A Pre-Pruning Method in Belief Decision Trees

Zied Elouedi
Institut Supérieur de Gestion de Tunis,
41 Avenue de la liberté,
2000 Le Bardo,
Tunis, Tunisia
zied.elouedi@isg.rnu.tn

Khaled Mellouli
Institut Supérieur de Gestion de Tunis,
41 Avenue de la liberté,
2000 Le Bardo,
Tunis, Tunisia
khaled.mellouli@ihec.rnu.tn

Philippe Smets
IRIDIA, Université Libre de Bruxelles,
50 av. Roosevelt,
1050 Brussels,
Belgium
psmets@ulb.ac.be

Abstract

The belief decision tree approach is a decision tree method adapted in order to handle uncertainty about the actual class of the objects in the training set. The uncertainty is represented by the Transferable Belief Model (TBM). We present two methods to build the tree.

In order to reduce the size and the complexity of the induced tree, we present a pre-pruning tool related to the stopping criteria used during the development of the paths.

Keywords: belief decision tree, decision tree, transferable belief model, pre-pruning, classification.

1 Introduction

Decision trees are considered as one of the efficient classification techniques applied in several fields, in particular in artificial intelligence applications. Basically, we have a training set composed of objects where each one is described by attributes and its assigned class which is unique. The output will be a decision tree ensuring the classification of new objects. We call the set of these new objects a testing set.

Classically the building of a decision tree follows a recursive top-down procedure, partitioning at each level of the tree the training (sub) set into subsets equal to the number of values of the chosen attribute. The choice of the attribute as the root of the induced (sub) tree is made according to an attribute selection measure.

In the C4.5 algorithm proposed by Quinlan [9], the selected attribute is the one presenting the highest gain ratio. Once an attribute is chosen, a branch relative to each value of the selected attribute will be created. The data are allocated to a node according to the value of the selected attribute. This node is declared as a leaf when the gain ratio values of the remaining attributes do not present any improvement or there is no attribute to test.

As pointed out in several researches [3] [4] [5] [6] [8], a major problem faced by the standard decision tree algorithms is related to the uncertainty that may affect data in the training set. In this paper, we consider the case where there is uncertainty about the actual class of the objects in the training set. This uncertainty is represented by a belief function as understood in the transferable belief model (TBM). In order to cope with such uncertainty, we have developed a belief decision tree method (BDT). In that tree, we implement a pre-pruning method in order to reduce the complexity of the tree. It is based on an idea found in [1] and used in a context of upper and lower probability . It turns out their idea corresponds to a discounting in the TBM and could thus be tailored for the BDT.

This paper is organized as follows: section 2 provides a brief description of the basics of TBM. In section 3, we describe the two attribute selection measures developed

for building a BDT. Then, in section 4, we present the description of the building and classification procedures and pre-pruning mechanism. Finally in section 5, we carry simulations to illustrate the effect of the pre-pruning method.

2 Transferable belief model

In this section, we briefly review the main concepts underlying the transferable belief model [13] [14] [15], one interpretation of the belief function theory [10].

2.1 Definitions

The TBM is a model to represent quantified beliefs based on belief functions. Let Θ be a finite set of elementary events, called the frame of discernment. The basic belief assignment (bba) is a function $m : 2^\Theta \rightarrow [0, 1]$ such that:

$$\sum_{A \subseteq \Theta} m(A) = 1$$

The value $m(A)$, named the basic belief mass (bbm), represents the portion of belief committed exactly to the event A and nothing more specific. The events having positive bbm's are called focal elements. Associated with m is the belief function [13] defined for $A \subseteq \Theta, A \neq \emptyset$ as:

$$bel(A) = \sum_{\emptyset \neq B \subseteq \Theta} m(B) \text{ and } bel(\emptyset) = 0$$

The degree of belief $bel(A)$ given to a subset A of the frame Θ is defined as the sum of all the basic belief masses given to subsets that support A without supporting its negation.

Another function is used basically to simplify computations namely the commonality function q is defined as:

$$q(A) = \sum_{A \subseteq B} m(B), \ \forall A \subseteq \Theta$$

2.2 Combination

In the transferable belief model, the basic belief assignments induced from distinct pieces of evidence are combined by the conjunctive rule of combination defined as [11]:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B,C \subseteq \Theta : B \cap C = A} m_1(B).m_2(C)$$

$m_1 \textcircled{\cap} m_2$ is the bba representing the combined impact of the two pieces of evidence.

2.3 Discounting

The technique of discounting allows to take in consideration the reliability of the information source that generates the bba m.

For $\alpha \in [0, 1]$, let $(1-\alpha)$ be the degree of 'confidence' ('reliability') we assign to the source of information. If the source is not fully relibale, the bba it generates is 'discounted' into a new less informative bba denoted m^α [12]:

$$m^\alpha(A) = (1 - \alpha).m(A) \text{ for } A \subset \Theta$$

$$m^\alpha(\Theta) = \alpha + (1 - \alpha).m(\Theta)$$

2.4 Decision making

The TBM considers that holding beliefs and making decision are distinct processes. Hence, it proposes a two level model:

- The credal level where beliefs are entertained and represented by belief functions.
- The pignistic level where beliefs are used to make decisions and represented by probability functions called the pignistic probabilities.

When a decision must be made, beliefs held at the credal level are transformed into a probability measure denoted $BetP$ [15].

The function building this probability is called the pignistic transformation and is defined as:

$$BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)}, \ \forall A \subseteq \Theta$$

3 Attribute selection measures in the context of belief decision trees

A belief decision tree is a decision tree in an uncertain environment where the uncertainty is represented by the TBM. The structure of the training set allowing to induce a belief decision tree is different from the traditional one. We assume that the values of the attributes of each training object are known with certainty, whereas its corresponding class is uncertain. The uncertainty on the classes of a training object is represented by a basic belief assignment defined on the set of possible classes. The major parameter ensuring the building of a decision tree is the attribute selection measure allowing to determine the attribute to assign to the decision node of the induced belief decision tree at each step. Within our framework, we propose two attribute selection measures.

1. The first one is an extension of the classical approach developed by Quinlan and based on the gain ratio criterion [9]. It is called the averaging approach.

2. The second one represents ideas behind the TBM itself and based on distance criterion. It is called the conjunctive approach.

For all the following sections, we will use the following notations:

- T: a given training set composed by p objects $I_j, j = 1, \ldots p$,
- S: a set of objects belonging to the training set T,
- A: an attribute,
- $\Theta = \{C_1, C_2, ..., C_n\}$: the frame of discernment made of the n possible classes related to the classification problem.
- $m^\Theta\{I_j\}(C)$: the bbm given to the hypothesis that the actual class of object I_j belongs to $C \subseteq \Theta$.

3.1 The averaging approach

As mentioned, under this approach the attribute selection measure is based on the extension of the gain ratio criterion to the uncertain context. It will be based on the entropy computed from the average pignistic probability taken into account the pignistic probabilities of each object in the node. We propose the following steps to choose the appropriate attribute:

1. Compute the pignistic probability, denoted $BetP^\Theta\{I_j\}$, of each training object I_j by applying the pignistic transformation to $m^\Theta\{I_j\}$.

2. Compute the average pignistic probability function $BetP^\Theta\{S\}$ taken over the set of objects S. For each $C_i \in \Theta$,
$$BetP^\Theta\{S\}(C_i) = \frac{1}{|S|} \sum_{I_j \in S} BetP^\Theta\{I_j\}(C_i)$$

3. Compute the entropy $Info(S)$ of the average pignistic probabilities in the set S. The entropy of a probability function P^Θ is given by:
$$Entr(P^\Theta) = -\sum_{\theta \in \Theta} P^\Theta(\theta) \log_2(P^\Theta(\theta))$$
In the present case we define:
$$Info(S) = Entr(BetP^\Theta\{S\})$$

4. For each value v of a given attribute A, define the subset S_v^A including objects having v as a value for the attribute A. Then, compute the average pignistic probability for objects in subset S_v^A. Let the result be denoted $BetP^\Theta\{S_v^A\}$.

5. Compute the entropy $Info(S_v^A)$ with the cases for which the attribute value is v:
$$Info(S_v^A) = Entr(BetP^\Theta\{S_v^A\})$$

6. Compute $Info_A(S)$, as in Quinlan:
$$Info_A(S) = \sum_{v \in D(A)} \frac{|S_v^A|}{|S|} Info(S_v^A)$$
where $D(A)$ is the domain of the possible values of the attribute A.

7. Compute the information gain provided by the attribute A in the set of objects S such that:

$$Gain(S, A) = Info(S) - Info_A(S)$$

8. Using the *Split Info* [9], compute the gain ratio relative to the attribute A:

$$Gain\ Ratio(S, A) = \frac{Gain(S, A)}{Split\ Info(S, A)}$$

where

$$Split\ Info(S, A) = -\sum_{v \in D(A)} \frac{|S_v^A|}{|S|} log_2 \frac{|S_v^A|}{|S|}$$

9. Repeat the same process for every attribute A belonging to the set of attributes that can be selected. Next, choose the one that maximizes the gain ratio.

3.2 The conjunctive approach

The conjunctive approach is based on an intra-group distance quantifying for each attribute value how strongly objects are close from each others. The different steps upon this attribute selection measure ensuring the building of a belief decision tree are the following ones:

1. For each training object, compute:

$$\kappa\{I_j\}(C) = -\ln q^{\Theta}\{I_j\}(C)\ \forall C \subseteq \Theta$$

from the bba $m^{\Theta}\{I_j\}$.

2. For each attribute value v of an attribute A, compute the joint $\kappa\{S_v^A\}$ defined on Θ, the set of possible classes by:

$$\kappa\{S_v^A\} = \sum_{I_j \in S_v^A} \kappa\{I_j\}$$

3. For each attribute value, the intra-group distance $SumD(S_v^A)$ is defined by:

$$SumD(S_v^A) = \frac{1}{|S_v^A|} \sum_{I_j \in S_v^A} \sum_{X \subseteq \Theta}$$

$$(\kappa\{I_i\}(X) - \frac{1}{|S_v^A|}\kappa\{S_v^A\}(X))^2$$

4. Compute $SumD_A(S)$ representing the weighted sum of the different $SumD(S_v^A)$ relative to each value v of the attribute A:

$$SumD_A(S) = \sum_{v \in D(A)} \frac{|S_v^A|}{|S|} SumD(S_v^A)$$

5. By analogy to our averaging approach, we may also compute $Diff(S, A)$ defined as the difference between $SumD(S)$ and $SumD_A(S)$:

$$Diff(S, A) = SumD(S) - SumD_A(S)$$

where

$$SumD(S) = \frac{1}{|S|} \sum_{I_j \in S} \sum_{X \subseteq \Theta}$$

$$(\kappa\{I_i\}(X) - \frac{1}{|S|}\kappa\{S\}(X))^2$$

6. Using the *Split Info*, compute the diff ratio relative to the attribute A:

$$Diff\ Ratio(S, A) = \frac{Diff(S, A)}{Split\ Info(S, A)}$$

7. For every attribute repeat the same process, and choose the one that maximizes the diff ratio.

Adaptations must be introduced if some $m^{\Theta}\{I_i\}$ are dogmatic (i.e., $m^{\Theta}\{I_j\}(\Theta) = 0$). The simplest consists in discounting every dogmatic bba.

3.3 Structure of leaves

Due to uncertainty in classes of training objects, each leaf in the induced tree will be characterized by a bba. According to the used attribute selection measure:

- Using the averaging approach, the leaf's bba is equal to the average of the bba's of the objects belonging to this leaf.
- Using the conjunctive approach, the leaf's bba is the result of the conjunctive combination of the bba's of objects belonging to this leaf.

4 Description of the belief decision tree approach

4.1 An algorithm to build belief decision trees

Building a decision tree in this context of uncertainty will follow the same steps presented by Quinlan in his C4.5 algorithm [9]. Our algorithm which uses a Top Down Induction of Decision Trees. Furthermore, our algorithm is generic since it offers two possibilities for selecting the attributes by using either the averaging approach or the conjunctive one. The different steps of our algorithm for building a belief decision tree are described as follows:

1. Create the root node of the belief decision tree with all the training objects of T.
2. Choose which approach will be used to select the 'best' attribute: either the averaging approach or the conjunctive one.
3. Verify if this node satisfies any stopping criteria:
   - The node is empty or contains only one object.
   - There is no further attribute to test. In other words, all the attributes are split.
   - If the value of the attribute selection measure is less or equal than zero i.e., another partition does not provide a better separation between objects.
4. If one of these conditions is satisfied,
   - then declare the node as a leaf node and compute its corresponding bba according to the chosen approach.
     Note that in both approaches when the leaf is empty, its corresponding bba is a vacuous bba[1].
   - else, look for the attribute having the highest value of the attribute selection measure. This attribute will be designed as the root of the decision tree related to the whole training set.
5. Develop a branch for each attribute value chosen as a root. This partition leads to several training subsets.
6. Create a root node relative to each training subset.
7. Repeat the same process for each training subset from the step 3, while verifying the stopping criteria.
8. Stop when all the nodes of the latter level of the tree are leaves.

4.2 Classification

To classify a new object described by an exact value for each one of its attribute, we have to start from the root of the belief decision tree, and follow the path leading to a leaf such that for each level of the tree, we test the specified attribute that allows us to move down the tree branch according to the attribute value of the object to classify. This process is repeated until a leaf is encountered.

As a leaf is characterized by a basic belief assignment on classes, the pignistic transformation is applied to get the pignistic probability on the classes of the object to classify in order to decide its class. For instance, one can choose the class having the largest pignistic probability.

4.3 Improvements of the stopping criteria

Inducing a decision tree, and consequently a belief decision tree, without applying any mechanism of pruning leads in most cases to very large trees with many nodes and leaves.

In this paper we are not interested with post-pruning of decision trees consisting to prune branches once the tree is built. However, our objective is to reduce the size of the tree and to avoid the overfitting[2] by improving the

[1] A vacuous bba is defined such that: $m(\Theta) = 1$ and $m(\theta) = 0$ for $\theta \subset \Theta$.

[2] It occurs when the size of the tree is too large compared to the number of training objects.

stopping rules. Such process is called prepruning. In other words, the objective is to minimize the levels of the induced tree and also to avoid as much as possible leaves with too few objects.

The two approaches for building belief decision trees often produce 'complete' trees. The stopping criteria are such that usually they do not stop the development of a path before using all the attributes. Therefore, we think the stopping criteria should be adapted.

As stated previously one condition to declare a node a leaf is when the value of the attribute selection measure is less or equal than zero (either for the gain ratio or the diff ratio). The gain ratio or the diff ratio depend on the aggregation of objects bba's of those belonging to the node (either using the averaging rule or the conjunctive rule of combination according the used approach). We suggest to discount the induced bba.

As the objective is to reduce the number of leaves in a tree and consequently not to have leaves with too few objects, we suggest that this discounting factor depends on the number of objects in the node. On the other hand, it should not 'badly' affect the qualtiy criteria (PCC, kappa) used to judge the quality of the classifier. We define the reliability factor $1-\alpha$ such that:

$$1-\alpha = \frac{N.\ \mathrm{OD}}{N.\ \mathrm{OD} + V}$$

where $N.$ OD is the number of objects in the considered decision node and V is a non negative real. This value of V should be chosen such that the number of leaves diminishes without reducing the quality criteria.

This discounting should be applied on the aggregated bba, i.e., once the combination is done (either through the averaging rule or the conjunctive rule). Note that discounting the aggregated bba increases uncertainty, bringing the pignistic probability function closer to the equi-probability function. The value of the attribute selection measure decreases and this increases the chances to declare the node as a leaf.

5 Simulation

We have applied the BDT approaches[3] in some known data bases. Since, there are several parameters to take into account and consequently several types of simulations, in this paper we will present results of simulation only on a modified breast cancer data base inspired by the breast cancer data base available in http://www.ics.uci.edu/ mlearn/MLRepository.html, but modified in order to satisfy the prerequisites of our methods: symbolic attributes and uncertain classes.

In table 1, we give a brief description of the modified breast cancer basis, hence we present the parameters composing this basis:

- $N.\ O$: number of the whole objects,
- $N.\ Tr\ O$: number of training objects,
- $N.\ Ts\ O$: number of testing objects,
- $N.\ Cl$: number of classes,
- $N.\ Att$: number of attributes,
- $N.\ Val\ Att$: number of values for each attribute.

Table 1: Modified breast cancer data basis parameters.

PARAMETERS	**VALUE**
N. O	690
$N.\ Tr\ O$	621
$N.\ Ts\ O$	69
$N.\ Cl$	2
$N.\ Att$	8
$N.\ Val\ Att$	[2 3 14 8 2 2 2 3]

Several criteria could be used to judge the quality of classification of our induced belief decision trees. In this paper, we use the PCC representing the percentile of the correct classification of the objects belonging to the testing set. It is computed by the cross validation method with a partition in 10 sub samples.

[3] All algorithms have been implemented using Matlab V6.0.

The PCC is computed after selecting for each testing set case the class with the largest pignistic probability.

Data bases with uncertainty in the class do not seem to be available. So we took a classical data base, and 'destroyed' the class variable.

We randomly generate a subset θ of Θ such that the actual class of the object under consideration belongs to θ, and every other class belongs to θ with probability p. We then build a simple support function with θ its focal element and its weight being a random number in [0,1]. We build several (2 here) such simple support functions and combine them conjunctively. The resulting bba is the bba describing our belief about the value of the actual class to which the object belongs.

Table 2 and table 3 present the mean PCC resulting from the application of our two approaches without and with pre-pruning ($V = 20$). We present PCC_a, PCC_c, $N.\ l\_a$ and $N.\ l\_c$, the PCC and the number of leaves induced from the averaging and conjunctive approaches, respectively. The p parameter (probability) used in the class 'destruction' varies from 0 to .9.

Table 2: PCC and number of leaves: without pre-pruning

p	PCC_a	PCC_c	$N.\ l\_a$	$N.\ l\_c$
0	84.38	83.25	337.2	336.2
0.1	83.11	82.09	316.2	316.1
0.2	84.84	83.93	295.7	295.7
0.3	86.93	86.25	269	268.9
0.4	83.9	83.31	246	245.9
0.5	87.04	85.9	222.4	222.4
0.6	83.01	82.47	225	225
0.7	85.64	86.94	146.1	146.1
0.8	81.27	81.46	109.8	109.8
0.9	80.43	81.15	63.4	63.4
Mean	*84.06*	*83.67*	*220.24*	*220.11*

As we note, both approaches (averaging and conjunctive) present high PCC's for all the values of p (the mean is almost 84%). However without pre-pruning, the induced belief decision are characterized by a great number of leaves (a mean of 220 leaves for both approaches).

Table 3: PCC and number of leaves: with pre-pruning ($V = 20$)

p	PCC_a	PCC_c	$N.\ l\_a$	**N. l_c**
0	87.73	83.38	67.2	180.4
0.1	86.66	82.22	35.8	198.2
0.2	87.68	85.67	37.9	199.4
0.3	87.38	85.42	37.3	161.9
0.4	86.57	84.91	37.2	147.4
0.5	87.44	87.93	31.7	160.7
0.6	87.33	85.2	38.1	141.2
0.7	86.46	88.19	23.6	103.2
0.8	87.98	87.57	13.3	70.7
0.9	86.41	84.09	36.2	54.1
Mean	*87.16*	*85.45*	*35.83*	*141.71*

Applying our pre-pruning mechanism with a value of V equals to 20, we keep a high value of PCC's (a mean of respectively 87% and 85%). On the other hand, we note that the number of leaves in both approaches have considerably decreased (a mean of 35 leaves for the averaging approach and 141 leaves for the conjunctive approach). Furthermore, the averaging approach seems more sensitive to this mechanism of pre-pruning than the conjunctive one in reducing the size of the induced trees. Other values of V have been considered. Smaller values were not that efficient in reducing the tree complexity, larger values reduce the PCC. The results were not sensible to local variations of V.

6 Conclusion

We have presented our approach of belief decision trees dealing with uncertainty about the actual class of those cases in the training set. Two attribute selection measures have been proposed and consequently two approaches are considered. We have developed a pre-pruning technique in order to reduce the complexity of the induced trees and to avoid the overfitting problem. Simulations have shown the efficiency of this method. In practical applications, the estimation of V can

be obtained by opportunistically optimizing the *PCC* on the testing set computed with the cross validation method.

References

[1] J. Abellan, S. Moral (2001). Building classification trees using the total uncertainty criterion. In *Proceedings of the International Symposium on Imprecise Probabilities and their Applications*, Newyork, USA, Juin 2001.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone (1984). Classification and regression trees. *Monterey, CA : Wadsworth & Brooks*, 1984.

[3] T. Denoeux, M. S. Bjanger (2000). Induction of decision trees from partially classified data using belief functions, In *proceedings of SMC Conference*, pages 2923-2928, Nashville, USA, October 2000.

[4] Z. Elouedi, K. Mellouli, P. Smets (2000). Decision trees using the belief function theory. In *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'2000*, volume 1, pages 141-148, Madrid, Sapin, July 2000.

[5] Z. Elouedi, K. Mellouli, P. Smets (2001). Belief Decision Trees: Theoretical Foundations. In *International Journal of Approxiamte Journal IJAR*, volume 28(2-3), pages 91-124, 2001.

[6] C. Marsala (1998). Apprentissage inductif en présence de données imprécises: Construction et utilisation d'arbres de décision flous. *Thèse de doctorat de l'Université Paris6, LIP6*, 1998.

[7] J. R. Quinlan (1986). Induction of decision trees, *Machine Learning 1*, pages 81-106, 1986.

[8] J. R. Quinlan (1987). Decision trees as probabilistic classifiers. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 31-37, June 1987.

[9] J. R. Quinlan (1993). C4.5 : Programs for machine learning. *Morgan Kaufmann*, San Mateo, Ca, 1993.

[10] G. Shafer (1976). A mathematical theory of evidence. *Princeton University Press, Princeton NJ*, 1976.

[11] P. Smets (1990). The combination of evidence in the Transferable Belief Model. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 12, pages 321-344, 1990.

[12] P. Smets (1992). The transferable belief model for expert judgments and reliability problems. In *Analysis and Management of uncertainty: Theory and Applications B.M. Ayyub, M.M. Gupta and L.N. Kanal (editors), Elsevier Science Publishers B.V.*, 1992.

[13] P. Smets, R. Kennes (1994). The transferable belief model. *Artificial Intelligence*, volume 66, pages 191-234, 1994.

[14] P. Smets (1998). The Transferable Belief Model for Quantified Belief Representation. *D.M. Gabbay and Ph. Smets (eds.), Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 1, pages 267-301, Kluwer, Doordrecht, 1998.

[15] P. Smets (1998). The Application of the Transferable Belief Model to Diagnostic Problems. *Int. J. Intelligent Systems*, volume 13, pages 127-158, 1998.