Induction of decision trees from partially classified data using belief functions*

T. Denœux and M. Skarstein Bjanger
Université de Technologie de Compiègne
UMR CNRS 6599 Heudiasyc
BP 20529 F-60205 Compiègne cedex, France

Abstract

A new tree-structured classifier based on the Dempster-Shafer theory of evidence is presented. The entropy measure classically used to assess the impurity of nodes in decision trees is replaced by an evidence-theoretic uncertainty measure taking into account not only the class proportions, but also the number of objects in each node. The resulting algorithm allows the processing of training data whose class membership is only partially specified in the form of a belief function. Experimental results with EEG data are presented.

1 Introduction

Most of the work in pattern recognition and machine learning has focused on the induction of decision rules from learning examples with known classification. In certain real-world problems, however, such "perfect" information is not always available. Instead, one may have an "uncertain" training set of objects with partially known classification. For instance, an expert or a group of experts may have expressed conflicting opinions regarding the class of objects contained in a data base. In Ref. [3, 4], the Dempster-Shafer theory of belief functions was shown to provide a convenient framework for dealing with such learning problems. A distance-based approach was proposed, whereby a belief function for a pattern is constructed by combining the evidence of neighboring prototypes in a data set. This method was shown to behave equally well in the presence of data with precise or imprecise class labels.

In this paper, the problem of learning from partially classified data is addressed from a different perspective using a new approach to decision tree (DT) induction based on the theory of belief functions [1]. Like most tree-based classification techniques [2, 7], our method recursively partitions the feature space into subregions corresponding to the leaves of the tree. The specificity of the proposed algorithm lies in the splitting rule applied at each step, and in the pruning strategy, which use concepts from Evidence theory. At each node t of the tree, we construct a belief function quantifying one's belief concerning the class of an example reaching t, using results by Smets concerning parametric inference for the Bernouilli distribution in an Evidential framework [10]. The impurity of each node is then assessed using an evidence-theoretic uncertainty measure, which happens to depend not only on the class proportions in the node, but also on its size, thus allowing to control the complexity of the tree. An interesting feature of this method is its ability to deal with training data, the class membership of which is uncertain or imprecise, and is described by a belief function.

The paper is organized as follows. The necessary background on belief functions and their use for statistical inference is recalled in Section 2. Our method is then explained in Section 3 and experimental results are presented in Section 4. Section 5 concludes the paper. Note that some familiarity of the reader with decision trees will be assumed throughout the paper. Two standard references on this topic are the books by Breiman *et al.* [2] and Quinlan [7].

2 Background

2.1 Belief functions

A belief function on a finite set Ω is a subadditive measure of the form

$$\mathrm{bel}(A) = \sum_{\emptyset \neq B \subseteq \Omega} m(B) \quad \forall A \subseteq \Omega, \tag{1}$$

where m is a basic belief assignment, also called a belief structure (BS), i.e., a function from 2^Ω to $[0, 1]$

*Proc. of SMC'2000, Nashville, Oct. 2000, IEEE.

verifying $\sum_{A\subseteq\Omega} m(A) = 1$. It can be shown that a belief function is induced by a unique belief structure, so that m and bel can be considered as different forms of a single mathematical object[1].

The idea of using belief functions for modeling partial belief and reasoning under uncertainty was introduced by Shafer [9]. In the last 25 years, Shafer's work gave rise to an important literature about the so-called Dempster-Shafer (D-S) theory, in which belief functions received different interpretations (e.g. in a random set or an imprecise probability setting), which sometimes obscured the debate about their usefulness and their relationship to Probability Theory [11]. In this paper, we shall adopt the point of view of Smets' Transferable Belief Model (TBM), a non probabilistic and subjectivist interpretation of D-S theory in which the state of belief of a rational agent, with respect to a certain question, is assumed to be represented by a belief function, defined independently from any probabilistic notion [11]. This model postulates the existence of two levels: a *credal* level at which beliefs are entertained and updated in view of incoming evidence, and a *pignistic* level in which belief functions are converted into probability functions for decision making purposes.

The basic mechanism for combining two belief functions induced by distinct information sources is Dempster's rule of combination [9]. This rule can be conveniently expressed by means of commonality functions. The commonality function q induced by a BS m is defined as

$$q(A) = \sum_{B\supseteq A} m(B) \quad \forall A \subseteq \Omega. \tag{2}$$

If q_1, q_2 and q are the commonality functions induced, respectively, by m_1, m_2 and $m = m_1 \cap m_2$ (the combination of m_1 and m_2), we have $q(A) = q_1(A)q_2(A)$ for all $A \subseteq \Omega$, and m or bel may be recovered from q using simple formula [9, p. 41].

At the pignistic level, a BS m is converted into a so-called *pignistic* probability function BetP defined as

$$\mathrm{BetP}(\omega) = \sum_{\{A\subseteq\Omega, \omega\in A\}} \frac{m^*(A)}{|A|}, \tag{3}$$

where m^* is the normalized BS induced by m (defined by $m^*(A) = m(A)/(1-m(\emptyset))$ for $A \neq \emptyset$ and $m^*(\emptyset) = 0$), and $|A|$ denotes the cardinality of A.

The D-S framework can be nicely extended to the case where $\Omega = \mathbb{R}$ by assigning basic probability masses to closed intervals $[x, y]$ by means of a basic belief density (BBD) function $m([x, y])$ (see, e.g., [10]). Belief and commonality densities are then defined, respectively, as

$$\mathrm{bel}([a,b]) = \int_a^b \int_x^b m([x,y])dxdy \tag{4}$$

$$q([a,b]) = \int_0^a \int_b^1 m([x,y])dxdy \tag{5}$$

and we have

$$m([a,b]) = -\frac{\partial^2 \mathrm{bel}([a,b])}{\partial a \partial b} = -\frac{\partial^2 q([a,b])}{\partial a \partial b}. \tag{6}$$

Further essential material on belief functions may be found in Refs. [9, 11].

2.2 Beliefs induced by Bernouilli trials

Let us assume that we have a random experiment (a Bernouilli trial) with two outcomes (success or failure, denoted by S and F), such as drawing a ball from an urn containing white and black balls. Associated with this experimental setting is an objective probability function P on $\Omega = \{S, F\}$. If it is known that $P(S) = p$, then one's belief that the outcome will be a success can reasonably be assumed to be equal to p (this is called the Hacking Frequency principle by Smets [10]). In most cases of interest, however, P is unknown, and all the available information resides in observed outcomes from n independent experiments (such as the colors of n balls drawn from the urn with replacement). Given such partial information, *how can we compute a belief function on Ω quantifying one's belief that the next outcome will be a success ?* An answer to this question was provided by Smets in the TBM framework [10]. The argument is technically involved and only the main findings will be summarized here. The reader is invited to refer to Smets' paper for a detailed presentation.

Let $\mathbb{P}_\Omega$ denote the set of probability functions on Ω, and $W = \mathbb{P}_\Omega \times \Omega$. Since Ω only has 2 elements, each probability function $P \in \mathbb{P}_\Omega$ can be indexed by $P(S) \in [0, 1]$, so that $\mathbb{P}_\Omega$ can be identified with the interval $[0, 1]$. The basic idea is to deduce from first principles (such as the Hacking principle) a BS m_W on W quantifying one's beliefs in the absence of any information. The impact of additional evidence (such as the observation of past outcomes) is then reflected by the updating of m_W using Dempster's rule of combination, and a belief function m_Ω is deduced

[1] The normality condition $m(\emptyset) = 0$ originally imposed by Shafer is not generally assumed in the TBM.

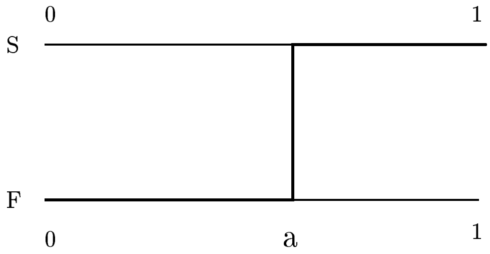

Figure 1: Structure of the domain of m_W and one example of a focal element $A(a)$ centered on a.

by marginalization over Ω. As shown by Smets, it follows from simple requirements (including the Hacking principle) that the focal element of m_W are of the form $A(a) = ([0,a) \times \{F\}) \cup ([a,1] \times \{S\})$, as illustrated in Figure 1. To simplify the notation, we shall denote by $m_W([a,1])$ the mass given to $A(a)$. It may be shown that $m_W([a,1]) = 1$ for all $a \in [0,1]$. The following facts then result from the definition of m_W.

(1) Suppose we learn that the "true" probability function P is such that $a \leq P(S) \leq b$. The impact of this evidence on m_W is obtained by conditioning[2] m_W on the cylindrical extension of $[a,b]$, defined by $\text{cyl}([a,b]) = [a,b] \times \Omega$; the BBD $m(A)$ for $A \subseteq W$ is thus transferred to $A \cap \text{cyl}([a,b])$. Let us denote by $\text{bel}_\Omega(S|[a,b])$ the belief that the outcome will be a success, given that $p \in [a,b]$. This degree of belief is equal to the integral of all the BBD that touch only S after conditioning on $\text{cyl}([a,b])$, which leads to

$$\text{bel}_\Omega(S|[a,b]) = \int_0^a m_W([x,1])dx = a. \tag{7}$$

(2) Suppose now that all you know is that an experiment has been carried out and a success has been observed. The impact of this evidence is reflected by conditioning m_W on S, which has the effect of transferring the non-null BBD to the intervals $[a,1]$. The commonality function induced on $\mathbb{P}_\Omega$ is then

$$q_{\mathbb{P}_\Omega}([a,b]|S) = a. \tag{8}$$

Similarly, if a single failure has observed, conditioning m_W on F leads to

$$q_{\mathbb{P}_\Omega}([a,b]|F) = 1 - b. \tag{9}$$

(3) If n independent experiments have been performed, and r successes and s failures have been observed, the resulting commonality function may be obtained, as a consequence of Dempster's rule, by multiplying the n corresponding commonality functions, leading to

$$q_{\mathbb{P}_\Omega}([a,b]|r,s) = a^r(1-b)^s. \tag{10}$$

By derivating this expression with respect to a and b according to Eq. (6), we obtain after appropriate normalization :

$$m_{\mathbb{P}_\Omega}([a,b]|r,s) = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(s)} a^{r-1}(1-b)^{s-1}, \tag{11}$$

where Γ is the gamma function.

(4) Finally, assume that you want to compute the belief that the next outcome will be a success, given that you have already observed r successes and s failures in $n = r + s$ independent trial. We have

$$\begin{aligned}\text{bel}_\Omega(S|r,s) &= \int_0^1 \int_a^1 \text{bel}_\Omega(S|[a,b]) m_{\mathbb{P}_\Omega}([a,b]|r,s)dbda \\ &= \frac{r}{n+1}.\end{aligned} \tag{12}$$

Similarly, we have $\text{bel}_\Omega(F|r,s) = s/(n+1)$, and, consequently, $m_\Omega(\Omega|r,s) = 1/(n+1)$. Note that $\text{bel}_\Omega(S|r,s)$ tends to the true probability of success when n tends to infinity.

3 Application to DT induction

3.1 Principles of DT induction

A decision tree (DT) is a sequential classification procedure in which the attributes describing an object are examined one at a time until one reaches a decision regarding the assignment of the object to a class [2, 7]. The root of a DT is the top node, and examples are passed down the tree with decisions being made at each node, until a terminal node, or leaf, is reached. Each leaf has a class label, and each example is classified by the label of the leaf it reaches. A DT thus partitions the attribute space in a hierarchical manner.

Usually, DT's are grown from a training set of examples with known classification, by successively splitting leaves. The process stops when the tree classifies correctly every learning example. In a noisy environment, a pruning rule is generally applied to prevent overfitting. DT induction algorithms differ essentially by the splitting rule and the pruning strategy used. A common strategy for splitting nodes is to define an "impurity" measure for each node and ask that child

[2] See [11] for a detailed presentation of the conditioning mechanism in the TBM.

nodes be "purer" than their parent. A common impurity measure is the entropy

$$i(t) = -\sum_j \frac{n_j(t)}{n(t)} \log_2 \frac{n_j(t)}{n(t)}, \qquad (13)$$

where $n_j(t)$ is the number of examples from class j in node t, and $n(t)$ is the size of node t. Consider a candidate split s which divides node t into t_L and t_R, such that a proportion p_L of the cases go to t_L and a proportion p_R goes to t_R. Then the goodness of the split may be defined as the decrease in impurity:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R). \qquad (14)$$

For each attribute, the best split is searched for, and the attribute allowing to reduce the degree of impurity by the largest amount is selected.

3.2 The TBM approach

In this paper, we propose to reconsider the problem of DT induction using the TBM framework[3]. Let $\Omega = \{\omega_1, \omega_2\}$ denote the set of classes (only two-class problems will be considered here; a way to deal with multi-class problems will be briefly discussed in the sequel). Let us assume for the moment that all training examples have known classification: among the learning examples which have reached node t, it is known that $n_1(t)$ belong to class 1, and $n_2(t)$ belong to class 2. Using the inference mechanism presented in Section 2.2, it is possible to use this information to determine our belief concerning the class of a previously unseen example, if we only know that it has reached note t. Using Eq. (12), this is defined as

$$\mathrm{bel}_\Omega(\omega_j|t) = \frac{n_j(t)}{n(t)+1}, \quad j = 1,2 \qquad (15)$$

$$m_\Omega(\Omega|t) = \frac{1}{n(t)+1}. \qquad (16)$$

This belief function, with clear interpretation, may be used to define a new impurity measure. In the same way that the Shannon entropy was used in Eq (13) to describe the empirical probability function $\{n_j(t)/n(t)\}$, we propose to use entropy-like criteria for quantifying the uncertainty of belief functions [6]. As remarked by Klir, a belief function actually models two different kinds of uncertainty: nonspecificity and conflict [6]. For instance, the vacuous belief function defined by $m(\Omega)$ has maximal nonspecificity but no conflict, whereas the uniform probability function on Ω has maximal conflict but is fully specific (since belief masses are assigned to singletons). A measure of nonspecificity that appears to be well justified is

$$N(m) = \sum_{A \subseteq \Omega} m(A) \log_2 |A|, \qquad (17)$$

which is maximal for the vacuous BF, and 0 for probability functions. To quantify the degree of conflict in a belief function, several extensions of the Shannon entropy may be defined. One of these extensions is the degree of discord, defined by

$$D(m) = -\sum_{A \subseteq \Omega} m(A) \log_2 \mathrm{BetP}(A), \qquad (18)$$

which is maximal for the uniform probability distribution on Ω. A composite measure of uncertainty may be defined using a convex sum of N and D, of the form

$$U_\lambda(m) = (1-\lambda) N(m) + \lambda D(m), \qquad (19)$$

where $\lambda \in [0,1]$ is a positive coefficient (Klir proposes to give equal weights to both terms, but we shall see that greater flexibility may be useful).

A measure of "impurity" or "uncertainty" $u_\lambda(t)$ for a node t is obtained by applying the uncertainty measure (19) to the BS defined by Eqs (15)-(16), which leads to

$$u_\lambda(t) = U_\lambda(m_\Omega(\cdot|t)) = \frac{1-\lambda}{n(t)+1} - \lambda \sum_{j=1}^{2} \frac{n_j(t)}{n(t)+1} \log_2 \left(\frac{2n_j(t)+1}{n(t)+1} \right). \qquad (20)$$

Note that the first term in the left-hand side of the above equation corresponds to the nonspecificity of $m_\Omega(\cdot|t)$ and is a decreasing function of $n(t)$, whereas the second term depends on both the size of t, and the class proportions in t. We define the goodness of a split s as

$$\Delta u_\lambda(s,t) = u_\lambda(t) - p_L u_\lambda(t_L) - p_R u_\lambda(t_R),$$

with the same notations as in Eq (14). Note that parameter λ allows to control the tree growing strategy: low values of λ penalize small nodes, which leads to early stopping of the tree growing process. In practice, λ may be determined by cross-validation (see Section 4).

Once the decision tree has been built, a BS $m_\Omega(\cdot|t)$ is associated to each leaf t. This BS quantifies one's beliefs regarding the class of an arbitrary pattern reaching that leaf.

[3] An alternative approach to decision tree generation in the TBM framework has been investigated independently by Elouedi *et al.* [5].

3.3 Handling of uncertain labels

Let us now assume that we have a learning set of the form $\{(\mathbf{x}_i, m_i), i = 1, n\}$, where $\mathbf{x}_i$ is the feature vector for example i, and m_i is a BS on Ω describing one's partial knowledge regarding the class of that example. The classical situation of precise labeling is recovered when $m_i(\{\omega\}) = 1$ for some $\omega \in \Omega$. Complete ignorance regarding the class membership corresponds to $m_i(\Omega) = 1$. We can thus model a whole range of situations from fully supervised to fully unsupervised learning.

To see how this more general learning problem can be solved by our method, let us return to the inferential framework defined in Section 2.2. Suppose that we have performed n independent Bernouilli experiments, but that the outcomes could only be partially observed (for example, the urn experiment was observed at a distance, so that the results of some trials could only be partially observed). Let m_i be the BS describing one's belief concerning the result of experiment i. Then Eqs (8) and (9) should be replaced by

$$q_{\mathbb{P}_\Omega}([a,b]|m_i) = am_i(S) + (1-b)m_i(F) + m_i(\Omega).$$

After combining the evidence from the n experiments by Dempster's rule we get

$$\begin{aligned} q_{\mathbb{P}_\Omega}([a,b]|m_1, \ldots, m_n) &= \prod_{i=1}^{n} q_{\mathbb{P}_\Omega}([a,b]|m_i) \\ &= \sum_{j+k\leq n} \alpha_{jk} a^j (1-b)^k, \end{aligned} \tag{21}$$

where the α_{jk} are coefficients depending only on the m_i (the coefficients can easily be computed by induction on n). After derivation and integration as in Eqs (11) and (12), we finally obtain

$$\mathrm{bel}(S|m_1, \ldots, m_n) = \sum_{j+k\leq n} \alpha_{jk} \frac{j}{j+k+1}, \tag{22}$$

and similar expression for $\mathrm{bel}(F|m_1, \ldots, m_n)$ and $m(\Omega|m_1, \ldots, m_n)$.

This result can be immediately transferred to the context of DT generation. Let us assume that we have $n(t)$ examples in node t, with labels m_i, $i = 1, n_t$. Then Eq (22) allows the calculation of a BF $\mathrm{bel}_\Omega(\cdot|t)$ quantifying our belief concerning the class of an example reaching node t. The impurity measure for node t is defined as above. It can be verified that unlabeled examples (i.e., examples such that $m_i(\Omega) = 1$) can be added to or removed from node t without changing the value of $u_\lambda(t)$, as it should be, since such examples carry no information regarding the classification problem at hand.

Table 1: Results with crisp and uncertain class labels.

	crisp labels	unc. labels
error rate	0.35	0.34
E	0.26	0.22

4 Results

Detailed results from preliminary experiments with the above method are given in [1]. These results are only briefly summarized here. The learning task considered was to detect different waveforms in sleep electroencephalogram (EEG) data, and in particular to discriminate between K-complex and delta waveforms. For a thorough presentation of this problem, see [8]. The data used in this experiment were EEG signals measured 64 times during 2-second intervals for one person during sleep[4]. Each object was then described by 64 attributes. Since the K-complex pattern is difficult to detect visually even by domain experts, five physicians were asked to inspect graphical displays of the data and state whether they believed a K-complex signal was present or not. As the experts did not always agree on the classification, this introduced uncertainty in the labeling of the objects. Our data base consisted of (1) EEG patterns classified in the K-complex class by at least one expert, and (2) delta wave patterns, which are known to bear some resemblance with K-complex waves, although they are related to different phenomena. One of the data sets considered in our study was composed of 50 delta wave patterns, and 100 K-complex pattern, of which only one half had been classified as such by a majority of experts.

Uncertain class labels were assigned to the training examples in the following manner. Let ω_1 and ω_2 denote, respectively, the delta wave and the K-complex class. Delta wave examples certainly belong to that class and were assigned labels m_i with $m_i(\{\omega_1\}) = 1$. For the K-complex patterns, the proportion q of experts classifying each example i in the K-complex class was used to define a BS m_i assigning the mass $m_i(\{\omega_2\}) = q$ to that class, and the rest of the mass to Ω.

The measurement of classification efficiency is not easy in such a context, because, in the case of data with uncertain class membership, disagreement between the

[4]These data come from the Foundation for Applied Neuroscience Research in Psychiatry in Rouffach, France, and were provided to us by C. Richard and R. Lengellé from Universit de Technologie de Troyes.

classifier output and the class label does not necessarily indicate an error. Intuitively, errors made for patterns whose class membership is uncertain should "count less" than errors made for patterns with completely known classification. With this in mind, the following error criterion was introduced:

$$E = \frac{1}{n'} \sum_{i=1}^{n'} (1 - \widehat{\mathrm{BetP}}_i(m_i)), \tag{23}$$

where n' is the size of the test set, $\widehat{\mathrm{BetP}}_i$ is the pignistic function induced by the output BS $\widehat{m}_i$ for test example i, and $\widehat{\mathrm{BetP}}_i(m_i)$ is defined as

$$\widehat{\mathrm{BetP}}_i(m_i) = \sum_{A \subseteq \Omega} m_i(A) \widehat{\mathrm{BetP}}_i(A). \tag{24}$$

Note that examples i such that $m_i(\Omega) = 1$ have zero contribution to the sum in Eq (23), and therefore do not participate in the performance evaluation.

We used a 5-fold cross-validation to find the value of λ providing the optimal tree in each case, and to evaluate the performance of that tree. Table 1 shows the cross-validation estimates of our method both in terms of standard error rate and generalized error rate defined in Eq (23). Also shown in this table are the results obtained with our method when the uncertainty in the class labels was ignored, i.e., the learning examples being then assigned crisp labels regardless of the uncertainty pertaining to the class membership of these examples. The error rate of our method applied to data with crisp label (35 %) is equivalent to that of Quinlan's C4.5 algorithm (not shown in Table 1). However, taking into account the uncertainty in class labels (which is not possible using standard DT generation techniques) does improve the classification performance for this problem, marginally in terms of error rate, but significantly in term of the more meaningful error measure defined here.

5 Conclusions

A new tree-structured classifier based on the Dempster-Shafer theory of evidence has been described. The method is applicable to partially classified data, in which the class labels are provided in the form of belief functions. Once a decision tree has been built, the method allows to compute a belief function describing the uncertainty pertaining to the class of any pattern under consideration. Although the method was presented in the case of two classes, it can be applied to more general situations by converting a c-class problem (with $c > 2$) into several two-class problems, and combining the results at the belief function level.

References

[1] M. Skarstein Bjanger. Induction of decision trees from partially classified data using belief functions. Master's thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, February 2000.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[3] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[4] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[5] Z. Elouedi, K. Mellouli, and P. Smets. Decision trees using the belief function theory. In *Proceedings of IPMU'2000*, Madrid, Spain, July 2000.

[6] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory.* Springer-Verlag, New-York, 1998.

[7] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[8] C. Richard. *Une méthodologie pour la détection à structure imposée. Applications au plan temps-fréquence.* PhD thesis, Université de Technologie de Compiègne, 1998.

[9] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.

[10] P. Smets. Belief induced by the partial knowledge of the probabilities. In D. Heckerman *et al.*, editor, *Uncertainty in AI'94*, pages 523–530. Morgan Kaufmann, San Mateo, 1994.

[11] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.