Classification trees based on belief functions

Nicolas Sutton-Charani, Sébastien Destercke and Thierry Denœux

Abstract Decision tree classifiers are popular classification methods. In this paper, we extend to multi-class problems a decision tree method based on belief functions previously described for two-class problems only. We propose three possible extensions: combining multiple two-class trees together and directly extending the estimation of belief functions within the tree to the multi-class setting. We provide experiment results and compare them to usual decision trees.

1 Introduction

Decision trees [2] (classification trees for categorical labels and regression trees for numerical ones) are popular classifiers, due to their simplicity, efficiency and readability. The construction of usual decision trees relies on probability theory. However, classical methods are not always fully adequate to deal with some problems. Among these problems are (1) the fact that all kinds of uncertainties (either in inputs or outputs) cannot be modeled faithfully by classical probabilities and (2) the fact that frequencies of occurrence are only sensible to proportions in a sample and not to its size.

Beyond the fact that the relationship between inputs and outputs may be non-deterministic, a classifier may have to deal with three different possible levels of uncertainty: in inputs, in outputs, and uncertainty due to the fact that the trained classifier is an estimation of the ideal one, due to a limited amount of knowledge or data. In this work, we mainly address the third issue, where the estimation quality translates into imprecision of belief functions.

Belief function theory [13] offers a convenient framework to deal with all these problems. For instance, Elouedi *et al.* [9] propose different ways to adapt decision

UMR CNRS 7253 Heudiasyc Université Technologique de Compiègne, BP 20529 - F-60205 Compiègne cedex - France, e-mail: nicolas.sutton-charani@hds.utc.fr, e-mail: sebastien.destercke@hds.utc.fr, e-mail: thierry.denoeux@hds.utc.fr

trees in the Transferable Belief model (TBM) framework to deal with uncertain outputs during the tree construction. In this work, we extend another approach also using belief functions proposed by Denœux and Skarstein Bjanger [8] that can cope with uncertain outputs and imprecision arising from limited sample size. In this sense, this approach is closer to some imprecise probabilistic approaches [1] that naturally integrate sample size information in their construction.

As Skarstein Bjanger's method only concerns two-class problems, we extend this methodology to any number of classes. For multi-class problems, we propose three ways of doing such an extension:

- combining belief functions provided by sets of two-class trees [12];
- building multinomial belief functions using the Imprecise Dirichlet Model (IDM) [14];
- building multinomial predictive belief functions using Denœux's approach [6].

Section 2 presents the needed background about decision trees and Skarstein Bjanger's method. Section 3 then extends this methodology to the multi-class case. Finally, in Section 4 we compare new classifiers with the usual CART algorithm and discuss the effects of parameters on experiment results.

2 Background

2.1 Decision trees

Let (X,Y) be a random vector where $X = (X_1,...,X_J) \in \mathscr{X} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_J$ represents the features (continuous or discrete) and $Y \in \mathscr{Y} = \{Y_1,\ldots,Y_K\}$ the class to predict. From a sample $E = \{(X^1,Y^1),\ldots,(X^n,Y^n)\}$, decision tree methods build iteratively a model of (X,Y) by building a partition of $\mathscr{X}$. Here, we consider binary trees (i.e., CART-like models), where each split provides two children.

The method works as follows: from a root node containing the whole learning sample, the optimal split (among all the variables and their values) in term of information gain is searched. The information gain IG corresponding to splitting on variable X_k with value α is computed as follows:

$$IG(k,\alpha) = i(t_0) - p_L i(t_1) - p_R i(t_2), \tag{1}$$

where $i(t)$ is an impurity measure of a node t, t_0 the root node, t_1 and t_2 its child nodes, p_L is the proportion of the samples in t_0 verifying the condition $X_k < \alpha$ (i.e., $p_L = {}^{n_L}/_{n}$ where n is the sample size in t_0 and n_L the number of cases such that $X_k < \alpha$). $p_R = 1 - p_L$ is the sample proportion not verifying it. The selected splitting value (k,α) is then the one maximizing IG (resulting in a gain in purity).

The method is then applied recursively to each child nodes until no possible information gain greater than a pre-established threshold can be made. In this case, the node becomes a leaf predicting the most frequent class of the leaf sample.

The information gain (or impurity measure) is calculated using the Gini-index for the CART algorithm or Shanon entropy for C4.5's (Quinlan [11]). Both of these functions measure the homogeneity in term of classes. They both use the frequencies of the different classes in the node samples; however, these frequencies do not depend on the sample size (provided class proportions remain the same). In contrast, Skarstein Bjanger's method impurity measure do change with the sample size.

2.2 Skarstein Bjanger's method for two-class datasets

This method shares CART principles, but differs in the computation of information gain: it uses mass functions instead of simple frequencies and the used impurity measure combines nonspecifity (imprecision) and conflict (variability).

To build the mass functions, Dempster's inference method applied to Bernoulli trials [5] induces the following mass function:

$$\begin{cases} m_{DaBt}(\{Y_1\}) = \frac{n_1}{n+1} \\ m_{DaBt}(\{Y_2\}) = \frac{n_2}{n+1} \\ m_{DaBt}(\mathscr{Y}) = \frac{1}{n+1}, \end{cases} \tag{2}$$

where n is the number of samples and n_1, n_2 are the number of samples whose class is Y_1, Y_2, respectively. Denœux and Skarstein Bjanger then propose to use the following impurity measure [10], applied to m_{DaBt}:

$$U_\lambda(m) = (1-\lambda)N(m) + \lambda D(m) \tag{3}$$

where $N(m) = \sum_{A \subseteq \mathscr{Y}} m(A) \log_2 |A|$ measures the non-specificity and

$$D(m) = -\sum_{A \subseteq \mathscr{Y}} m(A) \log_2 BetP(A)$$

the variability. The two parts are weighted by hyperparameter $\lambda \in [0,1]$. Note that as the size n of the sample increases, $m(\mathscr{Y})$ (the imprecision) decreases. When using U_λ as impurity measure $i(t)$, the information gain (1) can be negative. This gives a natural stopping criterion when building the tree, that is, no split is done if all possible information gains are negative. Usually, λ can be fixed by cross-validation (see Section 4).

Table 1 shows results obtained with CART-classification trees and with classification trees based on Skarstein Bjanger's method. The stopping criteria was the following: keep splitting while $IG > \beta$ for usual CART-trees ($IG > 0$ for the one based on U_λ) and while the children nodes of the split contains a minimum of 10 samples. The usual CART procedure and the U_λ-based algorithm were optimized with respect to the threshold β and parameter λ, respectively, using 10-fold cross-validation. Results show that the methods achieve comparable accuracies.

Table 1 Error rates of trees depending of the used impurity measure

Data set	Number of features	standard CART	trees based on U_λ
Blood transfusion	4	23.5%	24.2%
Statlog heart	13	28%	25.7%
Tic-tac	9	21.5%	11.5%
Breast-cancer	10	5.9%	4.7%
Pima	8	27.3%	25.1%

Dempster's method of inference cannot be easily extended from the binomial to the multinomial case. Therefore, we propose three ways to handle multiple classes: break up the classification problem containing K classes (K $\geq$ 3) into C_K^2 two-class problems using Quost's method for combining binary classifiers [12] and use the Imprecise Dirichlet Model (IDM) approach or Denœux's multinomial model.

3 Multi-class cases

3.1 Combinations of binary classifiers

In [12], Quost presents a method to handle multi-class classification problems by combining classifiers built on sub-samples containing only two classes. He proposes to learn (from the corresponding sub-sample) a conditional belief function for each pair $\{Y_i, Y_j\}$, $1 \leq i < j \leq K$ of classes and to combine them into a global belief function over $\mathscr{Y}$ using an optimisation procedure.

Here, we propose to use this method with decision trees issued from Skarstein Bjanger's method, using the latter as base classifier to learn conditional belief functions. This method is different from the one proposed by Vannoorenberghe and Denœux [15] in which K two-class trees corresponding to a "one vs all strategy" are built, their output being then combined by an averaging of obtained massess.

Decision trees are well adapted to this kind of combination, since they are simple classifiers. However, note that the optimization of λ becomes an issue, as $K(K-1)/2$ classifiers have to be learned at each optimization step.

3.2 IDM

The IDM was introduced in the "imprecise probability" framework by Walley [16]. Note that, although belief functions can be interpreted as imprecise probabilities, it is not their only possible interpretation. However, the IDM turns out to yield a belief function as output, hence it can be used in our framework. The IDM imprecision is controlled by a hyperparameter $s \in \mathbb{R}^+$. From a random sample $Y^1, ..., Y^n$, Walley

showed that the lower predictive probability distribution on $\mathscr{Y}$ is $\underline{P}(Y_k|N,s) = {}^{n_k}/_{n+s}$ where n_k is the number of times Y_k has been observed. The corresponding mass function is such that:

$$\begin{cases} m_{IDM}(Y_j) = n_j/(n+s) & j = 1,...,K \\ m_{IDM}(\mathscr{Y}) = s/(n+s) \end{cases} \tag{4}$$

Note that we recover equation (2) for $K = 2$ and $s = 1$. Using m_{IDM}, U_λ can be applied to measure the impurity in a node and multi-class trees can thus be created. The analytical form of U_λ applied to m_{IDM} can be derived as:

$$U_\lambda(m_{IDM}) = \frac{(1-\lambda)s}{n+s}\log_2(K) - \frac{\lambda}{n+s}\sum_{k=1}^{K} n_k \log_2\left[\frac{Kn_k+S}{K(n+s)}\right] \tag{5}$$

However, even if this model is simple, it is not easy to interpret it within the belief function framework. Also, the IDM imprecision only depends on the sample size n, and not on its distribution over $\mathscr{Y}$. This is not the case for Denœux's multinomial predictive belief function that offers an interesting alternative.

3.3 Denœux's multinomial model

Denœux [6] proposes to use Goodman's confidence intervals to build a predictive belief function. The first step is to build probability intervals [4] (probability lower and upper bounds over singletons) and then to transform them into belief functions.

Let $(X^1,Y^1),...,(X^n,Y^n)$ be an *iid* sample where $Y^k \in \mathscr{Y} = \{Y_1,...,Y_K\}$, those probability intervals $[P_k^-, P_k^+]$ are given, for Y_k (k=1,...,n), as:

$$P_k^- = \frac{q+2n_k-\sqrt{\Delta_k}}{2(n+q)} \quad \textit{and} \quad P_k^+ = \frac{q+2n_k+\sqrt{\Delta_k}}{2(n+q)} \quad , \tag{6}$$

where q is the quantile of order $1-\alpha$ of the chi-square distribution with one degree of freedom, and where $\Delta_k = q(q+\frac{4n_k(n-n_k)}{n})$. As shown in [6], the lower confidence measure (i.e., $P^-(A) = \max(\sum_{Y_k\in A} P_k^-, 1-\sum_{Y_k\notin A} P_k^-)$) built using these regions in the case where $K = 2$ or 3 is a belief function.

Note that the built belief functions follow Hacking's principle (see [6] for details), but the solution for $K = 2$ is not equivalent to that of Eq. (2).

In the case $K > 3$, the Möbius inverse of P^- may take negative values, so P^- is not a belief function in general. Different methods involving linear programming are proposed in [6] to approximate it into a belief function. Also, in the special case where the classes are ordinal, Denœux proposes an algorithm restricted to a certain set of focal elements. A valid predictive *bba* is obtained. These belief functions can then be used with U_λ to build multi-class trees.

4 Experiments

We start by comparing the classifier performances, and then discuss the effect of λ.

4.1 Comparison between classifiers

We compares the three proposed extensions with the usual CART algorithm. Table 2 shows three multi-class UCI datasets characteristics. Table 3 presents experimental results on the previous datasets comparing the accuracy of four types of classifiers:

- Standard CART trees based on Gini index (CART);
- Trees based on U_λ with m_{IDM} (IDM);
- Combination of two-class trees based on U_λ (combination);
- Trees based on U_λ with $m_{Multinomial}$ (multi).

The tree growing strategy is the following: keep splitting while $IG > \beta$ for CART and $IG > 0$ for the tree based on U_λ, the children nodes sample size is greater than 10 and the depth of the tree is smaller or equal to 5.

Because the aim of this experiment was to compare the different methods, none of the trees were optimized: for CART we fixed the threshold $\beta = 0$ and for trees based on U_λ we fixed $\lambda = 0.5$. None of the trees were post-pruned, as we are only interested in accuracies of each model, and not in their simplicity (defining a proper pruning strategy for U_λ based decision trees remains the matter of further research).

Table 2 UCI data sets used in experiments

Data set	Number of features	Number of classes	learning sets size	test sets size
Iris	4	3	113	37
Balance scale	4	3	469	156
Wine	13	3	134	44
Car	6	4	1152	576
Page blocks	10	5	3649	1824
Forest-fires	12	6	345	172

For the datasets with 3 classes we used the belief function induced by P^- whereas the linear programming and the ordinal approaches were used for *Page blocks* and $Forest - fires$, respectively.

The classifiers are competitive; however, as expected, computation times are longer with the multinomial model, due to its higher complexity.

Table 3 Accuracies (R=error rate T=time computation in seconds) of trees depending of the masses assignment model

	CART		IDM		Combi		Multinomial	
datasets	R	T	R	T	R	T	R	T
iris	2.0%	0	2.0%	0	2.0%	1	2.0%	6
balance-scale	20.2%	0	25.0%	0	17.8%	2	15.9%	29
wine	11.9%	0	8.5%	0	13.6%	1	13.6%	19
car	17.7%	1	17.7%	1	15.6%	9	32.3%	8
pageblocks	4.8%	53	4.7%	38	5.0%	140	5.2%	1801
forests-fire	43.6%	1	43.0%	1	43.0%	15	43.0%	81

4.2 Discussion about λ

Figure 1 shows the impact of λ in terms of tree complexity (using the usual number of leaf criterion) and in terms of accuracy on the UCI dataset "Pima". We can see that this complexity increases with λ, confirming that $1-\lambda$ can be interpreted as the importance given to the lack of samples in a node (i.e., to non-specifity $N(m)$) and to the propensity of IG to be negative. This suggests that optimization (here, a 10-fold cross-validation) should also integrate tree complexity as a criterion. The parameter λ seems to have only a small influence on accuracy.

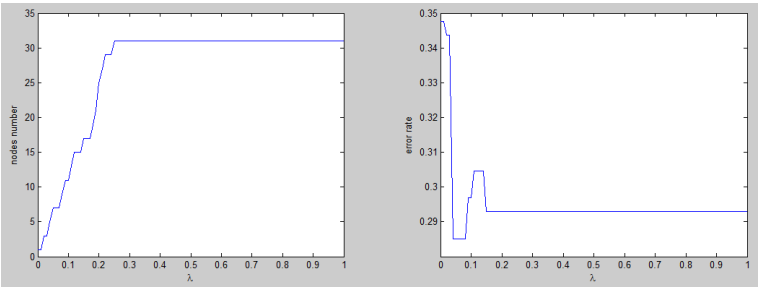

Fig. 1 Number of nodes as a function of λ (left) and error rate as a function of λ (right) for the PIMA dataset

5 Conclusion

In this paper, we have extended Skarstein Bjanger's method for building decision trees to the multi-class case, proposing three ways to do so. The IDM is not really based on the belief function theory and may result in too simple belief functions; Denœux's multinomial model is more elaborated, fits better with a belief function approach, but requires heavier computational efforts; two-class decomposition is

efficient, but makes the interpretation of results possibly harder (and, in any case, longer), as it builds a quadratic number of decision trees.

We have shown that the presented methods have a prediction power comparable to usual methods. However the present work is only a starting point with many perspectives: one of the major interest of using belief functions is the ability to handle uncertain data in inputs or outputs, a feature we shall integrate to the present methods in future works (using, for example, extensions of EM-algorithm to learn trees [3] [7]). Another interesting extension would be to adapt this model to continuous outputs and to regression problems.

References

1. J. Abellan and S. Moral. Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning*, 39(2-3):235–255, June 2005.
2. Breiman, Friedman, Olshen, and Stone. *Classification And Regression Trees*. Wadsworth, Belmont, CA, 1984.
3. A. Ciampi. Growing a tree classifier with imprecise data. *Pattern Recognition Letters*, 21(9):787–803, Aug. 2000.
4. L. de Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *Int. J. Uncertainty Fuzziness Knowledge-Based Syst.*, 1:167–196, 1994.
5. A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1966.
6. T. Denoeux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, Aug. 2006.
7. T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. on Know. and Data Eng. (to appear)*, 2011. doi:10.1109/TKDE.2011.201.
8. T. Denœux and M. S. Bjanger. Induction of decision trees from partially classified data using belief functions. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 4, pages 2923–2928. IEEE, 2000.
9. Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, 28(2-3):91–124, 2001.
10. G. J. Klir. *Uncertainty and information: foundations of generalized information theory*. Wiley-IEEE Press, 2006.
11. J. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, Oct. 1986.
12. B. Quost and T. Denœux. Pairwise Classifier Combination using Belief Functions. *Pattern Recognition Letters*, 28:644–653, 2006.
13. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
14. L. V. Utkin. Extensions of belief functions and possibility distributions by using the imprecise dirichlet model. *Fuzzy Sets and Systems*, 154(3):413–431, 2005.
15. P. Vannoorenbergue and T. Denœux. Handling uncertain labels in multiclass problems using belief decision trees. In *IPMU 2002*, volume 3, pages 1919–1926, Annecy, France, 2002.
16. P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society*, Series B,:3–57, 1996.