




Bayesian Smoothing of Decision Tree Soft Predictions and Evidential Evaluation

Nicolas Sutton-Charani^(✉) 

EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Alès, Alès, France
`nicolas.sutton-charani@mines-ales.fr`

Abstract. As for many classifiers, decision trees predictions are naturally probabilistic, with a frequentist probability distribution on labels associated to each leaf of the tree. Those probabilities have the major drawback of being potentially unreliable in the case where they have been estimated from a limited number of examples. Empirical Bayes methods enable the updating of observed probability distributions for which the parameters of the *prior* distribution are estimated from the data. This paper presents an approach of smoothing decision trees predictive binary probabilities with an empirical Bayes method. The update of probability distributions associated with tree leaves creates a correction concentrated on small-sized leaves, which improves the quality of probabilistic tree predictions. The amplitude of these corrections is used to generate predictive belief functions which are finally evaluated through the ensemble extension of three evaluation indexes of predictive probabilities.

Keywords: Smoothing · Correction · Predictive probabilities · Decision trees · Bayesian empirical methods · Predictive belief functions · Uncertain evaluation

1 Introduction

Even if the predictions provided by classifiers are generally considered in a *precise* or *crisp* form, they are often initially computed as *soft* predictions through probability distributions, most probable labels being used as *hard* predictions at the final predictive or decision-making step. Decision trees are basic classifiers and regressors that are at the basis of many famous supervised learning algorithms (random forest, XGBoost, etc.). Once a tree built, the proportions of labels contained in each leaf are used to compute these predictive probabilities. Small leaves, *i.e.* containing only a small number of examples, therefore produce unreliable predictive probabilities since they are computed from a limited amount of data. Those unreliable leaves are usually cut in a post-pruning step in order to avoid overfitting, but due to the complexity pruning methodologies often involve, users tend to choose pre-pruning strategies, *i.e.* set more conservative stopping criteria, instead.

In the classic machine learning literature some work focus on classifiers predictive probabilities *calibration* in order to make them smoother or to correct

some intrinsic biases typical of different predictive models [16]. These approaches often involve the systematic application of mathematical functions requiring a set of dedicated data at the stage of calibration and sometimes requiring heavy computations in terms of complexity [20] or only considering a subset of the learning data [27]. Statistical approaches such as Laplacian or additive smoothing provide tools that are known to correct estimators in order to increase the impact of classes for which only few or even no data are available. Those techniques have been largely used in Natural Language Processing [12] and Machine learning [21, 23]. From a Bayesian point of view, Laplacian smoothing equates to using a non-informative prior, such as the uniform distribution, for updating the expectation of a Dirichlet distribution. Other works based on evidential models enable local adjustments during the learning phase of decision trees. In these methodologies class labels estimates are carried out on small subsets of data independently to the dataset global distribution [7].

This work aims at providing basic adjustments of decision trees outputs based on the tree structure and the global distribution of the learning data without involving any additional complexity. The approach presented in this article allows the correction of classification trees predictive probabilities in the case of binary classes. To achieve this, an empirical Bayesian method taking into account the whole learning sample as prior knowledge is applied and results in the adjustments of the probabilities associated with leaves relatively to their size. Unlike Bayesian smoothing which takes benefits only from the size of sub-samples corresponding to leaves, empirical Bayesian smoothing uses the whole distribution of labels in regards to leaves which can be viewed as a rich piece of information and is therefore legitimate to be incorporated in any predictive evidential modelling. To this extent, the ranges of the resulting estimates corrections are finally used to generate predictive belief functions by discounting the leaves predictive probabilities which can be finally evaluated following extensions of existing evaluation metrics. It should be noticed that this work is out of the scope of learning decision trees from uncertain data as in [6, 25, 26]. In this paper all the learning data are precise, it is at the prediction step that the uncertainty is modelled by frequentist probabilities which are smoothed and transformed into belief functions from their correction ranges.

After recalling the necessary background in Sect. 2, the proposed approach is described in detail in Sect. 3. Taking into account the predictive probability adjustments amplitude enables, in Sect. 4, the formalisation of an evidential generative model and the extension of three evaluation metrics of predictive probabilities to the case of uncertain evidential predictions. In Sect. 6, a first set of experiments shows the contribution of the methodology on the one hand in a pragmatic point of view in terms of predictive performance and on the other hand by the flexibility of decision-making it offers.

2 Basics

The learning of a decision tree corresponds to a recursive partitioning of the attributes space aiming at separating labels as well as possible (classification) or

to decrease their variance (regression) [1]. The learning data are thus distributed in the different tree leaves, which are then associated with probability distributions over class variables according to the proportions of labels in the examples they contain. In this article we restrict ourselves to the case of binary classes noted $\{1, 0\}$ or $\{+, -\}$.

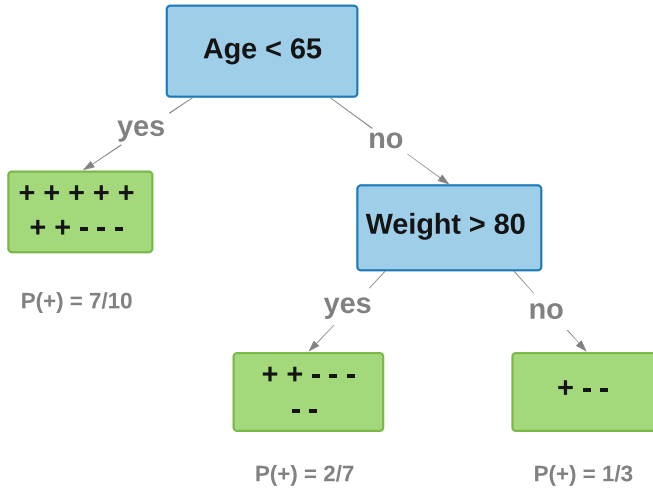


Fig. 1. Predictive probabilities

Fig. 1 is an example of a decision tree in which any example *younger* than 65 years old will have a positive class probability estimated by $P(+) = \frac{7}{10}$, for any *older* example (than 65 years old) and *heavier* than 80 kg we will have $P(+) = \frac{2}{7}$ and finally for examples older than 65 years old and weighing less than 80 kg we will predict $P(+) = \frac{1}{3}$. This last probability is here estimated from only 3 examples, it is therefore natural to consider it relatively unreliable (in comparison to the two others based on respectively 10 and 7 examples).

Various stopping criteria can be used during the learning of a decision tree depending on the structure of the tree (leaves number, depth, etc.) or in terms of information (impurity gain, variance). In order to avoid overfitting, pruning strategies are generally implemented to limit the number of leaves (which reduces variance and complexity).

2.1 Empirical Bayesian Methods

Bayesian inference is an important field of Statistics which consists in using some *prior* knowledge in order to update the estimations computed on data (according to Bayes theorem). This updates often result in predictive probabilities whose quality directly depends on the prior information. While Bayesian priors are generally constituted of probability distributions that the user subjectively express

about the phenomenon of interest (expert opinions are often at the basis of the prior’s modelling), for the Bayesian empirical methods [4, 22], the parameters of these prior distributions are estimated from the data. Some authors consider these methods as approximations of hierarchical Bayesian models [19].

Considering a sample of a binary variable $y = (y_1, \dots, y_n) \in \{0, 1\}^n$, for all subset $y^* \subseteq y$, a natural (and frequentist) estimate of the probability of label 1, i.e. of $p^* = P(Y = 1|Y \in y^*)$, is its observed frequency $\hat{p}^* = \frac{|\{y_i \in y^* : y_i = 1\}|}{|y^*|}$. This estimator makes implicitly the assumption that the sample y^* is large enough to estimate the probability of label 1 by a random draw in it.

In Bayesian statistics, for the binomial model corresponding to y^* ’s generation, a *prior* knowledge about p^* is usually modelled in terms of probability distribution by $p^* \sim \text{Beta}(\alpha, \beta)$ (*flexible* prior) and the model update from the data y^* results in a *posterior* distribution $p^*_|y^* \sim \text{Beta}(\alpha + n^*_1, \beta + n^* - n^*_1)$ with $n^*_1 = |\{y_i \in y^* : y_i = 1\}|$ and $n^* = |y^*|$.

The Bayesian (*posterior*) estimator of p^* is finally computed as its conditional expectation given y^* :

$$\hat{p}^* = \mathbb{E}[p^*|y^*] = \frac{n^*_1 + \alpha}{n^* + \alpha + \beta} \quad (1)$$

By doing so, \hat{p}^* is shifted toward its global expectation $E[\hat{p}^*] = \frac{\alpha}{\alpha + \beta}$ and this shift’s range relies on the value of n^* .

Whereas in standard Bayesian works the *prior*’s parameters are often set by an expert according to his knowledge, i.e. subjectively, or by default in a non-informative form as with Laplacian smoothing ($p^* \sim \text{Beta}(1, 1)$ is equivalent to $p^* \sim \mathcal{U}[0, 1]$), in empirical Bayesian approaches they are estimated from the whole sample y (by likelihood maximisation) with the hypothesis that $p \sim \text{Beta}(\alpha, \beta)$. One direct consequence is that the smaller the size of y^* , the greater the amplitude of the shift of p^* toward its global expectation on y . The Bayesian estimator of this approach, illustrated for the number of successful baseball shots per player in [2], is here applied to the probabilistic predictions attached with the leaves of binary classification trees.

2.2 Evaluation of Predictive Probabilities

Even if the evaluation of a classifier is often done from precise or *crisp* predictions by comparing them to the real class labels through different metrics (e.g. accuracy, precision, recall, etc.) it can however be done at the level of predictive probabilities, thus upstream. Three metrics for evaluating binary probabilistic predictions are presented hereinafter.

Let $y = (y_1, \dots, y_n) \in \{1, 0\}^n$ be the *true* labels of a given sample of size n and $p = (p_1, \dots, p_n) = [P(Y_1 = 1), \dots, P(Y_n = 1)]$ the class predictive probabilities of label $\{1\}$ according to a given predictive model M , applied to the sample x . The Table 1 summarizes the definitions of the *log-loss*, the Brier score and the area under the *ROC* curve (*AUC*). The log-loss can be interpreted as a Kullback–Leibler divergence between p and y that takes into account p ’s entropy and is therefore called *cross-entropy* by some authors. The Brier score is defined

as the mean squared difference between p and y . The log-loss and Brier score thus measure the difference between observations and predicted probabilities, penalizing the probabilities of the least probable labels. The ROC curve is a standard measure of a binary classifier's predictive power, it represents the rate of true positives or *sensibility* (*i.e.* the proportion of positive examples that are predicted as positive) as a function of the rate of false positives or $1 - \textit{specificity}$ (*i.e.* the proportion of negative examples that are predicted as positive), we have $ROC : \textit{sensibility}(1 - \textit{specificity})$. The AUC area under the ROC curve is a well known indicator of the quality of probabilistic predictions.

The three metrics thus defined lie in $[0, 1]$ and a good binary classifier will be characterized by log-loss and Brier score values close to 0 and an AUC value close to 1. It should be noted, however, that these three metrics are defined for standard uncertain predictions, *i.e.* probabilistic predictions.

Table 1. Evaluation metrics of binary probabilistic predictions (p_1, \dots, p_n) with respect to true labels (y_1, \dots, y_n)

Name	Definition
Log-loss	$-\frac{1}{n} \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$
Brier score	$\frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2$
AUC	$\int_0^1 ROC(p, y, \lambda) d\lambda$

3 Empirical Bayesian Correction of Decision Trees Predictive Probabilities

The approach presented in this paper consists in correcting the predictive probabilities of a binary classification tree containing H leaves denoted l_1, \dots, l_H with an empirical Bayesian method. It is assumed that the proportion of label $\{1\}$ within the leaves of the tree follows a $Beta(\alpha, \beta)$ distribution (without setting *a priori* the values of the parameters α and β). We can notice that the Beta distribution is both a special case of the Dirichlet distribution (widely used in Bayesian statistics) and a generalization of the uniform distribution ($\mathcal{U}_{[0,1]} = Beta(1, 1)$) which is supposed to model *non-informative* prior knowledge.

Once the hypothesis of the Beta law is formulated, its parameters α and β are estimated from the set $\{p_1^1, \dots, p_1^H\}$ of label proportions $\{1\}$ within the H leaves of the considered tree. In order to penalize small leaves, we will consider an artificial sample denoted E containing each p_1^h proportion repeated a number of times equal to the size of the considered leaf, *i.e.* to the number of examples it contains.

We have $E = (\underbrace{p_1^1, \dots, p_1^1}_{|l_1| \text{ times}}, \dots, \underbrace{p_1^H, \dots, p_1^H}_{|l_H| \text{ times}})$.

The estimation of the parameters α and β can then be performed on E by likelihood maximisation following different approaches (moments, least squares, etc.). This step makes α and β 's values take into account the leaves sizes through an artificial weighting that is defined according to their sizes. By doing so, the information of the tree structure is used through the leaves partitioning and their sizes to define the prior or the empirical Bayes model. The probabilities \hat{p}_1^h of the leaves l_1, \dots, l_H are finally corrected according to the Eq. (1):

$$\tilde{p}_1^h = \frac{n_1^h}{n^h} \rightsquigarrow \hat{p}_1^h = \frac{n_1^h + \alpha}{n^h + \alpha + \beta} \quad (2)$$

with n^h and n_1^h denoting respectively the size of the leaf l_h (i.e. the number of examples it contains) and the number of examples it contains that have the label $\{1\}$.

It should be recalled that other works [20, 27] allow a *calibration* of the predictive probabilities by different approaches using either only the distributions of the examples contained in the leaves independently from one another, or the distribution of the whole learning sample but applying systematic transformations based on estimations requiring many computations (often obtained by cross-validation). The method proposed in this paper uses both the whole distribution of the training data (once distributed in the different leaves of a tree) and remains very simple in terms of complexity, the α and β parameters of Eq. (2) being estimated only once for the whole sample and then used locally on leaves according to their sizes at the prediction step. Nevertheless the range of these probabilistic correction represent a piece of information by itself that should be incorporated into the leaves in order to express a confidence level on themselves. In the next section an evidential generative model based on these correction ranges is presented.

4 Generation of Predictive Belief Functions

The uncertainty expressed in the predictive probabilities of a classifier is mainly *aleatory*. It is based on the mathematical model underlying the classifier and on frequentist estimates. The knowledge of the empirical Bayesian adjustment and its range is a piece of information in itself that can allow epistemic uncertainty to be incorporated into the leaves predictive probabilities. It is indeed natural to consider unreliable the predictive probabilities that are estimated on a small number of examples (and therefore highly adjusted).

Unlike many works on belief functions generation where uncertain data are used in evidential likelihoods applied to parametric models [9, 14] or random generation is extended to mass function [3], the context of this article is restricted to the case of discounting leaves predictive probabilities according to the range of their empirical Bayesian adjustment. As a first approach, we propose to generate a belief function from a predictive probability using its empirical Bayesian correction range $|\tilde{p} - \hat{p}|$ as unreliability indicator (\tilde{p} and \hat{p} denoting respectively standard frequentist and empirical Bayesian adjusted estimates of p), assigning

its weight to ignorance and subtracting it uniformly to singletons probabilities. The resulting belief function can be viewed as a type of evidential discounted predictive probabilities. This modeling relies on the hypothesis that the more a predictive probability is corrected (*i.e.* the smaller the subsample considered), the less reliable it is.

$$\begin{cases} m(\{1, 0\}) = |\tilde{p} - \hat{p}| \\ m(\{1\}) = \tilde{p} - \frac{|\tilde{p} - \hat{p}|}{2} \\ m(\{0\}) = 1 - \tilde{p} - \frac{|\tilde{p} - \hat{p}|}{2} \end{cases} \quad (3)$$

In order to make as few assumptions as possible, once the mass of ignorance is defined, the mass of the two classes are symmetrically discounted. We can notice that we have $Bel(\{0\}) = 1 - Pl(\{1\})$ and $Bel(\{1\}) = 1 - Pl(\{0\})$.

Remark: This belief function modelling can also be written in terms of imprecise probability: $p \in [p^-, p^+]$ with

$$\begin{cases} p^- = \tilde{p} - \frac{|\tilde{p} - \hat{p}|}{2} \\ p^+ = \tilde{p} + \frac{|\tilde{p} - \hat{p}|}{2} \end{cases} \quad (4)$$

The generative model (3-4) can be interpreted as a reliability modelling of the corrected leaves predictive probabilities. The output nature of a classification tree can thus be considered through imprecise probabilities. In the next section, an imprecise evaluation model is presented that keeps an evidential uncertainty level until final outputs of decision trees evaluation.

5 Imprecise Evaluation

Evidential predictions evaluation remains a challenging task, some works have been presented based on evidential likelihood maximisation of evaluation model parameters (error rate or accuracy) [25] but in case of a predictive belief function the simplest evaluation solution is to convert it into standard probability with the pignistic transform for instance. The main drawback of such practice is that all the information contained in the uncertain modelling of the predictions is lost but it has the pragmatic advantage of providing *crisp* evaluation metrics that can be easily interpreted.

In order to keep the predictive uncertainty or imprecision provided by the model (4) until the stage of the classifiers evaluation, it is possible to consider the metrics defined in Table 1 in an ensemblist or *intervalist* perspective. Indeed, to a set of predictive probabilities naturally corresponds a set of values taken by these evaluation metrics. An imprecise probability $[p^-, p^+]$ computed according to the model (4) will thus be evaluated imprecisely by an interval defined as follows:

$$eval([p^-, p^+], y) = \left[\min_{p \in [p^-, p^+]} eval(p, y), \max_{p \in [p^-, p^+]} eval(p, y) \right]$$

where $eval$ is one of the metrics defined in Table 1.

This type of evaluation approach involves that, the smaller the leaves of the evaluated decision tree, the more the probabilities associated with these leaves will be adjusted and the more imprecise the evaluation of these trees will be (*i.e.* the wider the intervals obtained). This approach is therefore a means of propagating epistemic uncertainty about the structure of the tree into its evaluation metrics. It should be noted, however, that this solution requires an effective browse of the entire $[p^-, p^+]$ interval, which implies a significant computational cost.

6 Experiments

In this section, a set of experiments is implemented to illustrate the practical interest of the empirical Bayesian correction model presented in this paper. Using six benchmark datasets from the UCI¹ and Kaggle² sites, 10-fold cross-validations are carried out with, for each fold of each data set, the learning of decision trees corresponding to different complexities on the nine other folds and an evaluation on the fold in question using the three precise metrics defined in Table 1. The ‘ cp ’ parameter (of the ‘`rpart`’ function in R) allows to control the trees complexity, it represents the minimal relative information gain of each considered cut during the learning of the trees. Trees denoted ‘pruned’ are learned with a maximum complexity ($cp = 0$) and then pruned according to the classical *cost-complexity* criteria approach of the *CART* algorithm [1]. The evaluations consist of the precise metrics computations presented in Table 1 as well as their uncertain extensions defined in Sect. 5. These steps are repeated 150 times in order to make the results robust to fold random generation and only the mean evaluations of the trees predictive probabilities are represented here. The codes used for the implementation of all the experiments presented below are available at <https://github.com/lgi2p/empiricalBayesDecisionTrees>.

Table 2 represents the characteristics of the different datasets used in terms of number of examples (n), number of attributes or predictor variables (J) and number of class labels (K). The Tables 3, 4, 5, 6, 7 and 8 contain the mean evaluations computed for each dataset and for each tree type, over all 150 cross-validations performed, before and after empirical Bayesian smoothing. Figures 2 and 3 illustrate the distributions of these results with evaluation intervals for the log-loss on the ‘banana’ dataset and for the Brier score on the ‘bankLoan’ dataset with respect to the supports of the corresponding uncertain evaluations (only the extreme points of the uncertain metrics are represented).

The log-loss of corrected trees is almost always lower than that of original trees. This increase in performance is clear for large trees (learnt with a low

¹ <https://archive.ics.uci.edu/ml/datasets.html>.

² <https://www.kaggle.com/datasets>.

Table 2. Datasets dimensions

	n	J	K
banana	5300	2	2
bankLoan	5000	12	2
banknote	1372	4	2
mammo	830	5	2
pima	768	8	2
ticTacToe	958	9	2

Table 3. Log-loss before smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.410	0.351	0.355	0.419	0.425	0.359
bankLoan	1	0.665	0.568	0.571	0.571	0.566
banknote	0.330	0.329	0.322	0.331	0.347	0.329
mammo	0.444	0.437	0.414	0.409	0.450	0.416
pima	0.886	0.884	0.742	0.655	0.565	0.630
ticTacToe	0.221	0.221	0.213	0.210	0.549	0.214

Table 4. Log-loss after smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.287	0.300	0.334	0.418	0.424	0.358
bankLoan	0.640	0.582	0.568	0.571	0.571	0.566
banknote	0.173	0.174	0.180	0.225	0.346	0.328
mammo	0.419	0.419	0.410	0.406	0.450	0.416
pima	0.557	0.557	0.545	0.541	0.564	0.630
ticTacToe	0.188	0.188	0.188	0.191	0.550	0.214

Table 5. Brier score before smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.084	0.086	0.097	0.129	0.131	0.086
bankLoan	0.229	0.204	0.190	0.192	0.192	0.189
banknote	0.040	0.040	0.043	0.057	0.098	0.041
mammo	0.130	0.130	0.125	0.123	0.140	0.126
pima	0.188	0.188	0.184	0.183	0.189	0.186
ticTacToe	0.061	0.061	0.061	0.062	0.187	0.062

Table 6. Brier score after smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.083	0.086	0.097	0.129	0.131	0.086
bankLoan	0.219	0.198	0.190	0.192	0.192	0.189
banknote	0.040	0.040	0.042	0.057	0.098	0.041
mammo	0.129	0.129	0.125	0.123	0.140	0.126
pima	0.183	0.183	0.179	0.179	0.188	0.186
ticTacToe	0.060	0.060	0.060	0.062	0.187	0.061

Table 7. *AUC* before smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.946	0.936	0.920	0.872	0.867	0.937
bankLoan	0.778	0.772	0.747	0.722	0.720	0.755
banknote	0.843	0.838	0.821	0.799	0.777	0.827
mammo	0.852	0.848	0.835	0.819	0.789	0.839
pima	0.839	0.836	0.826	0.812	0.771	0.821
ticTacToe	0.861	0.858	0.850	0.838	0.758	0.846

Table 8. *AUC* after smoothing

dataset\cp	0	0.001	0.005	0.01	0.05	pruned
banana	0.948	0.937	0.920	0.872	0.867	0.937
bankLoan	0.781	0.773	0.747	0.722	0.720	0.755
banknote	0.845	0.840	0.822	0.799	0.777	0.827
mammo	0.854	0.850	0.836	0.819	0.789	0.838
pima	0.842	0.839	0.828	0.813	0.771	0.821
ticTacToe	0.864	0.860	0.851	0.839	0.758	0.845

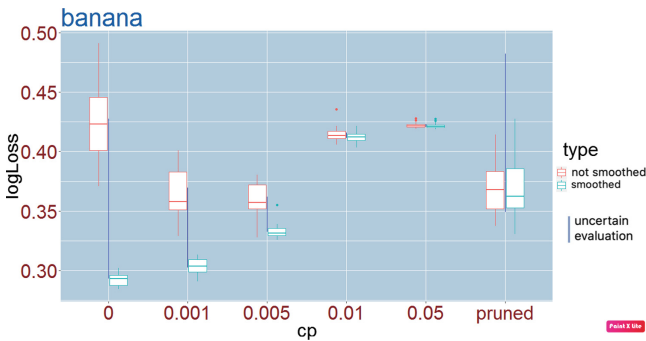


Fig. 2. Precise and uncertain log-loss as a function of the complexity

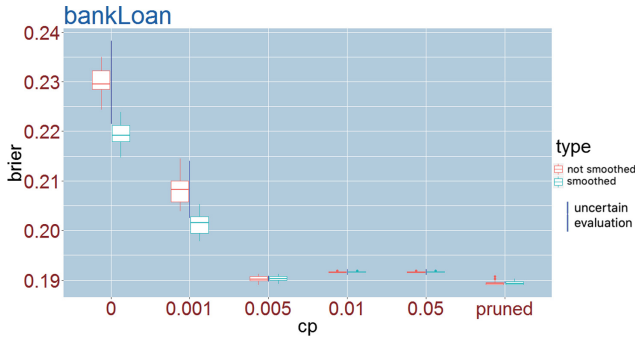


Fig. 3. Precise and uncertain Brier score as a function of the complexity

value of the hyper-parameter ‘ cp ’ and thus containing many leaves), slightly less visible for small trees and very limited for pruned trees. This difference in terms of performance gain in regards to the tree size is probably due to the fact that the bigger the trees, the smaller are their leaves. Indeed, if the examples of the same learning data are spread into a higher number of leaves, their number inside the leaves have to be lower (in average). This conclusion shows that empirical Bayesian adjustment makes sense especially for complex models. The same phenomenon of gain in performance proportional to tree size is globally observable for the Brier score and the AUC index but in smaller ranges.

Other works have already illustrated the quality increase of decision trees predictions by smoothing methods [5, 17, 18] but none of these neither illustrated nor explained the link between pruning and those increase ranges based on overfitting intuition. Moreover, using the learning sample labels distribution as prior knowledge is a new proposal that has not yet been studied in the context of decision trees leaves, especially in order to generate predictive belief functions (and their evaluation counterpart). Even if it was not in the scope of this article, some experiments have been carried on in order to compare the empirical Bayesian smoothing with the Laplacian one, and no significant differences were observed in terms of increase of the predictive evaluation metrics used in this paper.

The intervals formed by the uncertain evaluations correspond roughly to the intervals formed by the precise evaluations without and with Bayesian smoothing. However, it can be seen in Fig. 2 and 3 that uncertain evaluations sometimes *exit* from these natural bounds (pruned trees in Fig. 2 and *large* trees in Fig. 3), thus highlighting the non-convexity of the proposed uncertain metrics for evaluating evidential predictions.

7 Conclusion

The empirical Bayesian correction model presented in this paper for decision trees predictive probabilities is of interest in terms of predictive performance,

and this interest is particularly relevant for large trees. The fact that Bayesian corrections hardly improve the performance of *small* (i.e. with low complexity) or pruned trees suggests that the Bayesian correction represents an alternative to pruning. By shifting the predictive probabilities of small leaves to their global averages (i.e. calculated within the total learning sample) it reduces the phenomenon of overfitting.

In this paper Bayesian corrections are only performed at the predictive stage, i.e. at the leaf level. Adopting the same approach throughout the learning process is possible by proceeding in the same way at the purity gain computation level (i.e. for all the considered cuts). It would be interesting to compare the approach presented in this work with the one of [7] which pursues the same goal (penalizing small leaves) based on an evidential extension of purity gain computation where a mass of $\frac{1}{n+1}$ is assigned to *ignorance* in the impurity measure presented in [15] that combines *variability* and *non-specificity* computed on belief functions. It is important to note that the approach previously mentioned is based on the distribution of examples within the leaves individually, in case of unbalanced learning samples they do not allow correction in the direction of the general distribution as it is the case with the empirical Bayesian model.

The approaches of predictive belief functions generation based on the use of evidential likelihood [9, 13] also represent an interesting alternative to which it will be important to compare oneself both in terms of predictive performances and with respect to the underlying semantics. In the same vein, the approach presented in this paper for the binary classification context could be extended to the multi-class case following the approaches used in [24]. More generally, all classifiers whose learning at the predictive stage involves frequentist probability computations could potentially benefit from this type of Bayesian correction. Ensemble learning approaches could be enhanced by empirical Bayesian corrections at different levels. When their single classifiers are decision trees, it could be straightforward to correct trees with the correction model presented in this article. A global adjustment could also be achieved at the aggregation phase with the same type of artificial sample creation as for tree correction (leaves level could be extended to classifiers level).

The model for generating predictive belief functions and especially the extension of evaluation metrics to the credibility context proposed in this work could be greatly enriched by a more refined modeling of the uncertainty resulting from the initial predictive probabilities and their corrections. For example, it would be possible to use the distances proposed in [11] in order to make the representation of uncertain assessment metrics more complex beyond simple intervals. It would also be desirable to directly estimate the bounds of the uncertain evaluation intervals without having to browse them effectively from simulation-based approaches such as in [8] or by optimization results such as in [10].

References

1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Chapman & Hall, Boca Raton (1984)

2. Brown, L.P.: Empirical Bayes in-season prediction of baseball batting averages. *Ann. Appl. Stat.* **2**(1), 113–152 (2008)
3. Burger, T., Destercke, S.: How to randomly generate mass functions. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* **21**, 645–673 (2013)
4. Casella, G.: An introduction to empirical Bayes data analysis. *Am. Stat.* **39**(5), 83–87 (1985)
5. Chawla, N.V.: Many are better than one: improving probabilistic estimates from decision trees. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d’Alché-Buc, F. (eds.) *MLCW 2005. LNCS (LNAI)*, vol. 3944, pp. 41–55. Springer, Heidelberg (2006). https://doi.org/10.1007/11736790_4
6. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. *Int. J. Approximate Reasoning* **28**(2–3), 91–124 (2001)
7. Denœux, T., Bjanger, M.: Induction of decision trees from partially classified data using belief functions. In: *International Conference on Systems, Man And Cybernetics (SMC 2000)*, vol. 4, pp. 2923–2928 (2000)
8. Denœux, T., Masson, M.-H., Hébert, P.-A.: Non-parametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets Syst.* **153**(1), 1–28 (2005)
9. Denœux, T.: Likelihood-based belief function: justification and some extensions to low-quality data. *Int. J. Approximate Reasoning* **55**(7), 1535–1547 (2014)
10. Destercke, S., Strauss, O.: Kolmogorov-Smirnov test for interval data. In: *Information Processing and Management of Uncertainty (IPMU)*, pp. 416–425 (2014)
11. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: comprehensive survey and generalizations. *Int. J. Approximate Reasoning* **53**(2), 118–145 (2012)
12. Jurafsky, D., Martin, J.H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River (2000)
13. Kanjanatarakul, O., Sriboonchitta, S., Denœux, T.: Forecasting using belief functions: an application to marketing econometrics. *Int. J. Approximate Reasoning* **55**(5), 1113–1128 (2014)
14. Kanjanatarakul, O., Denœux, T., Sriboonchitta, S.: Prediction of future observations using belief functions: a likelihood-based approach. *Int. J. Approximate Reasoning* **72**, 71–94 (2016)
15. Klir, G.J.: *Uncertainty and Information: Foundations of gEneralized Information Theory*. Wiley- IEEE Press, New York (2013)
16. Kuhn, Max, Johnson, Kjell: *Applied Predictive Modeling*. Springer, New York (2013). <https://doi.org/10.1007/978-1-4614-6849-3>
17. Ferri, C., Flach, P.A., Hernández-Orallo, J.: Decision trees for ranking: effect of new smoothing methods, new splitting criteria and simple pruning methods. *Mathematics* (2003)
18. Margineantu, D.D., Dietterich, T.G.: Improved class probability estimates from decision tree models. In: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (eds.) *Nonlinear Estimation and Classification*, pp. 173–188. Springer, New York (1989). https://doi.org/10.1007/978-0-387-21579-2_10
19. Maritz, J.S., Lwin, T.: *Applied Predictive Modeling*, 2nd edn. Chapman and Hall, London (1989)
20. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, pp. 625–632 (2005)
21. Osher, S.J., Wang, B., Yin, P., Luo, X., Pham, M., Lin, A.T.: Laplacian smoothing gradient descent. In: *International Conference on Learning Representations (ICLR) 2019 Conference - Blind Submission* (2019)

22. Robbins, H.: An Empirical Bayes Approach to Statistics. University of California Press **1**, 157–163 (1956)
23. Sucar, L.E.: Probabilistic Graphical Models. ACVPR. Springer, London (2015). <https://doi.org/10.1007/978-1-4471-6699-3>
24. Sutton-Charani, N., Destercke, S., Denœux, T.: Classification trees based on belief functions. In: Proceedings of the 2nd International Conference on Belief Functions (BELIEF 2012) (2012)
25. Sutton-Charani, N., Destercke, S., Denœux, T.: Training and evaluating classifiers from evidential data: application to E2M tree pruning. In: Proceedings of the 3rd International Conference on Belief Functions (BELIEF 2014) (2014)
26. Trabelsi, A., Elouedi, Z., Lefevre, E.: Handling uncertain attribute values in decision tree classifier using the belief function theory. In: Dichev, C., Agre, G. (eds.) AIMS 2016. LNCS (LNAI), vol. 9883, pp. 26–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44748-3_3
27. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001) (2001)