Classification trees based on belief functions

Nicolas Sutton Charani, Sébastien Destercke, Thierry Denoeux

Classification trees based on belief functions

Nicolas Sutton-Charani and Thierry Denoeux and Sébastien Destercke

Abstract Decision trees 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 2-class problems only. We propose two ways to achieve this extension: combining multiple 2-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 classical 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 classical 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 input or output) cannot be modeled faithfully by classical probabilities and (2) the fact that frequencies of occurrence are only sensible to proportions in sample and not to its size.

Outside the fact that the relation between input and outputs may be non-deterministic, a classifier may have to deal with three different possible levels of uncertainty: in

Nicolas Sutton-Charani
UMR CNRS 6599 Heudiasyc Université Technologique de Compiègne, BP 20529 - F-60205 Compiègne cedex - France, e-mail: nicolas.sutton-charani@hds.utc.fr

Thierry Denoeux
UMR CNRS 6599 Heudiasyc Université Technologique de Compiègne, BP 20529 - F-60205 Compiègne cedex - France, e-mail: thierry.denoeux@hds.utc.fr

Sébastien Destercke
UMR CNRS 6599 Heudiasyc Université Technologique de Compiègne, BP 20529 - F-60205 Compiègne cedex - France, e-mail: sebastien.destercke@hds.utc.fr

inputs, in outputs, and uncertainty due to the fact that the learned 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 kind, 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.* [8] propose different ways to adapt decision trees in the TBM framework to deal with uncertain outputs during the tree construction. In this work, we extend another approach also using belief functions proposed by Denoeux and Skarstein Bjanger [7] 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 the Denoeux and Skarstein Bjanger method only concerns 2-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 2-class trees [12]
- direcltly building multinomial belief functions using the Imprecise Dirichlet Model (IDM) [14]
- direcltly building multinomial belief functions using Denoeux's proposal [5]

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

2 Background

In this section, we present the necessary 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 provide two children.

The method works as follow: 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 samples 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 (computed as a gain in purity).

The method is then applied recursively on each child nodes until no possible information gain superior to a preestablished 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 Gini-index for CART algorithm or Shanon entropy for C4.5's one (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, the method of Denoeux and Skarstein Bjanger and its impurity measure, that we recall now, do change with the sample size.

2.2 Denoeux and Skarstein Bjanger method for 2-class datasets

This method use the same principle as CART, but differs in the computation of information gain: first, they use mass functions instead of simple frequencies and second, they use an impurity measure mixing non-specifity (imprecision) and conflict (variability).

To build the mass functions, they use Dempster's approach to Bernouilli (*DaBt*) trials that 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. They 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 $N(m) = \sum_{A \subseteq \mathscr{Y}} m(A) \log_2 |A|$ is 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 $\lambda \in [0,1]$. 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 value can be fixed by an optimization procedure (see Section 4).

Table 1 shows results obtained with CART-classification trees and with classification trees based on Denoeux and Skarstein Bjanger method. The stopping criteria is the following: keep splitting while $IG > \beta$ for classical 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. Classical CART are optimized on the threeshold β and U_λ-based on the parameter λ, using a 10-fold cross-validation procedure on training samples. Results show that the methods achieve comparable accuracies.

Data set	Number of features	standard CART error rate	trees based on U_λ error rate
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%
Haberman	3	26.6%	26%

Table 1 Accuracies of trees depending of the used impurity measure

Dempster's approach to Bernouilli's trial is not adapted to cases where the output has more than 2 classes. Therefore, we propose 3 ways to handle such cases: break up the classification problem containing K classes (K $\geq$ 3) into C_k^2 2-classes problems using Quost combination of binary classifiers [12], use the IDM approach or the Denoeux multinomial model.

3 Multi-classes cases

3.1 Combinations of binary classifiers

In [12], Quost presents a way to handle multi-class classification problems by combining classifiers built on sub-samples containing only 2 classes. That is, 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 combine this method with decision trees issued from Denoeux and Skarstein Bjanger method, using the latter as base classifier to learn conditional belief functions (those belief functions are assigned by the DaBt applied to leaves class proportions). Decision trees are well adapted to this kind of combination, since they are simple classifiers. However, note that λ optimization becomes an issue, as $K(K-1)/2$ (i.e., a quadratic number) of classifiers have to be learned at each optimization step.

3.2 IDM

The IDM was introduced in the "imprecise probability" framework by Walley [15]. Note that although belief functions can be interpreted as imprecise probabilities, this is far from being their only possible interpretation (and it is the one adopted here). However the IDM turn out to yield a belief function as output, hence it can be used in our framework. The IDM imprecision is controlled by an hyper-parameter $s \in \mathbb{R}^+$. 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 find back 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. Analytical form of U_λ applied to m_{IDM} can be derived as:

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

However even if this model is simple, it's not easy to interpret it within the belief function frameworks. Also, it should be noted that the IDM imprecision only depends on the number of samples n, and not on their distribution over $\mathscr{Y}$. This is not the case for the multinomial construction of Denoeux that offers a tractable and interesting alternative.

3.3 Denoeux multinomial model

Denoeux [5] proposes to use Goodmans confidence intervals [9] to build a multinomial 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:

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

where q is the quantile of order $1-\alpha$ of the chi-square distribution with one degree of freedom, and where $\Delta_k = q(q+{}^{4n_k(n-n_k)}/_n)$. As shown in [5], 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 are belief functions, hence induce corresponding mass functions.

Note that the built belief functions follow the Hacking principle (see [5] for details), but that the solution for $K = 2$ is this time not equivalent to Eq. (2).

In the case $K > 3$, the Möbius inverse of P^{-} takes some negative values so P^{-} is not a belief function. Different methods involving linear programing are proposed in [5] to approximate it into a belief function. Also, in the special case where the classes are ordinal, Denoeux 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 classifiers performances, and then discuss the effect of the λ parameter in U_λ.

4.1 Comparison between classifiers

We compare the three proposed extensions with the classical probabilistic CART algorithm. Table 2 shows 3 multi-class UCI datasets characteristics. Table 3 presents experimental results on the previous datasets comparing the accuracy of 4 types of classifier:

- standard CART trees based on Gini index (CART)
- trees based on U_λ with m_{IDM} (IDM)
- combination of 2-classes 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

CART trees and the ones using U_λ are optimized by a 10-fold cross-validation procedure: for CART we optimize the threshold β and for trees based on U_λ we optimize λ. None of the trees are post-pruned, as we are only interested in accuracies of each models, and not in their simplicity (defining a proper pruning strategy for U_λ based decision trees remains the matter of further research).

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

Table 2 UCI data sets used in experiments

Data set	CART	IDM	combination	multi
Iris	4.1%	4.1%	4.1%	21.3%
Balance scale	23.9%	25.5%	25.5%	21.3%
Wine	13.6%	10.2%	11.9%	15.3%

Table 3 Accuracies of trees depending of the masses assignement model

As we can see the different classifiers' accuracies are competitive although the time computations are longer with the ones using belief functions probably because of the number of focal elements wich can be higher than with probabilities (where they are restricted only to singletons)

4.2 Discussion about λ

Figures 1 study the impact of λ in terms of tree complexity (using the classical number of leafs criterion) and in terms of accuracy on the UCI dataset "Breast Tissue". They show 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 procedure) should also integrate tree complexity as a criterion. The best accuracy level is here obtained for small lambdas which give small trees. Accuracy seems to be little influenced by lambda value, suggesting that lambda value should be kept low.

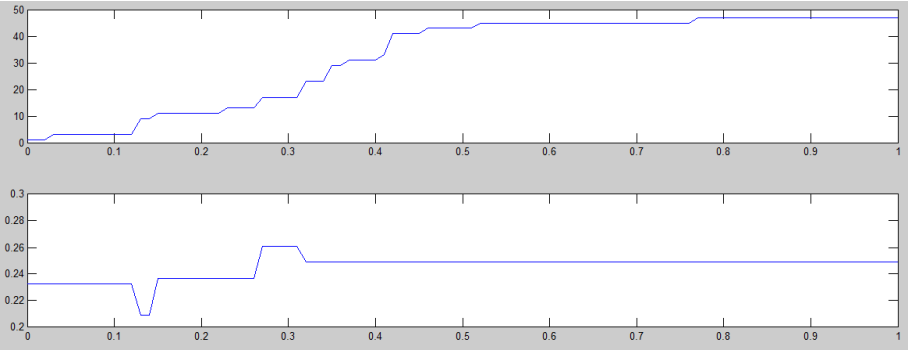

Fig. 1 number of nodes/λ (top) and error rate/number of nodes (bottom)

5 Conclusion

In this paper, we have extended Denoeux and Skarstein Bjanger method to the multi-class case, proposing three ways to do so. The IDM is not really based on belief function and may result in too simple belief functions; Denoeux multinomial model is more elaborated, fits better with a belief function approach, but requires heavy computational efforts (asking to solve a linear program for each split test); 2-class decomposition is efficient, but makes the interpretability 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 prediction power comparable to classical 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, 6]). Another interesting extension would be to adapt this model to continuous outputs and to regression problems. Additional experiments exploring the method behavior should also be done.

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*. 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. T. Denoeux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, Aug. 2006.
6. T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. on Know. and Data Eng. (to appear)*, 2011.
7. T. Denoeux and M. 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.
8. Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, 28(2-3):91–124, 2001.
9. L. A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, 1965.
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. Denœ ux. 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. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society*, Series B,:3–57, 1996.