $pl_i^1 = 0.6$, the indeterminate prediction $\hat{y}_i = \{0,1\}$ would cost $1-\gamma$.

Since the Heaviside function is neither continuous nor differentiable, we propose to use the sigmoid function as an approximation:

$$H(x) \approx U(x) = \frac{1}{1+\exp(-\alpha x)}; \tag{19}$$

the approximation is reasonable if α is large enough. The sigmoid function being nonconvex, this cost function is prone to local minima. A solution to this issue consists in minimizing a surrogate (upper bound) $\boldsymbol{J}_{\sup}(\boldsymbol{w})$ for $\boldsymbol{J}(\boldsymbol{w})$ (Dmochowski et al., 2010). Using the inequality $z \le -\ln(1-z),\ \forall z < 1$, the equality $U(-x) = 1 - U(x)$ and $U(x) < 1,\ \forall x \in \mathbb{R}$, we have

$$U(0.5 - bel_i^1) \le -\ln(U(bel_i^1 - 0.5)),$$

$$U(pl_i^1 - 0.5) \le -\ln(1 - U(pl_i^1 - 0.5)),$$

and

$$\begin{aligned} &-U\left((0.5 - bl_i^1)(pl_i^1 - 0.5)\right) \le \\ &\quad -\ln\left(1 - U\left((bel_i^1 - 0.5)(pl_i^1 - 0.5)\right)\right) - 1. \end{aligned}$$

Remarking that a regularization term should be taken into account in the cost function, so as to avoid overfitting, we finally obtain the following regularized upper bound:

$$\begin{aligned} \boldsymbol{J}_{\sup}(\boldsymbol{w}) = &-\frac{1}{N}\sum_{i=1}^{N}\left\{ y_i \ln\left(U(\boldsymbol{w}^\top \underline{\boldsymbol{\delta}}_i - 0.5)\right)\right. \\ &+ (1-y_i)\ln\left(1 - U(\boldsymbol{w}^\top \overline{\boldsymbol{\delta}}_i - 0.5)\right) \\ &\left.+\gamma \ln\left(1 - U((\boldsymbol{w}^\top \underline{\boldsymbol{\delta}}_i - 0.5)(\boldsymbol{w}^\top \overline{\boldsymbol{\delta}}_i - 0.5))\right)\right\} \\ &+ \frac{1}{2}\lambda \parallel \boldsymbol{w} \parallel_2^2 \\ &s.t.\ \sum_{t=1}^{T} w_t = 1, \quad w_t \ge 0,\ \forall t = 1, \dots, T. \end{aligned} \tag{20}$$

In Eq. (20), the first and the second terms within the summation correspond to the penalty incurred for not assigning an instance to the right class; however, should the classification fail because of an indeterminate decision, this penalty would be compensated by the third term of the summation (generally partially, depending on the γ value). The last term out of the summation is a regularization term to avoid overfitting. The error criterion defined by Eq. (20) is continuous and convex, as shown in Appendices A and B. It can therefore be easily minimized using any convex optimization solver.

5. Experiments

In this section, we detail the experiments conducted to show the interest of our proposed approach. The experiments were conducted on 25 public datasets from the UCI repository (Bache & Lichman, 2013), of which Table 1 provides a summary. They all correspond to binary classification problems, and cover a large range of sample sizes and number of features. Section 5.1 introduces the performance criteria used to assess the quality of the imprecise classification results, and the tests applied to compare multiple models over multiple datasets. Then, experiments are reported in two steps:

- in Section 5.2, the different tree aggregation strategies providing cautious predictions are compared on normal data, noisy data and small training data;
- Section 5.3 illustrates the advantage of our proposed strategy for learning tree weights compared to other weight assignment methods, and studies the influence of the hyper-parameter tuning the compromise between informativeness and cautiousness.

Table 1
Datasets used in the experiments, with abbreviation ABB, numbers of instances (N) and of features (nominal/numerical).

Dataset	ABB	N	Feat	Nom	Num
adult	ADT	45222	11	0	11
banknote	BKT	1372	4	0	4
biodeg	BID	1053	41	0	41
breast-cancer	BRC	568	30	0	30
cardiac	CAD	889	12	0	12
compas	COP	2652	6	0	6
credit	CRD	690	15	9	6
diabetes	DIB	768	8	0	8
german	GER	1000	24	0	24
heart	HRT	303	13	0	13
heloc	HLC	10459	23	0	23
ionosphere	INS	351	34	0	34
liver	LIV	345	6	0	6
magic	MGC	2300	57	0	57
mammographic	MMG	830	5	0	5
occupancy	OCP	2665	6	1	5
phishing	PHS	11054	30	0	30
pima	PMA	768	8	0	8
post-operative	POP	88	8	7	1
ringnorm	RNO	7400	20	0	20
seismic	SSC	2584	18	4	14
sonar	SNR	208	60	0	60
spam	SPM	4594	57	0	57
vote	VTE	435	16	16	0
wine	WNE	1599	11	0	11

5.1. Performance measures

Measuring the performance of a cautious (imprecise) classification system should reward both its accuracy and its determinacy, between which the imprecise classifier should achieve a reasonable compromise. When making precise decisions is paramount, errors being acceptable and/or human interventions being too expensive, determinacy primes over accuracy; the reverse corresponds to scenarios where errors should be avoided by all means, such as e.g. in medical diagnosis or autonomous driving.

Discounted accuracy is frequently used as a global evaluation metric for imprecise classifiers. Let $U(x)$ stand for the set of classes predicted for instance x. In a nutshell, discounted accuracy rewards a cautious prediction with $1/|U(x)|$ iff $U(x) \ni y$. Set accuracy rather measures the proportion of instances for which the prediction (be it indeterminate) contains the actual class. Therefore, in the binary case, whereas discounted accuracy rewards each indeterminate prediction with $1/2$ (which statistically corresponds to choosing at random between the two classes), set accuracy rewards them with 1 (as for a correct precise prediction).

As emphasized by Zaffalon et al. (2012), both are therefore not appropriate when it comes to evaluate cautious predictions in a binary setting; this problem may be overcome by using utility-discounted predictive accuracy measures, the two most popular of which are u_{65} and u_{80}:

$$\begin{aligned} u_{65}(z) &= -0.6z^2 + 1.6z, \\ u_{80}(z) &= -1.2z^2 + 2.2z, \end{aligned} \tag{21}$$

with $z = 1/|U(x)|$.

In binary classification, u_{65} rewards an indeterminate prediction with 0.65, and u_{80} with 0.80. Note that using this notation, discounted accuracy corresponds to u_{50}, and set accuracy to u_{100}. Since utility-discounted accuracy makes an assumption regarding the cost of indeterminate predictions, we will also consider in our experiments cautiousness (cau) and single-set accuracy (ssa) as performance evaluation metrics, in addition to u_{65} and u_{80}. Whereas the latter measures the proportion of instances correctly classified among those precisely classified, the former indicates the proportion of indeterminate predictions. These four indicators thus give a rather complete overview of the properties of the binary cautious classifiers evaluated.

In order to compare multiple models over multiple datasets, we followed the recommendation of Demšar (2006). First, the Friedman test (Friedman, 1940) is performed to determine whether all of the algorithms are equivalent or not; this non-parametric test scores the algorithms independently for each data set. The top performing algorithm receives a rank of 1, the second best algorithm receives a rank of 2, and so on. If the null hypothesis (all algorithms are equivalent) is rejected, a Nemenyi test (Nemenyi, 1963) can be used in a second step to identify significant differences.

5.2. Phase 1: tree aggregation procedure

Models compared

In this first phase of experiments, we benchmark different tree aggregation strategies in random forests, all tree weights being considered as equal. The methods compared are:

- AVE: AVErage, where, following Murphy (2000) and Fink (2012), we average the lower and upper probabilities provided by the trees at hand, i.e. $bel^1(x) = \frac{1}{T}\sum_{t=1}^{T} \underline{p}_t^1(x)$ and $pl^1(x) = \frac{1}{T}\sum_{t=1}^{T} \overline{p}_t^1(x)$, before applying interval dominance (14);
- MV: Majority Voting is adapted to our imprecise classification setting, by applying interval dominance to each tree, and considering indeterminate predictions $\{0, 1\}$ as a possible outcome when counting the votes (Fink, 2012);
- MVTH: in Majority Voting with THresholding, we first estimate the probability $p^1(x)$ of class 1 as the number of trees providing a probability $p_t^1(x) \geq 0.5$, and we predict class 1 whenever $p^1(x) > 0.5+\theta$, class 0 whenever $p^1(x) < 0.5-\theta$, and $\{0, 1\}$ otherwise (with θ being chosen arbitrarily);
- MVA: Minimum Vote Against counts the number of classifiers that predict a class as dominated (vote against), the final non-dominated set of classes being

Table 2
Comparison of aggregation strategies: metrics evaluated on each dataset (without label noise), and counts for each strategy giving the highest and lowest scores.

(a) Cautiousness

Data	AVE	MV	MVTH	MVA	CRF
ADT	13.43	15.79	4.82	0.26	**18.66**
BKT	**0.40**	0.03	0.37	0.02	0.26
BID	8.69	0.93	5.64	0.25	**10.32**
BRC	**2.36**	0.05	1.88	0.07	1.51
CAD	3.98	**8.73**	1.62	0.18	8.05
COP	35.58	31.03	8.91	0.72	**37.40**
CRD	10.67	8.15	4.48	0.17	**13.32**
DIB	18.30	2.59	9.74	0.50	**20.88**
GER	27.32	15.47	11.82	0.56	**33.13**
HRT	16.31	4.98	7.60	0.56	**20.07**
HLC	19.61	1.41	11.45	0.46	**21.90**
INS	2.45	0.23	1.77	0.00	**3.42**
LIV	**27.36**	0.46	13.50	0.58	17.56
MGC	**3.98**	0.28	2.39	0.13	2.95
MMG	14.65	**27.28**	3.40	0.24	25.03
OCP	0.58	0.67	0.30	0.02	**0.94**
PHS	**6.18**	1.86	2.42	0.14	5.63
PMA	18.59	2.40	10.29	0.48	**21.08**
POP	29.53	**39.14**	6.25	0.35	29.04
RNO	4.99	0.32	4.70	0.22	**5.41**
SSC	**1.98**	0.06	1.13	0.01	1.29
SNR	**18.48**	0.77	14.47	0.67	11.64
SPM	**3.71**	0.20	2.48	0.05	2.61
VTE	**3.95**	1.20	1.42	0.09	3.70
WNE	**14.42**	0.61	8.30	0.37	10.01
Average	12.30	6.59	5.65	0.28	**13.03**
#Highest	10	3	0	0	**12**
#Lowest	0	2	0	**23**	0

(b) Single-set accuracy

Data	AVE	MV	MVTH	MVA	RF	CRF
ADT	87.79	88.93	84.63	84.18	83.60	**89.73**
BKT	99.42	99.29	99.46	99.36	**99.63**	99.37
BID	90.10	87.36	89.19	87.02	87.04	**90.94**
BRC	**97.03**	95.90	96.87	96.04	96.02	96.93
CAD	78.98	79.75	78.42	77.90	77.82	**79.94**
COP	64.57	61.77	60.61	59.66	60.11	**64.64**
CRD	91.18	90.01	89.22	87.18	87.51	**92.15**
DIB	81.24	77.25	79.14	76.77	76.36	**82.26**
GER	83.51	79.47	79.63	76.26	77.28	**84.78**
HRT	87.30	84.11	84.67	82.75	82.66	**88.47**
HLC	74.59	71.01	72.75	70.87	70.75	**75.15**
INS	94.47	93.56	94.27	93.53	93.34	**94.81**
LIV	**78.65**	73.74	76.37	74.22	73.68	77.05
MGC	**95.92**	94.48	95.57	93.35	94.62	95.89
MMG	84.94	**88.00**	80.24	81.30	79.89	87.49
OCP	98.78	98.88	98.61	98.62	**99.09**	98.97
PHS	**96.39**	94.59	95.19	94.28	94.84	96.11
PMA	81.02	76.94	79.06	76.48	76.24	**81.79**
POP	67.24	62.98	65.14	65.00	65.05	**67.91**
RNO	**95.16**	93.13	95.09	93.27	93.72	95.03
SSC	**93.87**	93.25	93.64	93.25	93.76	93.73
SNR	**89.00**	83.27	88.26	83.40	84.69	87.58
SPM	**95.82**	94.49	95.43	94.42	95.00	95.37
VTE	97.53	96.40	96.40	96.27	95.86	**97.68**
WNE	**86.05**	82.32	84.88	82.25	82.43	85.17
Average	87.62	85.64	86.11	84.71	84.84	**87.96**
#Highest	9	1	0	0	2	**13**
#Lowest	0	6	1	9	**10**	0

(c) u_{65} score

Data	AVE	MV	MVTH	MVA	CRF
ADT	84.72	**85.13**	83.67	84.13	85.09
BKT	99.29	99.28	99.33	**99.35**	99.29
BID	87.93	87.14	87.82	86.96	**88.26**
BRC	96.27	95.88	96.26	96.02	**96.44**
CAD	78.38	78.43	78.18	77.88	**78.72**
COP	64.70	62.72	61.00	59.70	**64.74**
CRD	88.43	88.00	88.16	87.14	**88.57**
DIB	78.23	76.91	77.73	76.71	**78.62**
GER	**78.38**	77.18	77.86	76.19	78.14
HRT	83.66	83.16	83.21	82.64	**83.75**
HLC	72.69	70.93	71.85	70.85	**72.92**
INS	93.73	93.48	93.74	93.53	**93.77**
LIV	**74.89**	73.70	74.73	74.14	74.87
MGC	94.68	94.40	94.83	93.31	**94.98**
MMG	**82.00**	81.72	79.71	81.27	81.86
OCP	98.58	**98.65**	98.51	98.62	**98.65**
PHS	94.44	94.03	**94.46**	94.23	94.34
PMA	78.00	76.66	77.60	76.42	**78.22**
POP	65.86	63.02	65.11	65.02	**66.17**
RNO	93.65	93.04	**93.67**	93.21	93.40
SSC	93.30	93.24	93.31	93.25	**93.36**
SNR	84.53	83.12	84.86	83.28	**84.90**
SPM	**94.67**	94.43	**94.67**	94.41	94.57
VTE	96.25	96.02	95.96	96.24	**96.48**
WNE	83.01	82.21	**83.22**	82.19	83.15
Average	85.61	84.90	85.18	84.67	**85.73**
#Highest	4	2	4	1	**16**
#Lowest	0	9	4	**12**	0

(d) u_{80} score

Data	AVE	MV	MVTH	MVA	CRF
DT	86.74	87.50	84.40	84.17	**87.89**
BKT	99.35	99.29	**99.38**	99.35	99.32
BID	89.23	87.28	88.66	87.00	**89.81**
BRC	96.62	95.89	96.54	96.03	**96.67**
CAD	78.98	79.74	78.43	77.91	**79.93**
COP	70.04	67.38	62.34	59.81	**70.35**
CRD	90.03	89.23	88.83	87.17	**90.57**
DIB	80.97	77.30	79.19	76.78	**81.76**
GER	82.48	79.50	79.64	76.28	**83.11**
HRT	86.11	83.91	84.35	82.73	**86.76**
HLC	75.63	71.15	73.57	70.92	**76.21**
INS	94.10	93.52	94.01	93.53	**94.28**
LIV	**78.99**	73.77	76.75	74.23	77.51
MGC	95.28	94.44	95.19	93.33	**95.42**
MMG	84.20	**85.81**	80.22	81.30	85.61
OCP	98.67	98.75	98.56	98.62	**98.79**
PHS	**95.37**	94.31	94.83	94.26	95.19
PMA	80.79	77.02	79.14	76.49	**81.38**
POP	70.29	68.89	66.04	65.07	**70.52**
RNO	**94.40**	93.09	94.38	93.24	94.21
SSC	**93.60**	93.25	93.48	93.25	93.55
SNR	**87.30**	83.23	87.03	83.38	86.64
SPM	**95.23**	94.46	95.05	94.42	94.96
VTE	96.84	96.20	96.18	96.25	**97.03**
WNE	**85.17**	82.30	84.47	82.24	84.65
Average	87.46	85.89	86.03	84.71	**87.68**
#Highest	7	1	1	0	**16**
#Lowest	0	7	3	**16**	0

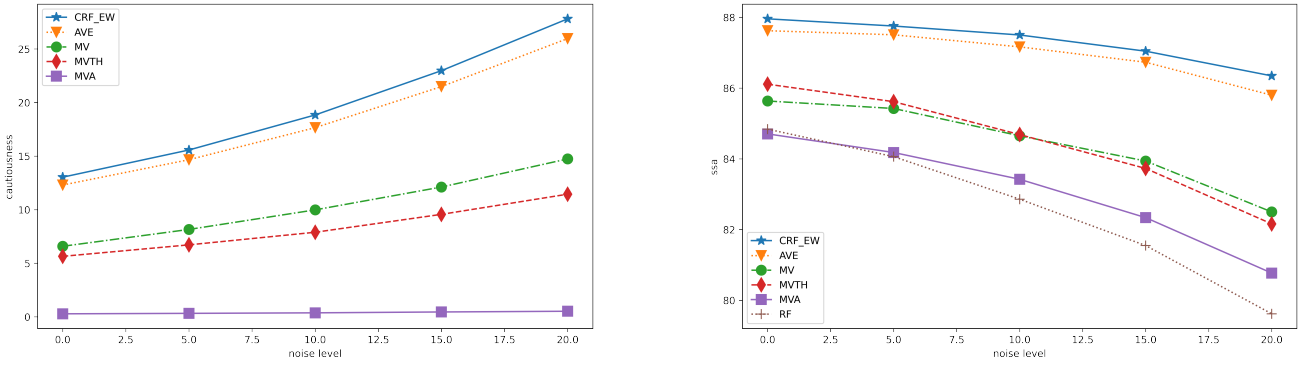
Table 3
Phase 1 (noise-free data): Friedman statistic and p-value (left), Nemenyi p-values for pairwise model comparison (right).

(a) Friedman rank and test

	AVE	MV	MVTH	MVA	CRF	p-value
cau	1.96	3.08	3.48	4.48	**1.64**	5.21E-08
ssa	2.08	3.08	3.84	4.44	**1.56**	7.99E-09
u65	2.36	2.96	3.80	4.12	**1.76**	1.74E-07
u80	2.16	3.28	3.68	4.40	**1.48**	9.01E-09

(b) Nemenyi test

CRF	vs. AVE	vs. MV	vs. MVTH	vs. MVA
cau	0.90	0.001	0.007	0.001
ssa	0.90	0.001	0.005	0.001
u65	0.49	0.001	0.020	0.001
u80	0.90	0.001	0.002	0.001



(a) Cautiousness

(b) Single-set Accuracy

(c) u_{65} score

(d) u_{80} score

Figure 1: Average cautiousness, single-set accuracy, u_{65}, and u_{80} scores computed over all datasets, as a function of label noise.

made of the classes with the lowest amount of votes against (Moral-García et al., 2020);

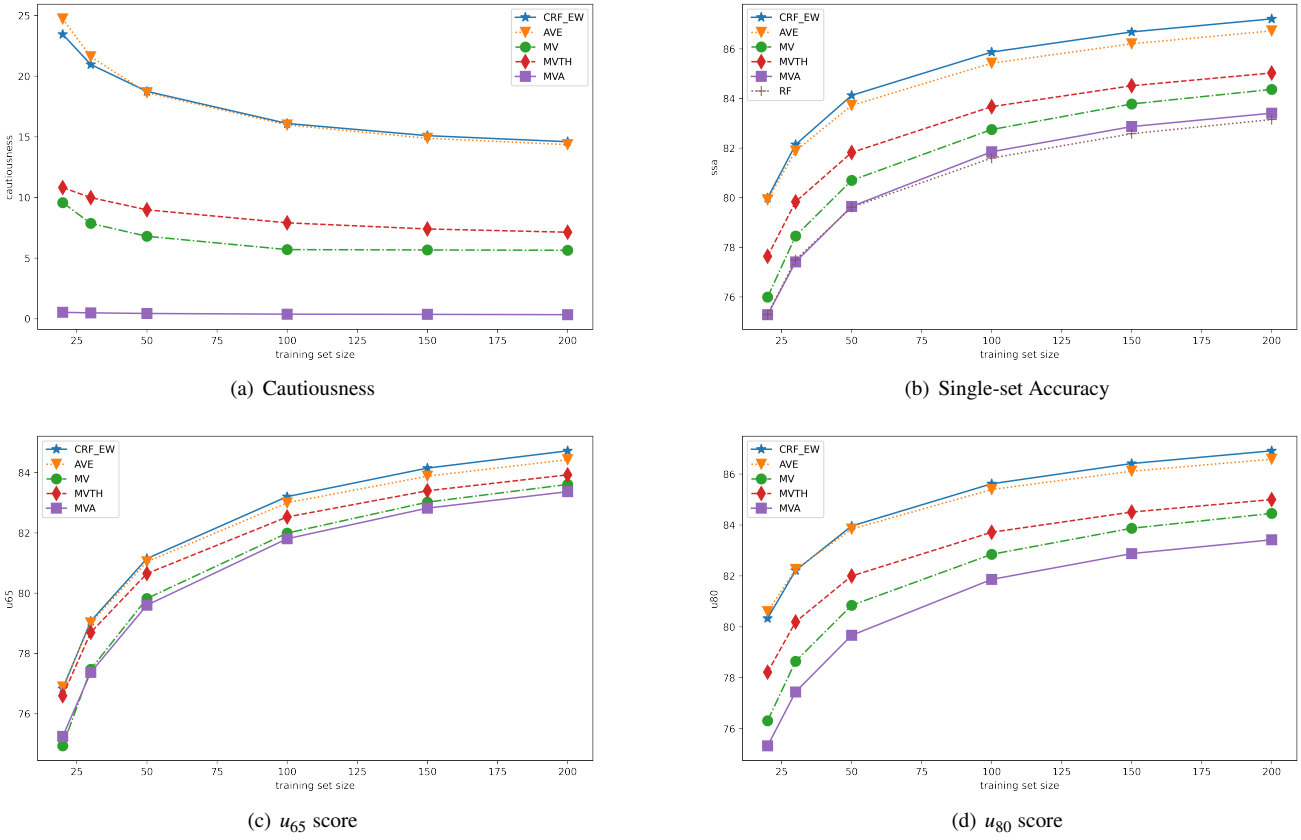
- CRF: our proposed cautious random forest strategy, where we first pool the trees by computing the belief and plausibility degrees according to Equations (12)-(13) (with equal tree weights), before applying interval dominance (14).

Experimental setting and protocol

The experiments were realized using the Random Forest classifier from the Scikit-Learn (Pedregosa et al., 2011) Python library. Each tree in the ensemble is trained to its full depth, i.e. the minimum number of training samples allowed in a leaf is one. Since the library made it possible to handle numeric features only, all categorical features were converted by one-hot encoding. The forest consists of $T = 100$ trees.

We implemented the following protocol to compare the aggregation strategies. For each dataset, for our method (CRF) we selected by cross-validation the value of the IDM parameter s which maximizes u_{65} score; we used the same s for the MV strategy. For AVE, the value was fixed to $s = 1$, following the recommendations in (Walley, 1996). For MVTH, the threshold was set to $\theta = 0.05$ for all datasets.

Tests have been carried out in three directions. First, we applied our protocol to the standard UCI datasets. In a second step, we introduced noise in the training data by flipping a fixed proportion of labels drawn at random. In the experiments, we considered various levels of label noise (0%, 5%, 10%, 15%, 20%). Average cautiousness, single-set accuracy, u_{65} and u_{80} values were computed by averaging the



(a) Cautiousness

(b) Single-set Accuracy

(c) u_{65} score

(d) u_{80} score

Figure 2: Average cautiousness, single-set accuracy, u_{65} and u_{80} scores computed over all datasets, as a function of training set size.

measures made on ten repetitions of 10-fold cross-validation according to the selected parameters. Last, we studied the effect of the training set size on the results. For different sizes of the training set ($N \in \{20, 30, 50, 100, 150, 200\}$), each metric was computed by averaging 100 independent repetitions according to the selected parameters. The training samples were randomly selected from the whole dataset and the remaining ones were used as test set.

Results and discussion

First, we discuss on the results obtained by applying the methods on standard datasets, which are reported in Tables 2(a) to 2(d). As can be seen from Table 2(a), CRF appears to be the most cautious of all models and yields very similar results to AVE. MVA is the least cautious on all datasets, reaching cautiousness less than 1%.

All cautious classifiers outperform the precise random forest (RF) — often by a significant amount — in terms of single-set accuracy, thanks to their ability to classify some difficult samples as indeterminate. However, according to the results in Table 2(b), CRF is able to achieve the highest single-set accuracy, which indicates that it is the most reliable model when determinate predictions are made. Tables 2(c) and 2(d) show that in terms of utility-discounted accuracy (both u_{65} and u_{80}), which measures a trade-off between cautiousness and single-set accuracy, CRF outperforms all other baselines in the great majority of cases. This is confirmed by using the Friedman test and Nemenyi test in Table 3(a) and 3(b). CRF outperforms significantly all other models (with a p-value less than 0.05) except AVE, for which the differences are not significant. This first round of experiments thus shows that our combination and decision strategy based on the theory of belief functions provides an interesting way of making cautious and reliable decisions.

We now move onto the second part of this first phase of experiments, designed to study the robustness of CRF against noisy data. The ability to adapt to noisy data is an important feature of a good classifier. In our case, the classifier is expected to become more cautious when faced with low-quality data. In these experiments, we investigate the impact of label noise on model performance, by introducing a given percentage of erroneous labels in the training samples. Figures 1(a) to 1(d) display the behavior of the four evaluation metrics for the compared models, averaged over all datasets, as a function of label noise.

As expected, the cautiousness of all models increases as the noise level increases. However, the effect is strongest for CRF and AVE: with 20% of noisy labels, cautiousness is increased by about 15%, which indicates that CRF and AVE perform better in presence of noise compared to MV, MVTH and MVA. For MV and MVTH, cautiousness is only increased by about 5%. Even worse, MVA seems to

be insensitive to noise and always maintains cautiousness around 0.5%. It should be noted that CRF is even more cautious than AVE for very high levels of label noise. From 0% to 20%, the single-set accuracy of the traditional random forest dropped by 5%, and by 3% for MV, MVTH and MVA, whereas the results of CRF and AVE suffered a decrease of about 1% only. Note also that CRF always keeps a slight advantage over AVE. The same can also be noticed with the u_{65} and u_{80} metrics.

These results show that CRF performs well in case of high aleatoric uncertainty in the data. Another crucial type of uncertainty is epistemic uncertainty, which is mainly caused by a lack of training data (Hüllermeier & Waegeman, 2021). In general, a cautious classifier faced to a high epistemic uncertainty should maintain a high degree of cautiousness to reduce the risk of making incorrect decisions. As the training set size increases (i.e., more data are collected), cautiousness should decrease and converge to a constant level caused by aleatoric uncertainty. Thus we have carried out some experiments so as to study this point. Figure 2(a) presents average cautiousness computed over all datasets for the four methods, when varying the size of the training set. It can be seen that all models tend to be more cautious as the size of the training set gets smaller, but CRF and AVE are far more sensitive to this parameter. This makes it possible for CRF and AVE to reach a higher single-set accuracy, so as to gain also a higher u_{65} and u_{80}, regardless of the size of the training data, as shown in Figures 2(b) to 2(d).

5.3. Phase 2: tree weight assignment strategy

Models compared

The second phase of experiments evaluates the interest of learning tree weights by optimizing the proposed cost function (20). For this purpose, the aggregation strategy used for all models is the one defined by Equations (12)-(14). Different weighting strategies are compared to each other:

- EW: the Equal Weight strategy assigns a weight $1/T$ to each tree;
- OOBACC: the Out-Of-Bag ACCuracy approach assigns a weight to each tree according to its accuracy, estimated using out-of-bag samples;
- OOBU65: this approach is similar to OOBACC, except that the performance of each tree is determined using the u_{65} criterion (see Section 5.1);
- IRF: tree weights are learnt using the cost function proposed by Utkin et al. (2019), which corresponds to Equation (17);
- AW: our proposed tree weight allocation strategy, where weights are obtained so as to minimize Equation (20).

Experimental setting and protocol

In order to evaluate the various tree-weighting strategies, we used the following protocol. For all weight assignment strategies, we used the same values of s as in the first phase of experiments. The parameter λ was set to 0.5 for all datasets in the experiments.

For CRF with AW, and for each dataset, we selected the value for the parameter γ in (20) that maximizes the u_{65} score and fix the parameter λ to 10 for all datasets. Regarding the IRF approach, Utkin et al. (2019) proposed to avoid overfitting by grouping the trees, and computing a weight for each group instead of each tree. We followed the procedure described and performed grid search cross-validation so as to select the best combination of the two hyperparameters $\epsilon \in \{0.25, 0.5, 0.75\}$ and $G \in \{5, 10, 20, 25, 100\}$; however, we maximized the u_{65} score instead of accuracy, since we compare here cautious classification strategies.

Cautiousness, single-set accuracy, u_{65} and u_{80} were evaluated by averaging the results obtained on 10 repetitions for each of the weight assignment methods compared, after the parameters were selected (in each repetition) using 10-fold cross-validation.

Results and discussion

In this section, the results obtained for various tree weight assignments in a cautious random forest are presented and analyzed. The influence of the parameter γ in the learning process is also discussed.

Tables 4(a) to 4(d) report the performances of CRF with different weight assignment methods. Thanks to the introduction of a specific utility value for indeterminate predictions, CRF with automatically-learnt weights (AW) always makes it possible to reach a good compromise between single-set accuracy and cautiousness: for all datasets, it yields the highest cautiousness degree, and at the same time the highest single-set accuracy, u_{65} and u_{80} values. The differences are significant (all p-values being less than 0.05), which is confirmed by the Friedman and Nemenyi tests reported in Tables 5(a) and 5(b).

It is noteworthy that the three weight assignment methods EW, OOBACC and OOBU65 achieve almost identical performances. This may be due to the fact that the differences between the trees are not significant enough to result in different decisions being made after normalization, especially since a voting mechanism is used. By contrast, the proposed weight assignment strategy better fits the decision trees in the forest, which results in higher accuracy scores. Remember that as illustrated by Utkin et al. (2020), the cost function in IRF is advantageous for precise classification problems: in an imprecise classification setting, considering only accuracy leads to designing classifiers that are not cautious enough, hence resulting in lower single-set accuracy, u_{65} and u_{80} values.

Our last experiment focuses on the influence of parameter γ, which was fixed using cross-validation in previous experiments. As explained above, this parameter has been introduced in the cost function to adjust the level of cautiousness in the model, so as to choose a specific behavior according to the user's needs: in general, the larger the value

Table 4
Comparison of wight assignment methods: metrics evaluated on each dataset, and counts for each strategy giving the highest and lowest scores.

(a) Cautiousness

Data	EW	OOBACC	OOBU65	IRF	AW
ADT	18.66	18.84	18.90	17.75	**20.27**
BKT	0.26	0.18	0.17	0.14	**0.27**
BID	10.32	10.13	10.20	9.97	**10.54**
BRC	1.51	1.65	1.63	1.69	**1.83**
CAD	8.05	8.03	8.02	7.94	**10.00**
COP	37.40	37.07	37.18	35.61	**38.77**
CRD	**13.32**	12.91	13.01	12.78	13.17
DIB	20.88	20.37	20.50	19.20	**20.90**
GER	33.13	33.20	33.28	32.95	**33.49**
HRT	**20.07**	20.00	**20.07**	18.65	19.84
HLC	21.90	21.72	21.79	21.33	**22.06**
INS	3.42	3.65	3.68	3.33	**3.73**
LIV	17.56	16.93	17.10	16.55	**18.76**
MGC	2.95	3.01	3.01	2.88	**3.17**
MMG	25.03	25.21	25.21	22.93	**25.25**
OCP	0.94	0.95	0.95	0.95	**1.26**
PHS	**5.63**	5.37	5.38	5.37	5.63
PMA	21.08	21.08	21.12	20.30	**21.84**
POP	29.04	28.25	28.37	23.36	**30.03**
RNO	5.41	5.51	5.51	5.43	**5.69**
SSC	1.29	1.17	1.15	1.26	**1.37**
SNR	11.64	11.70	11.65	10.67	**13.11**
SPM	2.61	2.53	2.51	2.41	**2.78**
VTE	3.70	3.54	3.54	3.13	**3.95**
WNE	**10.01**	9.49	9.55	9.58	9.85
Average	13.03	12.90	12.94	12.25	**13.50**
#Highest	4	0	1	0	**22**
#Lowest	3	2	1	**20**	0

(b) Single-set accuracy

Data	EW	OOBACC	OOBU65	IRF	AW
ADT	89.73	90.28	90.27	89.60	**91.69**
BKT	**99.37**	99.35	99.34	99.21	99.34
BID	90.94	90.88	90.89	90.50	**92.08**
BRC	96.93	97.20	97.16	96.91	**98.34**
CAD	79.94	80.03	80.05	79.68	**81.95**
COP	64.64	65.01	64.93	64.46	**66.67**
CRD	92.16	92.19	92.18	91.89	**93.45**
DIB	82.26	81.74	81.89	81.17	**83.08**
GER	84.78	84.81	84.83	84.37	**86.04**
HRT	88.47	88.64	88.65	87.80	**89.64**
HLC	75.15	75.14	75.21	74.77	**76.27**
INS	94.81	94.92	94.93	94.71	**96.00**
LIV	77.05	76.90	76.89	76.38	**78.71**
MGC	95.89	94.74	94.74	94.39	**95.90**
MMG	87.49	87.71	87.72	87.30	**88.84**
OCP	98.97	99.18	99.16	98.95	**99.31**
PHS	96.11	96.46	96.44	96.17	**97.67**
PMA	81.79	81.97	81.97	81.58	**83.35**
POP	67.91	65.84	66.23	65.39	**68.30**
RNO	95.03	95.41	95.46	95.14	**96.71**
SSC	93.73	94.05	94.07	93.77	**95.23**
SNR	87.58	86.95	86.82	86.26	**88.61**
SPM	95.37	95.53	95.56	95.21	**96.75**
VTE	97.68	97.76	97.75	97.37	**99.00**
WNE	85.17	85.02	85.05	84.75	**86.26**
Average	87.96	87.91	87.93	87.51	**89.17**
#Highest	1	0	0	0	**24**
#Lowest	3	0	0	**22**	0

(c) u_{65} score

Data	EW	OOBACC	OOBU65	IRF	AW
ADT	85.09	85.50	85.48	85.22	**86.26**
BKT	**99.29**	**99.29**	99.28	99.16	99.25
BID	88.26	88.25	88.25	87.94	**89.22**
BRC	96.44	96.67	96.63	96.37	**97.73**
CAD	78.72	78.81	78.84	78.51	**80.26**
COP	64.74	64.96	64.91	64.62	**65.99**
CRD	88.57	88.68	88.65	88.45	**89.71**
DIB	78.62	78.32	78.40	78.02	**79.28**
GER	78.14	78.20	78.20	77.96	**78.96**
HRT	83.75	83.92	83.89	83.54	**84.77**
HLC	72.92	72.92	72.97	72.67	**73.76**
INS	93.77	93.82	93.82	93.71	**94.82**
LIV	74.87	74.79	74.75	74.42	**76.08**
MGC	**94.98**	93.85	93.85	93.54	94.92
MMG	81.86	81.95	81.96	82.16	**82.79**
OCP	98.65	98.92	98.89	98.63	**99.07**
PHS	94.34	94.76	94.74	94.49	**95.82**
PMA	78.22	78.36	78.36	78.20	**79.31**
POP	66.17	65.62	65.81	65.44	**67.26**
RNO	93.40	93.74	93.78	93.51	**94.91**
SSC	93.36	93.70	93.73	93.40	**94.81**
SNR	84.90	84.35	84.27	84.07	**85.51**
SPM	94.57	94.76	94.79	94.47	**95.87**
VTE	96.48	96.60	96.59	96.36	**97.66**
WNE	83.15	83.11	83.13	82.86	**84.16**
Average	85.73	85.75	85.76	85.51	**86.73**
#Highest	2	1	0	0	**23**
#Lowest	5	0	0	**20**	0

(d) u_{80} score

Data	EW	OOBACC	OOBU65	IRF	AW
ADT	87.89	88.32	88.32	87.88	**89.30**
BKT	**99.32**	99.31	99.31	99.18	99.29
BID	89.81	89.77	89.78	89.44	**90.80**
BRC	96.67	96.92	96.88	96.62	**98.00**
CAD	79.93	80.02	80.04	79.70	**81.76**
COP	70.35	70.52	70.49	69.96	**71.81**
CRD	90.57	90.62	90.60	90.37	**91.68**
DIB	81.76	81.37	81.48	80.90	**82.42**
GER	83.11	83.18	83.19	82.90	**83.98**
HRT	86.76	86.92	86.90	86.33	**87.75**
HLC	76.21	76.18	76.23	75.87	**77.07**
INS	94.28	94.36	94.37	94.21	**95.39**
LIV	77.51	77.33	77.32	76.90	**78.89**
MGC	**95.42**	94.30	94.30	93.98	95.40
MMG	85.61	85.73	85.74	85.60	**86.58**
OCP	98.79	99.04	99.02	98.77	**99.07**
PHS	95.19	95.56	95.54	95.29	**96.66**
PMA	81.38	81.52	81.53	81.24	**82.58**
POP	70.52	69.86	70.06	68.94	**71.76**
RNO	94.21	94.56	94.60	94.32	**95.76**
SSC	93.55	93.88	93.90	93.59	**95.02**
SNR	86.64	86.11	86.02	85.67	**87.48**
SPM	94.96	95.13	95.17	94.84	**96.29**
VTE	97.03	97.13	97.12	96.83	**98.25**
WNE	84.65	84.53	84.56	84.29	**85.64**
Average	87.68	87.69	87.70	87.34	**88.75**
#Highest	2	0	0	0	**23**
#Lowest	3	0	0	**22**	0

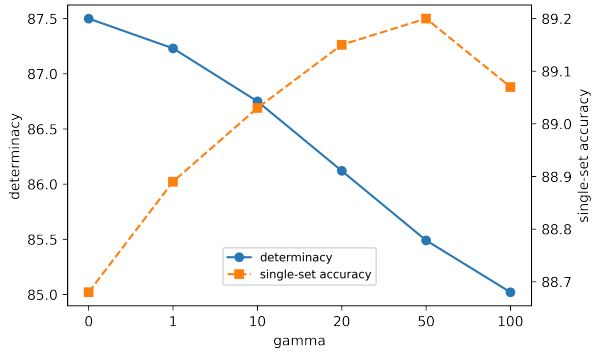
Table 5
Phase 2: Friedman statistic and p-value (left), Nemenyi p-values for pairwise model comparison.

(a) Friedman rank and test

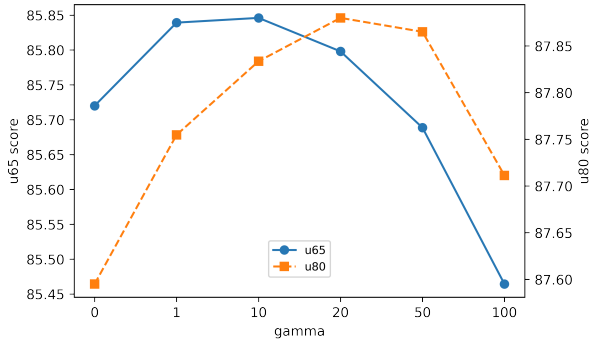
	EW	OOBACC	OOBU65	IRF	AW	p-value
cau	2.64	3.64	3.08	4.52	**1.12**	5.85E-08
ssa	3.08	3.36	2.88	4.64	**1.04**	3.89E-09
u65	2.96	3.60	2.96	4.40	**1.08**	1.32E-08
u80	3.04	3.48	2.76	4.64	**1.08**	2.20E-08

(b) Nemenyi test

AW	vs. EW	vs. OOBACC	vs. OOBU65	vs. IRF
cau	0.004	0.001	0.001	0.001
ssa	0.001	0.001	0.002	0.001
u65	0.001	0.003	0.001	0.001
u80	0.001	0.001	0.005	0.001



(a) Determinacy and single-set accuracy



(b) u_{65} and u_{80}

Figure 3: Average metrics computed over the 25 datasets, as a function of γ.

of γ, the more cautious the model, i.e. the lower the determinacy. Ideally, for each dataset, picking an appropriate value of γ would make it possible to reach the best compromise between determinacy and single-set accuracy.

Figures 3(a) and 3(b) illustrate the influence of γ on average determinacy, single-set accuracy, u_{65} and u_{80}, computed over all datasets. These metrics behave as expected: determinacy appears to be a decreasing function of γ, whereas single-set accuracy is increasing. When the value of γ is too large (for example for $\gamma = 100$), single-set accuracy slightly decreases; an explanation to this behavior would be that the cost function then highly favors indeterminate predictions: turning determinate, correct predictions into indeterminate ones then leads to a decrease in accuracy. The u_{65} and u_{80} both present an optimum, obviously attained for different values of γ, which could be determined for instance by cross-validation.

6. Conclusion

In this paper, we have proposed a new aggregation method to construct a cautious random forest in an imprecise classification setting. The method is based on a pre-trained random forest. Each tree in the forest provides intervals of probabilities obtained via the imprecise Dirichlet model, rather than point estimates.

Our strategy consists in aggregating the tree outputs using an extension of the weighted voting mechanism: all probability intervals generated for a test instance are used to compute the belief and plausibility of the event that the probability of class 1 belongs to the interval [0.5,1]. Finally, the interval dominance principle is applied, which may lead to making indeterminate decisions. We have also proposed a strategy for assigning weights to trees, based on a cost function which takes both determinacy and single-set accuracy into account. Optimizing this cost function thus allows to reach a compromise between cautiousness and accuracy.

Our experiments on 25 classical datasets showed that our aggregation method compares favorably to other aggregation operators leading to cautious decisions, such as averaging, majority voting (with indeterminate predictions), and majority voting with threshold. Experiments also show that our approach is robust to label noise and to scarcity of training data. Overall, in the very large majority of cases, our strategy performs better than averaging in terms of all evaluation metrics considered. Second, it is reasonably more cautious than majority voting (with or without threshold), but also more accurate on determinate predictions, which results in a lower risk for the model. Therefore, it seems a good candidate in classification problems where cautiousness is paramount, or when data are scarce or of a low quality.

Through a second series of experiments, we showed that our strategy for learning tree weights results in a more cautious model compared to the other four baselines, and achieves the best performances in terms of single-set accuracy, as well as u_{65} and u_{80} measures. In a nutshell, our strategy makes it possible to reach a good compromise between informativeness and cautiousness, by avoiding mistakes when the tree outputs appear to be too conflicting or too indeterminate.

Our model also has some limitations. First, as for previous strategies involving the IDM, we have no principled way for automatically choosing the value of the parameter s: as a consequence, without any preliminary tests, the resulting classifier may be over-cautious — note however that our strategy was shown to better resist to high s values than the other aggregation schemes (Zhang et al., 2021). Second, the cost function optimized (20) is only a surrogate for the criterion akin to the u_α score (18). The experimental results nevertheless suggest that our optimization procedure effectively leads to reach a good compromise between accuracy and cautiousness.

In future works, we plan to extend the aggregation strategy to multi-class problems. We will also work on providing explanations for the model, based on our aggregation strategy and the weights obtained in the optimization procedure, for questions such as why a particular sample is predicted imprecisely or what would be the minimal changes to bring to an instance to enable it to be precisely classified.

Appendix

A. Gradient of the cost function

In this section, we provide the expression for the gradient of the proposed cost function. Considering the sigmoid function $U(x)$ defined by Equation (19), it should be noted that

$$U'(x) = \alpha U(x)(1 - U(x)).$$

If we write

$$\underline{u}_i = U(\boldsymbol{w}^\top \underline{\boldsymbol{\delta}}_i - 0.5), \overline{u}_i = U(\boldsymbol{w}^\top \overline{\boldsymbol{\delta}}_i - 0.5),$$
$$u_i = U(\boldsymbol{w}^\top \underline{\boldsymbol{\delta}}_i - 0.5)(\boldsymbol{w}^\top \overline{\boldsymbol{\delta}}_i - 0.5),$$

the cost function (20) can be rewritten as:

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\Big\{y_i \ln(\underline{u}_i) + (1-y_i)\ln(1-\overline{u}_i) + \gamma \ln(1-u_i)\Big\} + \frac{1}{2}\lambda||\boldsymbol{w}||_2^2. \quad (22)$$

Obviously, we have

$$\nabla_{\boldsymbol{w}} y_i \ln(\underline{u}_i) = \alpha y_i (1-\underline{u}_i)\underline{\boldsymbol{\delta}}_i,$$
$$\nabla_{\boldsymbol{w}}(1-y_i)\ln(1-\overline{u}_i) = -\alpha(1-y_i)\overline{u}_i\overline{\boldsymbol{\delta}}_i,$$

and

$$\nabla_{\boldsymbol{w}} \ln(1-u_i) = -\alpha u_i\left[(\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top)\boldsymbol{w} - 0.5(\underline{\boldsymbol{\delta}}_i + \overline{\boldsymbol{\delta}}_i)\right].$$

If we write $\boldsymbol{\delta}_i = (\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top)\boldsymbol{w} - 0.5(\underline{\boldsymbol{\delta}}_i + \overline{\boldsymbol{\delta}}_i)$, the gradient for the cost function writes as

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = -\frac{\alpha}{N}\sum_{i=1}^{N}\Big\{y_i(1-\underline{u}_i)\underline{\boldsymbol{\delta}}_i - (1-y_i)\overline{u}_i\overline{\boldsymbol{\delta}}_i - \gamma u_i\boldsymbol{\delta}_i\Big\} + \lambda\boldsymbol{w}. \quad (23)$$

B. Hessian and convexity

In this section, we provide the Hessian matrix of the cost function 20 and the proof that it is positive semi-definite, so as to prove the convexity of the cost function. First, the Hessian matrix can be calculated separately for each part of the cost function:

$$H(y_i \ln(\underline{u}_i)) = -\alpha^2 y_i \underline{u}_i(1-\underline{u}_i)\underline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top,$$
$$H((1-y_i)\ln(1-\overline{u}_i)) = -\alpha^2(1-y_i)\overline{u}_i(1-\overline{u}_i)\overline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top,$$
$$H(\frac{1}{2}\lambda||\boldsymbol{w}||_2^2) = \lambda I,$$

and

$$H(\gamma \ln(1-u_i)) = -\alpha^2\gamma u_i(1-u_i)\boldsymbol{\delta}_i\,\boldsymbol{\delta}_i^\top - \alpha\gamma u_i(\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top).$$

Consequently, the complete Hessian matrix writes as:

$$H(J(\boldsymbol{w})) = \frac{\alpha^2}{N}\sum_{i=1}^{N}\Big\{y_i\underline{u}_i(1-\underline{u}_i)\underline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top + (1-y_i)\overline{u}_i(1-\overline{u}_i)\overline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \gamma u_i(1-u_i)\boldsymbol{\delta}_i\,\boldsymbol{\delta}_i^\top + \frac{1}{\alpha}\gamma u_i(\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top)\Big\} + \lambda I.$$

All the matrices of the form $\xi\boldsymbol{a}\,\boldsymbol{a}^\top$, where ξ is a non-negative real number and $\boldsymbol{a}$ is a vector, are symmetric positive semi-definite. Moreover, λI is obviously symmetric positive definite. According to the theorem stating that the sum of two symmetric positive semi-definite matrices is also symmetric positive semi-definite, a sufficient and necessary condition for $H(J(\boldsymbol{w}))$ to be a symmetric positive semi-definite matrix is that the last term in the sum be symmetric positive semi-definite as well.

Since $(\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top)^\top = \underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top + \overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top$, it is symmetric. Suppose we have two non-zero vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, and let $A = \boldsymbol{a}\,\boldsymbol{b}^\top$. All rows of A are linearly dependent. Therefore, $\det A = 0$, and $\operatorname{rank} A = 1$. Since the rank of a matrix is equal to the number of non-zero eigenvalues and its trace is equal to the sum of its eigenvalues, matrix A has only one non-zero eigenvalue and its value is equal to its trace, which is $\operatorname{tr} A = \boldsymbol{a}^\top\boldsymbol{b}$.

It should be noted that all elements of $\underline{\boldsymbol{\delta}}_i$ and $\overline{\boldsymbol{\delta}}_i$ are either 0 or 1. Therefore, whenever both of these vectors are non-zero, $\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top$ and $\overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top$ each have only one positive eigenvalue and the other eigenvalues are zeros. If at least one of them is zero, all eigenvalues will be zero.

In conclusion, $\underline{\boldsymbol{\delta}}_i\,\overline{\boldsymbol{\delta}}_i^\top$ and $\overline{\boldsymbol{\delta}}_i\,\underline{\boldsymbol{\delta}}_i^\top$ have non- negative eigenvalues, and their sum is semi-positive definite, which completes the proof that the Hessian matrix is positive semi-definite and that the cost function is therefore convex.

CRediT authorship contribution statement

Haifei Zhang: Conceptualization, Methodology, Writing - original draft. **Benjamin Quost:** Conceptualization, Methodology, Writing - review & editing. **Marie-Hélène Masson:** Conceptualization, Methodology, Writing - review & editing.

Declaration of competing interest

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

References

Abellán, J. (2013). Ensembles of decision trees based on imprecise probabilities and uncertainty measures. *Information Fusion*, *14*, 423–430. https://doi.org/10.1016/j.inffus.2012.03.003.

Abellán, J., & Masegosa, A. R. (2010a). Bagging decision trees on data sets with classification noise. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 248–265). Springer, Berlin, Heidelberg volume 5956 LNCS. https://doi.org/10.1007/978-3-642-11829-6_17.

Abellán, J., & Masegosa, A. R. (2010b). An ensemble method using credal decision trees. *European Journal of Operational Research*, *205*, 218–226. https://doi.org/10.1016/j.ejor.2009.12.003.

Abellán, J., & Masegosa, A. R. (2012). Imprecise classification with credal decision trees. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *20*, 763–787. https://doi.org/10.1142/s0218488512500353.

Abellán, J., Moral, S., Gómez, M., & Masegosa, A. (2006). Varying parameter in classification based on imprecise probabilities. *Advances in Soft Computing*, *37*, 231–239. https://doi.org/10.1007/3-540-34777-1_28.

Ambika, & Biradar, S. (2021). Survey on prediction of loan approval using machine learning techniques. *International Journal of Advanced Research in Science, Communication and Technology*, (pp. 449–454). https://doi.org/10.48175/ijarsct-1165.

Bache, K., & Lichman, M. (2013). Uci machine learning repository. https://archive.ics.uci.edu/ml/index.php.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*, 627–635. https://doi.org/10.1057/palgrave.jors.2601545.

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1007/bf00058655.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *Wadsworth Inc*, *67*. https://doi.org/10.2307/2530946.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning* (p. 18). https://doi.org/10.1145/1015330.1015432.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SigKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

De Campos, L. M., Huete, J. F., & Moral, S. (1994). Probability interval; A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *2*, 167–196. https://doi.org/10.1142/S0218488594000146.

Dempster, A. P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, *38*, 325–339. https://doi.org/10.1214/aoms/1177698950.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30. https://www.jmlr.org/papers/v7/demsar06a.html.

Denœux, T. (2009). Extending stochastic ordering to belief functions on the real line. *Information Sciences*, *179*, 1362–1376. https://doi.org/10.1016/j.ins.2009.01.009.

Dmochowski, J. P., Sajda, P., & Parra, L. C. (2010). Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, *11*. https://www.jmlr.org/papers/v11/dmochowski10a.html.

Fink, P. (2012). *Ensemble methods for classification trees under imprecise probabilities*. Master's thesis Ludwig-Maximilians University.

Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical engineering online*, *13*, 1–9. https://doi.org/10.1186/1475-925x-13-94.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*, 1612. http://www.yorku.ca/gisweb/eats4400/boost.pdf.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*, 119–139. https://doi.org/10.1006/jcss.1997.1504.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, *11*, 86–92. https://doi.org/10.1214/aoms/1177731944.

Grandvalet, Y. (2004). Bagging equalizes influence. *Machine Learning*, *55*, 251–270. https://doi.org/10.1023/b:mach.0000027783.34431.42.

Haddouchi, M., & Berrado, A. (2019). A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)* (pp. 1–6). https://doi.org/10.1109/icssd47982.2019.9002770.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, *110*, 457–506.

Isinkaye, F. O., Folajimi, Y., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, *16*, 261–273. https://doi.org/10.1016/j.eij.2015.06.005.

Kim, H., Kim, H., Moon, H., & Ahn, H. (2011). A weight-adjusted voting algorithm for ensembles of classifiers. *Journal of the Korean Statistical Society*, *40*, 437–449. https://doi.org/10.1016/j.jkss.2011.03.002.

Li, H. B., Wang, W., Ding, H. W., & Dong, J. (2010). Trees weighting random forest method for classifying high-dimensional noisy data. In *2010 IEEE 7th international conference on e-business engineering* (pp. 160–163). IEEE. https://doi.org/10.1109/icebe.2010.99.

Mangili, F., & Benavoli, A. (2015). New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, *56*, 278–306. https://doi.org/10.1016/j.ijar.2014.08.005.

Mantas, C. J., & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, *41*, 4625–4637. https://doi.org/10.1016/j.eswa.2014.01.017.

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2016). *Autonomous driving: technical, legal and social aspects*. Springer Nature. https://doi.org/10.1007/978-3-662-48847-8.

Moral-García, S., Mantas, C. J., Castellano, J. G., Benítez, M. D., & Abellan, J. (2020). Bagging of credal decision trees for imprecise classification. *Expert Systems with Applications*, *141*, 112944. https://doi.org/10.1016/j.eswa.2019.112944.

Murphy, C. K. (2000). Combining belief functions when evidence conflicts. *Decision support systems*, *29*, 1–9. https://doi.org/10.1016/s0167-9236(99)00084-6.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine learning*, *42*, 203–231. https://doi.org/10.1023/A:1007601015854.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/bf00116251.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. https://doi.org/10.1007/BF00993309.

Sage, A. J., Genschel, U., & Nettleton, D. (2020). Tree aggregation for random forest class probability estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *13*, 134–150. https://doi.org/10.1002/sam.11446.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press. https://doi.org/10.2307/j.ctv10vm1qb.5.

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications* (pp. 253–260). Springer. https://doi.org/10.1007/978-981-13-2354-6_27.

Troffaes, M. C. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29. https://doi.org/10.1016/j.ijar.2006.06.001.

Utkin, L. V., Kovalev, M. S., & Coolen, F. P. (2020). Imprecise weighted extensions of random forests for classification and regression. *Applied Soft Computing*, *92*, 106324. https://doi.org/10.1016/j.asoc.2020.106324.

Utkin, L. V., Kovalev, M. S., & Meldo, A. A. (2019). A deep forest classifier with weights of class probability distribution subsets. *Knowledge-Based Systems*, *173*, 15–27.

Walley, P. (1996). Inferences from Multinomial Data: Learning About a Bag of Marbles. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 3–34. https://doi.org/10.1111/j.2517-6161.1996.tb02065.x.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1.

Zaffalon, M., Corani, G., & Mauá, D. (2012). Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, *53*, 1282–1301. https://doi.org/10.1016/j.ijar.2012.06.022.

Zhang, H., Quost, B., & Masson, M.-H. (2021). Cautious random forests: a new decision strategy and some experiments. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)* (pp. 1–4). https://proceedings.mlr.press/v147/zhang21a.html.