Cautious Random Forests: a New Decision Strategy and some Experiments

Haifei Zhang HAIFEI.ZHANG@HDS.UTC.FR
Benjamin Quost BENJAMIN.QUOST@HDS.UTC.FR
UMR UTC-CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, France

Marie-Helène Masson MYLENE.MASSON@HDS.UTC.FR
Université de Picardie Jules Verne, France

Abstract

Random forest is an accurate classification strategy, which estimates the posterior probabilities of the classes by averaging frequencies provided by trees. When data are scarce, this estimation becomes difficult. The Imprecise Dirichlet Model can be used to make the estimation robust, providing intervals of probabilities as outputs. Here, we propose a new aggregation strategy based on the theory of belief functions. We also propose to assign weights to the trees according to their amount of uncertainty when classifying a new instance. Our approach is compared experimentally to the baseline approach on several datasets.

Keywords: Imprecise random forests, imprecise Dirichlet model, belief functions

1. Introduction

Ensemble learning can significantly improve the performance of basic machine learning models. We may cite three main approaches: bagging [5], stacking [20] and boosting [12]. In a random forest [6], a variation of bagging, a large number of unpruned decision trees are trained by introducing sample and feature randomness; the tree outputs (decisions or class probability distributions) are then aggregated either by voting or by averaging. Different aggregation schemes have been compared in [15]. Random forests have been successfully applied in many settings. However, making precise predictions is questionable when the available information is scarce, or when there is a large conflict between the decision tree outputs. Then, an alternative consists in keeping the model cautious by producing sets of decisions, or probabilities, so as to achieve robustness.

In this paper, we deal with the case where random forests provide imprecise predictions using Walley's Imprecise Dirichlet Model (IDM) [19]. Whereas classical inference is based solely on posterior probability estimates, obtained by calculating the class frequencies over the instances falling into a leaf node, the IDM incorporates an imprecise information with regard to these class frequencies in the form of a set of Dirichlet distributions; consequently, by conjugacy, the posterior information on the classes is an updated set of Dirichlet distributions [4]. The information inferred from the data falling into a leaf node can then be described by lower and upper bounds on the posterior probabilities of the classes. The interest of using the IDM in decision trees so as to provide cautious decisions has been demonstrated in several studies [2, 14, 18].

Since the prediction generated by the cautious random forest is a set of probability intervals, aggregating the tree outputs is more difficult than with precise predictions. Several strategies may be used for this purpose. One consists in first obtaining the prediction for each tree (using for example interval dominance [17]), and then obtaining a final prediction by majority voting or weighted majority voting [1]. Another consists in directly merging all probability intervals in the ensemble, either using disjunction or conjunction [7] or by averaging [11], and then making a prediction based on the resulting probability intervals.

In this paper, we propose an approach to combine the tree outputs based on theory of the belief functions [8, 16]. We consider that the trees provide pieces of evidence about the value of the true probability in the form of closed random intervals defined on $[0;1]$. It is then easy to compute the belief and plausibility of any event defined on $[0;1]$, and to use them in a cautious decision-making process [10]. We also study the interest of weighting the trees in the ensemble, via two strategies: weights based on the number of samples in the leaves, or weights equal to the length of the intervals (epistemic uncertainty of the intervals). This approach can be seen as a generalization of voting.

2. Preliminaries

2.1. Imprecise Dirichlet Model (IDM)

Let us assume a sample space $\Omega = \{\omega_1, \ldots, \omega_K\}$ with $K \geq 2$ elements, and let $\pi_1, \ldots, \pi_K$ be an unknown multinomial distribution over Ω (with $\pi_j = \mathbb{P}(\omega_j)$ for $j = 1, \ldots, K$. Let N instances be sampled independently from this distribution: we obtain a vector $n = (n_1, \ldots, n_K)$, of numbers of instances in each class (with $\sum_{j=1}^{K} n_j = N$). An imprecise Dirichlet prior is updated by the likelihood of observations into the following lower and upper posterior probabilities:

$$\underline{\mathbb{E}}(\pi_j|n) = \frac{n_j}{N+s}, \quad \overline{\mathbb{E}}(\pi_j|n) = \frac{n_j+s}{N+s}. \tag{1}$$

The parameter s can be interpreted as a number of virtual samples with unknown class information. Although several studies have been conducted with regard to choosing an appropriate value [3], this problem remains open.

2.2. Belief Functions Induced by Random Intervals

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 $\mathbb{R}$:

$$bel(A) = \mathbb{P}([U,V] \subseteq A), \quad pl(A) = \mathbb{P}([U,V] \cap A \neq \emptyset), \quad (2)$$

for any element A of the Borel sigma-algebra $\mathscr{B}(\mathbb{R})$ of the real line [9]. Let $I_i = [u_i, v_i]$ with $i = 1,\ldots,n$, and let m be the mass function from the set $\mathscr{I}$ of closed real intervals of $[0,1]$ such that $m(I_i) = m_i$ with $i = 1,\ldots,n$ and $\sum_{m=1}^{n} m_i = 1$. Under this setting, the belief and plausibility functions are

$$bel(A) = \sum_{I_i \subseteq A} m_i, \quad pl(A) = \sum_{I_i \cap A \neq \emptyset} m_i, \quad \forall i = 1,\ldots,n. \quad (3)$$

The intervals I_i are called focal intervals of m [10]. This definition provides a basis for pooling pieces of information provided by the trees with respect to the class probabilities.

3. Combining Credal Decision Trees

We focus here on a binary classification problem. Assuming a training data set (x_i, y_i) with $i = 1,\ldots,N$, where $y_i = \in \{0,1\}$, the probability that sample x_i belongs to category 0 (respectively, 1) is written $p_{i,0}$ (resp., $p_{i,1}$).

We consider a random forest composed of T trees $\{C_1,\ldots,C_t,\ldots,C_T\}$. Each instance x_i in the feature space belongs to a certain region of the random forest, defined by the set of regions $R = \{L_i^1,\ldots,L_i^t,\ldots,L_i^T\}$ with L_i^t (the region associated with) the leaf in which the instance falls for tree C_t. The leaf information is summarized by (n_i^t, N_i^t), where n_i^t is the number of samples of category 1 and N_i^t is the total number of training samples (the information being available for all leaves L_i^t, $t = 1,\ldots,T$).

The IDM gives an interval-valued estimate $I_t^i = [\underline{p}_{i,1}^t, \overline{p}_{i,1}^t]$ of $p_{i,1}$:

$$I_i^t = \left[\frac{n_i^t}{N_i^t + s}, \frac{n_i^t + s}{N_i^t + s}\right] \quad t = 1,\ldots,T. \quad (4)$$

We propose to aggregate these intervals by computing the belief and the plausibility of the interval $[0.5; 1]$, i.e. that the available evidence points towards class 1 for instance x_i. According to Equation (2), we have

$$\begin{aligned} bel_{i,1} &= bel(p_{i,1} \in [0.5,1]) \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(\underline{p}_{i,1}^t \geq 0.5) = \sum_{t=1}^{T} m_i^t \mathbb{I}(\underline{p}_{i,1}^t \geq 0.5), \end{aligned} \quad (5)$$

$$\begin{aligned} pl_{i,1} &= pl(p_{i,1} \in]0.5,1]) \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(\overline{p}_{i,1}^t > 0.5) = \sum_{t=1}^{T} m_i^t \mathbb{I}(\overline{p}_{i,1}^t > 0.5), \end{aligned} \quad (6)$$

where m_i^t $(t = 1,\ldots,T)$ is the mass of interval I_i^t in the aggregation process. Note that, by duality, we have $bel_{i,0} = 1 - pl_{i,1}$ and $pl_{i,0} = 1 - bel_{i,1}$. A natural choice is $m_i^t = 1/T$; we propose here two alternatives, which depend on the number of instances in the leaves. The first one is

$$m_i^t = \frac{N_i^t}{\sum_{j=1}^{T} N_i^j}, \quad \forall t = 1,\ldots,T; \quad (7)$$

using Equation (7), leaves with fewer samples bring weaker evidence. The second one defines, for $t = 1,\ldots,T$,

$$m_i^t = \frac{1 - u_i^t}{\sum_{j=1}^{T}(1 - u_i^j)}, \quad \text{with} \quad u_i^t = \frac{s}{N_i^t + s}, \quad (8)$$

where u_i^t is the level of epistemic uncertainty for instance x_i and for the tth tree. Intuitively, with this proposal, leaves with a smaller epistemic uncertainty get larger weights.

A decision can then be made by applying interval dominance to $bel_{i,1}$ and $pl_{i,1}$: we would choose class 1 whenever $bel_{i,1} > 0.5$, class 0 whenever $pl_{i,1} < 0.5$, and leave the decision as indeterminate otherwise.

4. Experimental Results

We compared the performance of our strategy with a baseline approach, which consists in averaging the lower and upper probability bounds over all trees: $\underline{p}_{i,1} = ave(\underline{p}_{i,1}^t)$ and $\overline{p}_{i,1} = ave(\overline{p}_{i,1}^t)$. This baseline also produces indeterminate predictions: class 0 is chosen if $\overline{p}_{i,1} < 0.5$ and class 1 if $\underline{p}_{i,1} > 0.5$, the decision being indeterminate otherwise. This baseline approach can be seen as a generalization of averaging precise probabilities and has been shown to produce good results [11]. In our experiments, we compared both proposals on eight data sets from the UCI Machine Learning Repository [13]. Different values were considered for the IDM parameter: $s \in \{1,3,5\}$, so as to study the behaviour of the various aggregation schemes with respect to epistemic uncertainty.

We compared the models using the u_{65} criterion [21], which is commonly used for comparing imprecise classification results. In the case of binary classification, u_{65} rewards precise and correct decisions with $\frac{1}{N}$, and indeterminate ones with $\frac{0.65}{N}$, thus making being cautious attractive while still penalizing indeterminate predictions. The mean value of u_{65} was computed using 10-fold cross-validation, due to the small amount of data. We statistically assessed the difference in u_{65} by repeating this procedure 50 times. Throughout the experiments, the random forest always consisted of 100 decision trees which were always trained to the maximum possible depth.

Tables 1, 2 and 3 display u_{65} values as a percentage. In these tables, mass1 refers to $m_i^t = 1/T$ (all trees have an equal weight), mass2 to Equation (7) (weights are proportional to the number of samples in the leaf) and mass3 to Equation (8) (weights depend on epistemic uncertainty). We compared the best result obtained with either of the credal approaches to that of the baseline, the best result being indicated in bold. We used a Student t-test on the 50 results obtained in each case, the level of significance being indicated using stars.

Table 1: Comparison in terms of u_{65} with s=1

Dataset	mass1	mass2	mass3	baseline
Pima	77.37	76.71	77.30	**78.11****
Heart	82.98	83.17	83.13	**83.56**
Biodeg	87.26	85.49	87.11	**87.67***
B-cancer	**96.19**	95.15	96.07	96.14
Cardiac	77.96	77.98	77.96	**78.12**
Wine	82.42	79.46	82.34	**82.74**
Magic	94.52	93.53	94.41	**94.68**
Spam	95.36	94.33	95.22	**95.50***

Table 2: Comparison in terms of u_{65} with s=3

Dataset	mass1	mass2	mass3	baseline
Pima	**78.33**	77.32	78.19	78.22
Heart	83.06	83.70	**83.89*****	82.56
Biodeg	**87.99**	85.69	87.64	87.80
B-cancer	96.15	95.10	**96.03**	95.96
Cardiac	78.59	78.80	**78.90**	78.75
Wine	81.86	79.06	**82.07*****	81.45
Magic	**94.32**	93.57	94.28	94.23
Spam	**95.27**	94.32	95.14	95.26

Table 3: Comparison in terms of u_{65} with s=5

Dataset	mass1	mass2	mass3	baseline
Pima	77.82	77.62	**78.59*****	76.86
Heart	81.17	**83.82*****	83.57	80.47
Biodeg	87.58	85.83	**87.73*****	86.91
B-cancer	95.78	95.08	**95.81**	95.58
Cardiac	78.03	**78.80**	78.44	78.41
Wine	80.57	79.71	**81.53*****	79.62
Magic	93.76	93.52	**94.09*****	93.60
Spam	94.92	94.32	**95.04*****	94.81

The belief function-based strategy performs slightly worse than the baseline for small values of s. However, as s increases, it outperforms baseline. Although we cannot find a weighting strategy consistently giving the best performances, weights based on epistemic uncertainty seem to guarantee better results for high s values. Recall that in the IDM, the parameter s is interpreted as number of virtual samples with unobserved category. Therefore, larger values induce a larger uncertainty. The belief-theoretic approach combined with a weighting strategy seems to withstand greater epistemic uncertainty, keeping the performances at a good level while remaining determinate, probably due to using appropriate weights.

The three different weighting strategies can be compared based on these preliminary results. Giving equal weights to the trees (mass1) seems to yield roughly the same results as the baseline — including regarding robustness to epistemic uncertainty. Weighting the trees based on the number of samples in their leaves (mass2) does not seem to be consistently better than the baseline in terms of u_{65}, even for high values of s. Weighting according to the level of epistemic uncertainty (mass3) proves to be much more fruitful in this case, even if both of these strategies (mass2 and mass3) are based on the same information (number of samples in the leaves attained in each tree by the test instance).

5. Conclusion

In this short contribution, we describe a new approach to aggregate probability intervals on the posterior probabilities of the classes produced by an imprecise random forest. We focus on binary decision trees with outputs obtained using the Imprecise Dirichlet Model. The approach is formalized within the theory of belief functions. The belief that the instance belongs to the positive class is estimated as the proportion of intervals supporting exclusively the assumption $p_1 \geq 0.5$ (i.e., with lower bound greater than 0.5); the plausibility of the positive class, as the proportion of intervals which do not contradict this assumption (upper bound greater than 0.5). A decision is then made by applying interval dominance. We propose two variants by assigning specific weights to the trees, based on the amount of information associated with the leaf reached by the test sample. The variants of our strategy are compared with the baseline approach (which averages the probability intervals) on several datasets. In presence of low uncertainty, the baseline seems to perform better, whereas our approach based on weights derived from the amount of uncertainty seems to be much more robust when uncertainty increases.

Future work may be conducted in several directions. First, we will consider learning automatically the weights of the trees using the training data, similarly to the approach proposed in [18]. Second, our aim is to provide a way to interpret the results of the random forest, and in particular to explain why a decision is indeterminate. Finally, we shall study the extension of our approach to the multiclass case.

References

[1] Joaquín Abellán and Andrés R. Masegosa. 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)*, volume 5956 LNCS, pages 248–265. Springer, Berlin, Heidelberg, 2010.

[2] Joaquín Abellán and Andrés R. Masegosa. Imprecise classification with credal decision trees. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, 20(5):763–787, 2012.

[3] Joaquín Abellán, Serafín Moral, Manuel Gómez, and Andrés Masegosa. Varying parameter in classification based on imprecise probabilities. *Advances in Soft Computing*, 37:231–239, 2006.

[4] Jean Marc Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-3):123–150, 2005.

[5] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] Luis M De Campos, Juan F Huete, and Serafin Moral. Probability interval; A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.

[8] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[9] A. P. Dempster. Upper and Lower Probabilities Generated by a Random Closed Interval. *The Annals of Mathematical Statistics*, 39(3):957–966, 1968.

[10] Thierry Denœux. Extending stochastic ordering to belief functions on the real line. *Information Sciences*, 179(9):1362–1376, 2009.

[11] Paul Fink. *Ensemble methods for classification trees under imprecise probabilities*. PhD thesis, Ludwig-Maximilians University, 2012.

[12] Yoav Freund and Robert E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[13] Moshe Lichman. Uci machine learning repository, 2013.

[14] Carlos J. Mantas and Joaquín Abellán. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, 41: 4625–4637, 2014.

[15] Andrew J. Sage, Ulrike Genschel, and Dan Nettleton. Tree aggregation for random forest class probability estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(2):134–150, 2020.

[16] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[17] Matthias C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[18] Lev V. Utkin, Maxim S. Kovalev, and Frank P.A. Coolen. Imprecise weighted extensions of random forests for classification and regression. *Applied Soft Computing Journal*, 92:106324, 2020.

[19] Peter Walley. Inferences from Multinomial Data: Learning About a Bag of Marbles. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):3–34, 1996.

[20] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[21] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. In *International Journal of Approximate Reasoning*, volume 53, pages 1282–1301. Elsevier, 2012.