World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:3, 2010

Learning and Evaluating Possibilistic Decision Trees using Information Affinity

Ilyes Jenhani, Salem Benferhat and Zied Elouedi

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

Abstract—This paper investigates the issue of building decision trees from data with imprecise class values where imprecision is encoded in the form of possibility distributions. The Information Affinity similarity measure is introduced into the well-known gain ratio criterion in order to assess the homogeneity of a set of possibility distributions representing instances's classes belonging to a given training partition. For the experimental study, we proposed an information affinity based performance criterion which we have used in order to show the performance of the approach on well-known benchmarks.

Keywords—Data mining from uncertain data, Decision Trees, Possibility Theory.

I. Introduction

Machine learning and data mining researches have rapidly emerged in the last decade. Especially, classification is considered as one of the most successful branches of Artificial Intelligence and it is playing a more and more important role in real-world applications.

Classification tasks are ensured by several approaches such as: discriminant analysis, artificial neural networks, k-nearest neighbors, Bayesian networks, decision trees. etc. The latter, namely, decision trees, is considered as one of the most popular classification techniques. They are able to represent knowledge in a flexible and easy form which justifies their use in decision support systems, intrusion detection systems, medical diagnosis, etc.

For many real-world problems and particulary for classification problems, imprecision is often inherent in modeling these applications and should be considered when building classifiers. For example, for some instances, an expert or a sensor may be unable to give the exact class value: an expert in ballistics in the scientific police who is unable to provide the exact type of a gun used in a crime, a mechanic who is unable to provide the exact fault of an engine, a doctor who cannot specify the exact disease of a patient, etc.

An interesting real example emphasizing the problem of having imprecise class labels is the one given in [5]. It consists in detecting certain transient phenomena (e.g. k-complexes and delta waves) in electroencephalogram (EEG) data. Such phenomena are usually difficult to detect, hence doctors are not always able to recognize them with full certainty. Consequently, it may be more easy for doctors to assess the possibility that certain phenomena are present in the data.

Hence, in these different examples, the expert can provide imprecise or uncertain classifications expressed in the form of a ranking on the possible classes. Obviously, rejecting these pieces of information in a learning process is not a good practice. A suitable theory dealing with such situations is possibility theory which is a non-classical theory of uncertainty proposed by [18] and developed by [6].

Let us note that some decision tree approaches have already dealt with the problem of uncertainty and imprecision by using other uncertainty formalisms. We can mention, fuzzy decision trees induced from instances with vaguely defined linguistic attributes and classes [11], [12], [17] and belief decision trees induced from data with partially defined classes presented in the form of basic belief assignments (belief decision trees) [4], [7].

In this paper, we propose a new decision tree approach that allows the induction of decision trees from imprecisely labeled instances, i.e., whose class labels are given in the form of possibility distributions. We introduced the concept of similarity into the attribute selection step of the proposed approach.

It is important to mention that existing possibilistic decision trees do not deal with uncertainty in classes, except, the work we have proposed in [10] using the concept of non-specificity in building possibilistic decision trees. A work proposed by Borgelt and al. [3] deals with crisp (standard) training sets: the authors encode the frequency distributions as possibility distributions (an interpretation which is based on the context model of possibility theory [3]) in order to define a possibilistic attribute selection measure. The possibilistic decision tree approach proposed by Hüllermeier [8] uses a possibilistic branching within the lazy decision tree technique. Again, this work does not deal with any uncertainty in the classes of the training objects. A work by Ben Amor et al. [1] have dealt with the classification of objects having possibilistic uncertain attribute values within the decision tree technique.

Ilyes Jenhani, LARODEC, Institut Supérieur de Gestion de Tunis, email: ilyes.j@lycos.com
Salem Benferhat, CRIL, Université d'Artois, Lens, France, email: benferhat@cril.univ-artois.fr
Zied Elouedi, LARODEC, Institut Supérieur de Gestion de Tunis, email: zied.elouedi@gmx.fr

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

This paper is organized as follows. Section 2 gives necessary background on possibility theory. Section 3 describes some basics of decision trees. Then, in Section 4, we present our approach, so-called Aff-PDT. Section 5 presents and analyzes experimental results carried out on modified versions of commonly used data sets from the U.C.I. repository [14]. Finally, Section 6 concludes the paper.

II. Possibility Theory

Possibility theory represents a non-classical uncertainty theory, first introduced by Zadeh [18] and then developed by several authors (e.g., Dubois and Prade [6]). In this section, we will give a brief recalling on possibility theory.

Possibility distribution

Given a universe of discourse $\Omega = \{\omega_1,\ \omega_2,\ ...,\ \omega_n\}$, a fundamental concept of possibility theory is the *possibility distribution* denoted by π. π corresponds to a function which associates to each element ω_i from the universe of discourse Ω a value from a bounded and linearly ordered valuation set (L,<). This value is called a *possibility degree*: it encodes our knowledge on the real world. Note that, in possibility theory, the scale can be numerical (e.g. L=[0,1]): in this case we have numerical possibility degrees from the interval [0,1] and hence we are dealing with the quantitative setting of the theory.

By convention, $\pi(\omega_i)$ = 1 means that it is fully possible that ω_i is the real world, $\pi(\omega_i)$ = 0 means that ω_i cannot be the real world (is impossible). Flexibility is modeled by allowing to give a possibility degree from]0,1[. In possibility theory, extreme cases of knowledge are given by:

-*Complete knowledge*: $\exists \omega_i,\ \pi(\omega_i) = 1\ and\ \forall\ \omega_j \neq \omega_i,\ \pi(\omega_j) = 0$.

- *Total ignorance*: $\forall\ \omega_i \in \Omega,\ \pi(\omega_i) = 1$ (all values in Ω are possible).

Possibility and Necessity measures

From a possibility distribution, two dual measures can be derived: *Possibility* and *Necessity* measures. Given a possibility distribution π on the universe of discourse Ω, the corresponding possibility and necessity measures of any event $A \subseteq 2^{\Omega}$ are, respectively, determined by the formulas: $\Pi(A) = \max_{\omega \in A}\ \pi(\omega)$ and $N(A) = \min_{\omega \notin A}\ (1 - \pi(\omega)) = 1 - \Pi(\overline{A})$.

$\Pi(A)$ evaluates at which level A is *consistent* with our knowledge represented by π while $N(A)$ evaluates at which level A is *certainly* implied by our knowledge represented by π.

Normalization

A possibility distribution π is said to be *normalized* if there exists at least one state $\omega_i \in \Omega$ which is totally possible. In the case of sub-normalized π,

$$Inc(\pi) = 1 - \max_{\omega \in \Omega}\{\pi(\omega)\} \tag{1}$$

is called the *inconsistency degree* of π. It is clear that, for normalized π, $\max_{\omega \in \Omega}\{\pi(\omega)\} = 1$, hence Inc($\pi$)=0. The measure Inc is very useful in assessing the degree of conflict between two distributions π_1 and π_2 which is given by $Inc(\pi_1 \wedge \pi_2)$. We take the $\wedge$ as the minimum operator. Obviously, when $\pi_1 \wedge \pi_2$ gives a sub-normalized possibility distribution, it indicates that there is a conflict between π_1 and π_2 $(Inc(\pi_1 \wedge \pi_2) \in]0,1])$.

Information Affinity: a possibilistic similarity measure

Comparing pieces of uncertain information given by several sources has attracted a lot of attention for a long time. This could be ensured by the use of similarity indexes. After a deep study of existing possibilistic similarity measures, we have proposed in a recent work [9], a new similarity index satisfying interesting properties (non-negativity, upper bound and non-degeneracy, lower bound, symmetry, inclusion, permutation).

The information affinity index, denoted by $InfoAff$ takes into account a classical informative distance, namely, the Manhattan distance along with the well known inconsistency measure. $InfoAff$ is applicable to any pair of normalized possibility distributions.

Definition 1: Let π_1 and π_2 be two possibility distributions on the same universe of discourse Ω. We define a measure InfoAff(π_1, π_2) as follows:

$$InfoAff(\pi_1, \pi_2) = 1 - \frac{d(\pi_1, \pi_2) + Inc(\pi_1 \wedge \pi_2)}{2} \tag{2}$$

where $d(\pi_1, \pi_2) = \frac{1}{n}\sum_{i=1}^{n} |\pi_1(\omega_i) - \pi_2(\omega_i)|$ represents the Manhattan distance between π_1 and π_2 and $Inc(\pi_1 \wedge \pi_2)$ tells us about the degree of conflict between the two distributions (see Equation (1)).

Two possibility distributions π_1 and π_2 are said to have a strong affinity (resp. weak affinity) if $InfoAff(\pi_1, \pi_2) = 1$ (resp. $InfoAff(\pi_1, \pi_2) = 0$).

For sake of simplicity, in the rest of the paper, a possibility distribution π on a finite set $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ will be denoted by $\pi[\pi(\omega_1), \pi(\omega_2), ..., \pi(\omega_n)]$.

III. Decision trees

Decision trees, also called classification trees, are graphical models with a tree-like structure: they are composed of three basic elements: decision nodes corresponding to attributes, edges or branches which correspond to the different possible attribute values. The third component consists of leaves including objects that typically belong to the same class or that are very similar.

Several algorithms for building decision trees have been developed. The most popular and applied ones are: **ID3** [15] and its successor **C4.5** "the state-of-the-art" algorithm developed by Quinlan [16]. These algorithms have many components to be defined:

a) Attribute selection measure generally based on information theory, serves as a criterion in choosing among a list of

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

candidate attributes at each decision node, the attribute that generates partitions where objects are distributed less randomly, with the aim of constructing the smallest tree among those consistent with the data. The well-known measure used in the **C4.5** algorithm of Quinlan [16] is the gain ratio.

Given an attribute A_k, the information gain relative to A_k is defined as follows:

$$Gain(T, A_k) \ = \ E(T) \ - \ E_{A_k}(T) \tag{3}$$

$where$

$$E(T) = -\sum_{i=1}^{n} \frac{n(C_i, T)}{|T|} \ log_2 \ \frac{n(C_i, T)}{|T|} \tag{4}$$

and

$$E_{A_k}(T) \ = \ \sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} \ E(T_v^{A_k}) \tag{5}$$

n(C_i,*T*) denotes the number of objects in the training set *T* belonging to the class C_i, $D(A_k)$ denotes the finite domain of the attribute A_k and $|T_v^{A_k}|$ denotes the cardinality of the set of objects for which the attribute A_k has the value *v*. Note that $\frac{n(C_i,T)}{|T|}$ corresponds to the probability of the class C_i in T. Thus, $E(T)$ corresponds to the *Shannon entropy* of the set T. The gain ratio is given by:

$$Gr(T, A_k) \ = \ \frac{Gain(T, A_k)}{SplitInfo(T, A_k)} \tag{6}$$

where $SplitInfo$(*T*,A_k) represents the potential information generated by dividing *T* into n subsets. It is given by:

$$SplitInfo(T, A_k) = -\sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} \ log_2 \ \frac{|T_v^{A_k}|}{|T|} \tag{7}$$

b) Partitioning strategy consisting in partitioning the training set according to all possible attribute values (for symbolic attributes) which leads to the generation of one partition for each possible value of the selected attribute. For continuous attributes, a discretization step is needed.

c) Stopping criteria stopping the partitioning process. Generally, we stop the partitioning if all the remaining objects belong to only one class, then the node is declared as a leaf labeled with this class value. We, also, stop growing the tree if there is no further attribute to test. In this case, we take the majority class as the leaf's label.

IV. Affinity based possibilistic decision trees

An affinity based possibilistic decision tree (Aff-PDT) has the same representation of a standard decision tree, i.e., it is composed of *decision nodes* for testing attributes, *branches* specifying attribute values and *leaves* dealing with classes of the training set.

A. Classification from imperfect data

As models of the real world, databases, or more specifically, training sets are often permeated with forms of imperfections, including imprecision and uncertainty. The topic of imperfect databases is gaining more and more attention the last years [13] since commercial database management systems are not able to deal with such kind of information.

Examples of imperfect class values include the exact type of an attack in an intrusion detection system, the exact cancer class of a patient in cancer diagnosis applications, the exact location or type of a detected aerial engine in military applications, etc. These imperfections might result from using unreliable information sources, such as faulty reading instruments, or input forms that have been filled out incorrectly (intentionally or inadvertently).

In order to deal with such kind of imperfection, in this work, we used a convenient mathematical model, namely, possibility theory [6], [18]. More formally, a possibility degree will be assigned to each possible class value indicating the possibility that the instance belongs to a given class [5]. These possibility degrees can be obtained from direct expert's elicitation, i.e., each expert is asked to quantify by a real number between 0 and 1 the possibility that a training instance belongs to each one of the different classes of the problem.

B. Building procedure

In the possibilistic setting, instances classes in the training set will be represented by possibility distributions over the different classes of the problem instead of exact classes. Hence, one must find a way to assess homogeneity of a given training partition. The idea consists in measuring the entropy of each partition weighted by the mean similarity degree of the possibility distributions in the corresponding partition. Let us define the basic components for the Aff-PDT approach:

a) Meta-classes and wrapper possibility distributions
Given a training set T (the initial partition) containing n instances and given the set of attributes, let us denote by π_i the possibility distribution labeling the class of the instance i in T.

In standard decision trees, homogeneity of a partition is determined by the entropy of that partition. However, in our context, π_i's are most of the time very different, so it has no sense to directly compute their frequencies in order to determine the entropy of T (the entropy will be equal to 1). Moreover, one cannot simply view each π_i as a new class. First, because the number of classes will be exponential. Second, there are similar distributions that should be considered as globally expressing same or similar pieces of information. For instance, we cannot simply consider the distributions [1, 0.2] and [1, 0.21] as two different exclusive classes, but we will consider them as similar.

Hence, we need a finite set of Meta-Classes $MC_{j=1..m}$. Each MC_j corresponds to a meta-class which gathers together

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

all possibility distributions similar to a predefined wrapper possibility distribution (say WD_j). More precisely, wrapper possibility distributions are binary possibility distributions (i.e., $\forall \omega \in \Omega, \pi(\omega) \in \{0, 1\}$) representing special cases of complete knowledge, partial ignorance and total ignorance representing the set of reference distributions.

After specifying the set of Meta-Classes MC, we will assign to each possibility distribution π_i labeling an instance i a meta-class MC_j such that: $MC_j = \arg\max_{j=1}^{m}\{InfoAff(\pi_i, WD_j)\}$ where m is the total number of meta-classes and $InfoAff$ corresponds to the information Affinity index (Equation (2)). Note that, as in standard decision trees, ties are broken arbitrarily.

b) Affinity-Gain ratio

After mapping the different π_i's to their corresponding MC_j's , it becomes possible to assess the discriminative power of each attribute in partitioning a set into homogeneous subsets by extending the well-known gain ratio criterion [16]. First, we define the Affinity-Entropy Gain ($AGain$) of an attribute A_k by:

$$AGain(T, A_k) \;=\; AE(T) \;-\; AE_{A_k}(T) \tag{8}$$

$where$

$$AE(T) = -\sum_{j=1}^{m}(AvgAff(MC_j)) * (\frac{|MC_j|}{|T|}\ log_2 \frac{|MC_j|}{|T|}) \tag{9}$$

and

$$AE_{A_k}(T) \;=\; \sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|}\ AE(T_v^{A_k}) \tag{10}$$

where $|MC_j|$ in Equation (9) denotes the number of objects in the training set *T* belonging to the meta-class MC_j. Obviously, to compensate for the information loss resulting from grouping resemblant π_i's into their corresponding MC_j's, we have introduced the $AvgAff(MC_j)$ factor which corresponds to the average similarity between the original possibility distributions $\pi_{p=1..n}$ assigned to MC_j:

$$AvgAff(MC_j) = \frac{\sum_{p=1}^{n-1}\sum_{q=p+1}^{n} InfoAff(\pi_p, \pi_q)}{\frac{n*(n-1)}{2}} \tag{11}$$

Proposition 1: When dealing with crisp training sets, i.e., with precise classes ($MC_j \equiv C_j$), we will always have $AvgAff(MC_j) = 1$ and $|MC_j|$ will correspond to number of instances labeled by the same class C_j, thus we recover the standard C4.5 approach.

Then, the Affinity-gain ratio is expressed in the same way as the classical gain ratio using $SplitInfo$ (Equation (7)):

$$AGr(T, A_k) \;=\; \frac{AGain(T, A_k)}{SplitInfo(T, A_k)} \tag{12}$$

Obviously, the attribute maximizing AGr will be assigned to the decision node at hand.

c) Partitioning strategy

Since we only deal with nominal attributes, the partitioning strategy will be the same as with standard decision trees.

d) Stopping criteria

We will stop growing the tree if:

1. There is no further attribute to test.

2.$AGain \leq 0$, i.e., no information is gained.

3. $|T_p|$=0, i.e., the generated partition does not contain any instance.

e) Structure of leaves

Leaves of our induced Aff-PDT trees will be labeled by possibility distributions on the different classes rather than crisp classes.

In fact, when the above stopping criterion 1 or 2 is satisfied for a training partition T_p containing n possibility distributions, we will declare a leaf labeled by the representative possibility distribution of that set (π_{Rep}), that is, the possibility distribution which corresponds to the closest distribution to all the remaining distributions in the set T_p:

$\pi_{Rep}(T_p) = \arg\max_{i=1}^{n}\{\frac{\sum_{j \neq i} InfoAff(\pi_i, \pi_j)}{(n-1)}\}$

Hence, when considering the special case of a leaf with only certain possibility distributions, if we take the fully possible class of ($\pi_{Rep}(T_p)$) as a final decision, we join the solution of majority class adopted by standard decision trees.

Finally, when stopping criterion 3 is satisfied, we declare an empty leaf labeled by a randomly chosen wrapper possibility distribution from WD.

It is clear that fusion cannot be applied to combine the possibility distributions belonging to a given leaf. In fact, in the decision tree context, in each leaf, we have possibility distributions of distinct training instances reaching that leaf. These instances have some common attribute values (those values labeling edges of the path leading to that leaf) and the remaining attributes may have different values. So, it is clear that we cannot merge possibility distributions which are not dealing with the same "object": a necessary condition for information fusion problems.

Example 1: Let us use a modified version of the golf data set [14] to illustrate the notion of wrapper distributions and show the computation of the affinity-gain ratio of a given attribute. Let T be the training set composed of fourteen instances $i_{=1..14}$. A possibility distribution was given for each possible class of each instance of T.

The set of wrapper distributions relative to this example is $WD = \{[1, 0], [0, 1], [1, 1]\}$. Consequently, $MC = \{MC_1, MC_2, MC_3\}$ such that $MC_1 = \{i_4, i_9, i_{10}, i_{11}, i_{12}, i_{13}\}$, $MC_2 = \{i_1, i_2, i_6, i_8, i_{14}\}$ and $MC_3 = \{i_3, i_5, i_7, \}$.

The Affinity-Entropy of the set T is computed (using Equation (9)) as follows : $AE(T) = -0.956 * (\frac{6}{14} * log_2 \frac{6}{14}) - 0.95 * (\frac{5}{14} * log_2 \frac{5}{14}) - 0.966 * (\frac{3}{14} * log_2 \frac{3}{14}) = 1.464$.

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

TABLE I
IMPRECISELY LABELED TRAINING SET

	Outlook	**Temp**	**Humidity**	**Wind**	C_1	C_2
i_1	sunny	hot	high	weak	0.2	1
i_2	sunny	hot	high	strong	0.4	1
i_3	overcast	hot	high	weak	1	0.7
i_4	rainy	mild	high	weak	1	0
i_5	rainy	cool	normal	weak	1	0.8
i_6	rainy	cool	normal	strong	0.4	1
i_7	overcast	cool	normal	strong	1	0.9
i_8	sunny	mild	high	weak	0.3	1
i_9	sunny	cool	normal	weak	1	0.3
i_{10}	rainy	mild	normal	weak	1	0
i_{11}	sunny	mild	normal	strong	1	0.2
i_{12}	overcast	mild	high	strong	1	0
i_{13}	overcast	hot	normal	weak	1	0.3
i_{14}	rainy	mild	high	strong	0	1

Let us show a detailed computation of the affinity-gain ratio of the "Wind" attribute. Let us compute $AE(T^{Wind}_{strong})$ and $AE(T^{Wind}_{weak})$ using Equation (9):

$AE(T^{Wind}_{strong}) = -(0.95) * (\frac{2}{6} * log_2 \frac{2}{6}) - (0.93) * (\frac{3}{6} * log_2 \frac{3}{6}) - (1) * (\frac{1}{6} * log_2 \frac{1}{6}) = 1.397.$

$AE(T^{Wind}_{weak}) = -(0.95) * (\frac{4}{8} * log_2 \frac{4}{8}) - (0.97) * (\frac{2}{8} * log_2 \frac{2}{8}) - (0.97) * (\frac{2}{8} * log_2 \frac{2}{8}) = 1.445.$

⇒ Using Equation (10), we obtain:
$AE_{Wind}(T) = \frac{6}{14} * 1.397 + \frac{8}{14} * 1.445 = 1.424.$

⇒ Using Equation (8): $AGain(T, Wind) = 1.464 - 1.424 = 0.04.$

⇒ Using Equation (7): $SplitInfo(T, Wind) = -\frac{6}{14} * log_2 \frac{6}{14} - \frac{8}{14} * log_2 \frac{8}{14} = 0.985.$

⇒ Finally, using Equation (12): $AGr(T, Wind) = \frac{0.04}{0.985} = 0.0406.$

Similarly, we should compute $AGr(T, Outlook)$, $AGr(T, Temp)$ and $AGr(T, Humidity)$, then choose the attribute that maximizes AGr which will be assigned to the decision node at hand.

C. Classification procedure

Once the Aff-PDT is constructed, we can classify any new object given values of its attributes. We start with the root of the constructed tree and follow the path corresponding to the observed value of the attribute in the interior node of the tree. This process is continued until a leaf is encountered. As mentioned above, each leaf of our decision tree will be labeled by a possibility distribution over the different class values. Hence, to make a decision about the class of a given object, the decision maker can take the fully possible class label (i.e. the class having a possibility degree equal to 1).

V. EXPERIMENTAL RESULTS

Our experimental studies are divided in two parts. First, we evaluate our Aff-PDT approach. Second, we compare our results with those of the C4.5 algorithm if we ignored uncertainty. Note that we do not intend to compare Aff-PDT with C4.5 since this latter do not deal with uncertainty: the aim of the comparison is to show whether ignoring uncertainty in training data is a good practice or not.

The experimental study is based on several data sets selected from the U.C.I repository of machine learning databases [14]. A brief description of these data sets is given in Table 2. #Data, #attributes, #classes denote respectively the total number of instances, the number of attributes and the number of classes.

TABLE II
DESCRIPTION OF DATABASES

Database	**#Data**	**#attributes**	**#classes**
Wisconsin Breast Cancer	699	8	2
Voting	497	16	2
Solar Flare	1389	10	3
Balance scale	625	4	3
Nursery	12960	8	5

We have modified these data sets by transforming the original crisp classes by possibility distributions over the different classes. We have used levels of uncertainty ($L\%$) when generating these possibilistic training sets: for each training instance from the $L\%$ randomly chosen instances, we have assigned a possibility degree equal to 1 to the original class and a random possibility degree to the remainders in an uniform way. To each one of the remaining $(100 - L)\%$ instances, we have assigned a completely sure possibility distribution corresponding to the original crisp instance's class.

In order to determine the accuracy of the induced trees, we have used two criteria, the first is relative to the percentage of correct classification ($PCC = \frac{number\ of\ well\ classified\ instances}{total\ number\ of\ classified\ instances} \times 100$) and the second corresponds to a similarity based criterion (PCC_Aff) which we have proposed in [9] as a new criterion that is more appropriate to the possibilistic context:

$$PCC\_Aff = \frac{\sum_{j=1}^{n} InfoAff(\pi^{res}, \pi^{j})}{total\_nbr\_classified\_inst} \times 100 \qquad (13)$$

Let us recall that the output of a possibilistic decision tree is given in the form of a possibility distribution (π^{res}). Thus, the standard PCC is computed by choosing for each instance to classify the class having the highest possibility degree (equal to 1). If more than one class is obtained, then one of them is chosen randomly. Finally, this class label is compared with the true class label. This is not a good practice. In fact, ignoring the rest of the degrees implies ignoring a part of the information given by the resulting possibility distribution (π^{res}).

Hence, we were inspired by the work in [2] to define the PCC_Aff criterion which takes into account the mean similarity relative to all the classified testing instances: the average of the similarities between the resulting possibility distribution (π^{res}_j) and the real (completely sure) possibility distribution

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf

(π_j) of each classified instance j. When PCC_Aff is close to 100$\%$, the classifier is good whereas when it falls to 0$\%$, it is considered as a bad classifier.

TABLE III
AFF-PDT (PCC_Aff AND STANDARD DEVIATION)

$L\%$	0%	30%	50%
W.B.cancer	93.37(1.3)	91.2(1.7)	88.71(2.1)
Voting	96.76 (1.7)	95.35(1.9)	94.11(1.9)
Solar Flare	87.26(2.2)	84.90(1.8)	83.77(1.6)
Balance	83.54 (1.5)	73.83 (1.2)	72.88(1.2)
Nursery	98.74 (0.8)	96.96(1.1)	95.95(1.4)

Table 3 reports the different obtained results after varying the training sets' level of uncertainty $L\%$ for each database. PCC_Aff values of the induced Aff-PDT trees are complemented by standard deviations after the use of a 10-fold cross validation testing process.

Note that high values of the PCC_Aff criterion do not only imply that the induced trees are accurate but also imply that the possibility distributions provided by the induced Aff-PDT trees are of high quality and faithful to the original possibility distributions. From Table 3, we can see that PCC_Aff values decrease when $L\%$ increases. This can be explained by the fact that the higher the level of uncertainty ($L\%$), the less informative the training set becomes (consequently, the harder the learning becomes), and therefore the less accurate the predictions are.

Now, let us see what happens when ignoring imprecisely labeled training instances when building decision trees. To respond to this question, we have conducted our experimentations as follows: for each training set and for each uncertainty level $L\%$, we have induced an Aff-PDT tree. On the other hand, a C4.5 tree was induced from the corresponding training set, i.e., the standard training set from which we have discarded the $L\%$ instances to which we have assigned imprecise class labels since the C4.5 algorithm cannot deal with such instances. Then, both approaches are evaluated on the same testing sets: standard testing sets for C4.5 trees have been used and their corresponding testing sets (with completely sure possibility distributions on the original class labels) for Aff-PDT trees: this corresponds to one iteration of the 10-fold cross validation process used for the evaluation of the approach.

Table 4 reports the different obtained results after varying the training sets' level of uncertainty $L\%$ for each database. $MPCC$ denotes the mean PCC (complemented by standard deviation) of the induced decision trees for the 10-fold cross validation process.

Table 4 shows that the Aff-PDT approach gives interesting results when compared with the C4.5 algorithm. Again, we can see that classification accuracies of both approaches decrease when the level of uncertainty increases (for the same explanation provided above for Table 3).

In spite of this decrease in accuracy, we can see that the classification rate of Aff-PDT is always (even slightly) greater than the one of C4.5. Note that the aim of this comparison is not to directly compare the two approaches. In fact, the C4.5 is used only in certain environments: it is trained from reduced training sets (imprecisely labeled instances are omitted) while the Aff-PDT approach deals with both certain and uncertain environments: it is trained from complete training sets (including both precisely and imprecisely labeled instances). Besides, Table 4 confirms Proposition 1. In fact, our approach recovers the C4.5 one when dealing with crisp instances (with precise labels, i.e., $L\%$=0).

TABLE IV
C4.5 AND AFF-PDT (MPCC AND STANDARD DEVIATION)

Database	**Method**	$L = 0\%$	$L = 30\%$	$L = 50\%$
W.B.cancer	C4.5	94.54(1.1)	91.05(2.5)	90.11(3.2)
	Aff-PDT	94.54(1.1)	91.63(2.3)	90.73(2.6)
Voting	C4.5	94.56(3.2)	90.15(3.8)	87.27(4.6)
	Aff-PDT	94.56(3.2)	91.62(3.5)	88.52(4.0)
Solar flare	C4.5	81.96(3.3)	77.03(3.7)	74.37(3.9)
	Aff-PDT	81.96(3.3)	80.57(3.7)	78.48(3.9)
Balance	C4.5	78.48(4.2)	74.78(5.3)	70.38(5.7)
	Aff-PDT	78.48(4.2)	77.06(4.9)	74.82(5.4)
Nursery	C4.5	98.78(0.8)	94.45(1.6)	92.81(2.6)
	Aff-PDT	98.78(0.8)	97.11(1.2)	94.37(2.2)

From the results given in this table, we can conclude that, generally, rejecting training instances, classes of which are imprecisely defined, is not a good practice and reduces the accuracy of the induced classifier. This issue can be avoided and well handled by the use of the proposed Aff-PDT approach which can exploit the information contained in imprecise labels.

VI. CONCLUSION

This paper proposes a generalization of the C4.5 approach to the imprecise setting. The new approach has the advantage of allowing the induction of decision trees from training instances having possibilistic class labels. The proposed Aff-PDT approach blends information affinity with entropy in order to asses the homogeneity of a given training partition. Experiments have shown that rejecting training instances, classes of which are imprecisely defined, is not a good practice and reduces the accuracy of the induced classifier. We plan to add an automatic clustering phase for the specification of the wrapper distributions which could enhance the performance of the approach.

REFERENCES

[1] N. Ben Amor, S. Benferhat, Z. Elouedi: Qualitative classification with possibilistic decision trees, *(IPMU'04)*, Perugia, Italy, 2004.

[2] N. Ben Amor, S. Benferhat, Z. Elouedi: Qualitative classification and evaluation in possibilistic decision trees, *(FUZZ-IEEE'04)*, Hungary, 2004, 653-657.

[3] C. Borgelt, J. Gebhardt, R. Kruse: Concepts for Probabilistic and Possibilistic Induction of Decision Trees on Real World Data. *(EUFIT'96)*, 1996, 1556-1560.

[4] T. Denoeux and M. S. Bjanger: Induction of decision trees from partially classified data. *SMC'00*, Nashville, TN, 2000, 2923-2928.

[5] T. Denoeux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3), 2001, 47-62.

[6] D. Dubois and H. Prade: Possibility theory: An approach to computerized processing of uncertainty, *Plenum Press*, New York, 1988.

[7] Z. Elouedi, K. Mellouli and P. Smets. Belief decision trees: Theoretical foundations. *International Journal of Approximate Reasoning*, 28, 2001, 91-124.

[8] E. Hüllermeier. Possibilistic Induction in decision tree learning. *ECML'02*, Helsinki, Finland, 2002, 173-184.

[9] I. Jenhani, N. Ben Amor, Z. Elouedi, S. Benferhat and K. Mellouli: Information Affinity: a new similarity measure for possibilistic uncertain information, *ECSQARU'07*, Hammamet, Tunisia, 2007, 840-852.

[10] I. Jenhani, N. Ben Amor, Z. Elouedi: Decision Trees as Possibilistic Classifiers, *International Journal of Approximate Reasoning*, 48(3), 2008, 784-807.

[11] C. Z. Janikow. Fuzzy decision trees: issues and methods. *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics* 28(1),1998, 1-14.

[12] C. Marsala: Apprentissage inductif en présence de données imprécises: construction et utilisation d'arbres de décision flous, PhD thesis, University P. et M. Curie, Paris, France, 1998.

[13] A. Motro: Sources of Uncertainty, Imprecision and Inconsistency in Information Systems. *In Uncertainty Management in Information Systems: From Needs to Solutions*, 1996, 9-34.

[14] P. M. Murphy, D. W. Aha: UCI repository of machine learning databases, 1996.

[15] J. R. Quinlan: Induction of decision trees, *Machine Learning*, 1, 1986, 81-106.

[16] J. R. Quinlan: C4.5: Programs for machine learning, Morgan Kaufmann, 1993.

[17] Y. Yuan, M.J. Shaw: Induction of fuzzy decision trees, *Fuzzy Sets and Systems*, 69, 1995, 125-139.

[18] L. A. Zadeh: Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets ans Systems*, 1, 1978, 3-28.

Open Science Index, Computer and Information Engineering Vol:4, No:3, 2010 publications.waset.org/5039.pdf