Maximum of Entropy in Credal Classification*

J. ABELLÁN
Universidad de Granada, Spain

S. MORAL
Universidad de Granada, Spain

Abstract

We present an application of the measure of maximum entropy for credal sets: as a branching criterion for classification trees based on imprecise probabilities. We also justify the use of maximum entropy as a global uncertainty measure for credal sets, and a deduction of this measure, based on the best lower expectation of the logarithmic score, is presented. We have also carried out several experiments in which credal classification trees are built taking a global uncertainty measure as a basis. The results show that there is a lower degree of error when maximum entropy is used as a global uncertainty measure.

Keywords

imprecise probabilities, uncertainty, maximum entropy, imprecision, non-specificity, classification, classification trees, credal sets

1 Introduction

Classification is an important problem in the area of machine learning in which classical probability theory has been extensively used. Basically, we have an incoming set of observations, called the training set, and we want to obtain a set of rules to assign a value of the variable to be classified to any new case. The set used to assess the quality of this set of rules is also called the test set. Classification has notable applications in medicine, recognition of hand-written characters, astronomy, banks, etc. The learned classifier can be represented as a Bayesian network, a neural network, a classification tree, etc. These methods normally use the Theory of Probability to estimate the parameters with a stopping criterion to limit the complexity of the classifier and to avoid overfitting.

*This work has been supported by the Spanish Ministry of Science and Technology, project Elvira II (TIC2001-2973-C05-01).

In some previous papers [4, 5, 6], we have introduced a new procedure to build classification trees based on the use of imprecise probabilities. Classification trees have their origin in Quinlan's ID3 algorithm [18], and a basic reference is the book by Breiman et al. [8]. We also applied decision trees for classification, but as in Zaffalon [25], the imprecise Dirichlet model is used to estimate the probabilities of belonging to the respective classes defined by the variable to be classified. In classical probabilistic approaches, information gain is used to build the tree, but then other procedures must subsequently be used to prune it, since information gain tends to build structures which are too complex. We have shown that if imprecise probabilities are used and the information gain is computed by measuring the total amount of uncertainty of the associated credal sets (a closed and convex set of probability distributions), then the problem of overfitting disappears and results improve.

In Abellán and Moral [1, 2, 3], we studied how to measure the uncertainty of a credal set by generalizing the measures used in the Theory of Evidence, Dempster [10] and Shafer [20]. We considered two main sources of uncertainty: entropy and non-specificity. We proved that the proposed functions verify the most basic properties of these types of measures (Abellán and Moral [2], Dubois and Prade [12], Klir and Wierman [15]).

We previously proved that by using a global uncertainty measure which is the result of adding an entropy measure and a non-specificity measure, classification results are better than those obtained by the C4.5 classification method, based on Quinlan's ID3 algorithm. In this paper, we have carried out some experiments in which the maximum entropy of the probability distributions of a credal set is used to measure its uncertainty, and we show that the results obtained are even better. We consider two methods of building classification trees. In the first method, Abellán and Moral [4], we start with an empty tree and in each step, a node and a variable are selected for branching which give rise to a greater decrease in the final entropy of the variable to be classified. In classical probability, a branching always implies a decrease in the entropy. It is necessary to include an additional criterion so as not to create models which are too complex and therefore overfit the data. With credal sets, a branching will produce a lower entropy but, at the same time, a greater non-specificity. Under these conditions, we follow the same procedure as in probability theory, but measuring the total uncertainty of a branching. The stopping criterion is very simple: when every possible branching produces an increment of the total uncertainty.

Finally, in order to carry out the classification given a set of observations, we use a strong dominance criterion to obtain the value of the variable to be classified and a maximum frequency criterion when we want to classify all the cases.

The extended method quantifies the uncertainty of each individual variable in each node in the same way, but also considers the results of adding two variables at the same time. In this way, we aim to discover relationships involving more than two variables that were not seen when investigating the relationships of a

single variable with the variable to be classified.

In Section 2, we present the necessary previous concepts on uncertainty on credal sets. We place special emphasis on the maximum of entropy as a global uncertainty measure. In Section 3, we introduce the necessary notation and definitions for our procedure of building classification trees. In Section 4, we describe the methods based on imprecise probabilities. In Section 5, we test our procedure with known data sets used in classification by comparing the use of two global uncertainty measures.

2 Total Uncertainty on Credal Sets

Dempster-Shafer's theory is based on the concept of basic probability assignment (bpa), and it defines a special type of credal set [10, 20]. In this theory, Yager [24] distinguishes two types of uncertainty: one is associated with cases where the information is focused on sets with empty intersections; and the other is associated with cases where the information is focused on sets with a greater than one cardinality. We call these *randomness* and *non-specificity*, respectively. In Abellán [6] we justify that a general convex set of probability distributions (a credal set) may contain the same type of uncertainty as a bpa: we consider similar randomness and non-specificity measures.

In Abellán and Moral [2], we define a measure for non-specificity for convex sets that generalizes Dubois and Prade's measure of non-specificity in the theory of evidence [11]. Using the Möbius inverse function for monotonic capacities [9], we can define:

Definition 1 *Let $\mathcal{P}$ be a credal set on a finite set X. We define the following capacity function,*

$$f_{\mathcal{P}}(A) = \inf_{P \in \mathcal{P}} P(A), \ \ \forall A \in \wp(X),$$

where $\wp(X)$ is the power set of X. This function is also known as the minimum lower probability which represents $\mathcal{P}$.

Theorem 1 *(Shafer [20]) For any mapping $f_{\mathcal{P}} : \wp(X) \rightarrow \mathbf{R}$ another mapping $m_{\mathcal{P}} : \wp(X) \rightarrow \mathbf{R}$ can be associated by*

$$m_{\mathcal{P}}(A) = \sum_{B \subseteq A} (-1)^{|A-B|} f_{\mathcal{P}}(B), \forall A \in \wp(X),$$

Where $|A-B|$ is the cardinal of the set $A-B$. This correspondence is one-to-one, since conversely, we can obtain

$$f_{\mathcal{P}}(A) = \sum_{B \subseteq A} m_{\mathcal{P}}(B), \ \forall A \in \wp(X).$$

These functions, $f_{\mathcal{P}}$ and $m_{\mathcal{P}}$, are Möbius inverses.

Definition 2 *Let $\mathcal{P}$ be a credal set on a frame X, $f_{\mathcal{P}}$ its minimum lower probability as in Definition 1 and let $m_{\mathcal{P}}$ be its Möbius inverse. We say that function $m_{\mathcal{P}}$ is an assignment of masses on $\mathcal{P}$. Any $A \in X$ such that $m_{\mathcal{P}}(A) \neq 0$ will be called a focal element of $m_{\mathcal{P}}$.*

We can now define a general function of non-specificity.

Definition 3 *Let $\mathcal{P}$ be a credal set on a frame X. Let $m_{\mathcal{P}}$ be its associated assignment of masses on $\mathcal{P}$. We define the following function of non-specificity on $\mathcal{P}$:*

$$IG(\mathcal{P}) = \sum_{A \subset X} m_{\mathcal{P}}(A) \ln(|A|).$$

In Abellán and Moral [3], we proposed the following measure of randomness for general credal sets:

$$G^*(\mathcal{P}) = Max \left\{ -\sum_{x \in X} p_x \ln p_x \right\},$$

where the maximum is taken over all probability distributions on $\mathcal{P}$, and $\mathcal{P}$ is a general credal set. This measure generalizes the classical Shannon's measure [21] verifying similar properties. It can be used either as one of the components of a measure of total uncertainty, or as a total uncertainty measure, Harmanec and Klir [14]. We have proved that this function is also a good randomness measure for credal sets and possesses all the basic properties required in Dempster-Shafer's theory [3].

We define a measure of total uncertainty as $TU(\mathcal{P}) = G^*(\mathcal{P}) + IG(\mathcal{P})$. This measure could be modified by the factor introduced in Abellán and Moral [1], but this will not be considered here, due to its computational difficulties (it is a supremum that is not easy to compute). The properties of this measure are studied in Abellán and Moral [2, 3] and these are similar to the properties verified by total uncertainty measures in Dempster-Shafer's theory [17].

In this paper, we shall also consider $G^*(\mathcal{P})$ as a measure of total uncertainty. In the particular case of belief functions, Harmanec and Klir [14] consider that maximum entropy is a measure of total uncertainty. They justify it by using an axiomatic approach: it possesses some basic properties. However, uniqueness is not proved. But perhaps the most compelling reason is given in Walley's book [22]. Walley calls this measure the upper entropy. We start by explaining the case of a single probability distribution, P. If You are subject to the logarithmic scoring rule, that means that You are forced to select a probability distribution Q on X that if the true value is x, then You must pay $-\log(Q(x))$. For example, if You say that $Q(x)$ is very small and finally x is the true value, You must pay a lot. If $Q(x)$ is close to one, then you must pay a small amount. Of course, You should choose Q so that $E_P[-\log(Q(x))]$ is minimum, where E_P is the mathematical expectation with respect to P. This minimum is obtained when $Q = P$ and the value

of $E_P[-\log(P(x))]$ is the entropy: the expected loss or the amount that You could accept to be subject to the logarithmic scoring rule. In the case of a credal set, $\mathcal{P}$, we can also have the logarithmic scoring rule, but now we choose Q in such a way that the upper loss $E^*_{\mathcal{P}}[-\log(Q(x))]$ (the supremum of the expectations with respect to the probabilities in $\mathcal{P}$) is minimum. Walley shows that this minimum is obtained for the distribution $P_0 \in \mathcal{P}$ with maximum entropy. Furthermore, $E^*_{\mathcal{P}}[-\log(P_0(x))]$ is equal to the maximum entropy in $\mathcal{P}$: $G^*(\mathcal{P})$. This is the minimum payment You require before being subject to the logarithmic scoring rule. This argument is completely analogous with the probabilistic one, except that we change the expectation for the upper expected loss. This is really a measure of uncertainty, as the better we know the true value of x, then the less we should need to accept the logarithmic scoring rule (lower value of $G^*(\mathcal{P})$). We are not saying that $\mathcal{P}$ can be replaced by the distribution of maximum entropy, only that its uncertainty can be measured by considering maximum entropy in the credal set.

3 Notation and Previous Definitions

For a classification problem we shall consider that we have a data set $\mathcal{D}$ with values of a set $\mathcal{L}$ of discrete and finite variables $\{X_i\}_1^n$. Each variable will take values on a finite set $\Omega_{X_i} = \{x_i^1, x_i^2, ..., x_i^{|\Omega_{X_i}|}\}$. Our aim will be to create a classification tree on the data set $\mathcal{D}$ of one target variable C, with values in $\Omega_C = \{c^1, c^2, ..., c^{|\Omega_C|}\}$.

Definition 4 *A configuration of $\{X_i\}_1^n$ is any m-tuple*

$$(X_{r_1} = x_{r_1}^{t_{r_1}}, X_{r_2} = x_{r_2}^{t_{r_2}}, ..., X_{r_m} = x_{r_m}^{t_{r_m}}),$$

where $x_{r_j}^{t_{r_j}} \in \Omega_{r_j}$, $j \in \{1, ..., m\}$, $r_j \in \{1, ..., n\}$ and $r_j \neq r_h$ with $j \neq h$. That is, a configuration is an assignment of values for some of the variables in $\{X_i\}_1^n$.

If $\mathcal{D}$ is a data set and σ is a configuration, then $\mathcal{D}[\sigma]$ will denote the subset of $\mathcal{D}$ given by the cases which are compatible with configuration σ (cases in which the variables in σ have the same values as the ones assigned in the configuration).

Definition 5 *Given a data set and a configuration σ of variables $\{X_i\}_1^n$ we consider the credal set $\mathcal{P}_C^\sigma$ for variable C with respect to σ defined by the set of probability distributions, p, such that*

$$p_j \in \left[\frac{n_{c^j}^\sigma}{N+s}, \frac{n_{c^j}^\sigma + s}{N+s} \right],$$

for every $j \in \{1,...,|\Omega_C|\}$*, where for a generic state* $c^j \in \Omega_C$*,* $n^{\sigma}_{c^j}$ *is the number of occurrences of* $\{C = c^j\}$ *in* $\mathcal{D}[\sigma]$*,* N *is the number of cases in* $\mathcal{D}[\sigma]$*, and* $s > 0$ *is a parameter.*

We denote this interval as

$$\left[\overline{P}(c^j|\sigma), \underline{P}(c^j|\sigma)\right].$$

This credal set is the one obtained on the basis of the imprecise Dirichlet model, Walley [23], applied to the subsample $\mathcal{D}[\sigma]$.

The parameter s determines how quickly the lower and upper probabilities converge as more data become available; larger values of s produce more cautious inferences. Walley [23] suggests a candidate value for s between $s = 1$ and $s = 2$, but no definitive statement is given.

4 Classification Procedure

We have proposed two methods to build a classification tree: the simple method [4] and the double method [5]. Here we describe the double procedure and give the simple as a particular case.

A classification tree is a tree where each interior node is labeled with a variable of the data set X_j with a child for each one of its possible values: $X_j = x_j^t \in \Omega_{X_j}$. In each leaf node, we shall have a credal set for the variable to be classified, $\mathcal{P}_C^{\sigma}$, as defined above, where σ is the configuration with all the variables in the path from the root node to this leaf node, with each variable assigned to the value corresponding to the child followed in the path. We use a measure of total uncertainty to determine how and when to carry out a branching of the tree. The method starts with a tree with a single node, which will have an empty configuration associated. This node will be open. In this node the set of variables $\mathcal{L}^*$ is equal to the list of variables in the database.

I. For each open node already generated, we compute the total uncertainty of the credal set associated with the configuration, σ, of the path from the root node to that node: $TU(\mathcal{P}_C^{\sigma})$. Then we calculate the values of α and β with

$$\alpha = \min_{X_i \in \mathcal{L}^*} \left(\sum_{r \in \{1,...,|\Omega_{X_i}|\}} \rho_{\{x_i^r\}|\sigma} TU(\mathcal{P}_C^{\sigma \cup (X_i = x_i^r)}) \right)$$

$$\beta = \min_{X_i, X_j \in \mathcal{L}^*} \left(\sum_{r \in \{1,..,|\Omega_{X_i}|\}, t \in \{1,..,|\Omega_{X_j}|\}} \rho_{\{x_i^r, x_j^t\}|\sigma} TU(\mathcal{P}_C^{\sigma \cup (X_i = x_i^r, X_j = x_j^t)}) \right),$$

where $\mathcal{L}^*$ is the set of variables of the data set minus those that appear on the path from the actual node to the root node, $\rho_{\{x_i^r\}|\sigma}$ is the relative

frequency with which X_i takes the value x_i^r in $\mathcal{D}[\sigma]$, $\rho_{\{x_i^r,x_j^t\}|\sigma}$ is the relative frequency with which X_i and X_j take values x_i^r and x_j^t, respectively, in $\mathcal{D}[\sigma]$, and $\sigma \cup (X_i = x_i^r)$ is the result of adding the value $X_i = x_i^r$ to configuration σ (analogously for $\sigma \cup (X_i = x_i^r, X_j = x_j^t)$).

II. If the minimum of $\{\alpha, \beta\}$ is greater or equal than $TU(\mathcal{P}_C^\sigma)$ (including the case in which $\mathcal{L}^*$ is empty), then the node is closed and the credal set $\mathcal{P}_C^\sigma$ is assigned to it.

III. If the minimum of $\{\alpha, \beta\}$ is smaller than $TU(\mathcal{P}_C^\sigma)$, then if $\alpha \leq \beta$, we choose the variable that attains the minimum in α as branching variable for this node; and if $\alpha > \beta$ we consider the pair of variables X_i, X_j for which the value of β is attained, and select as branching variable that from X_i, X_j with a minimum value of uncertainty (calculated in an individual way as in α computation).

If X_{i_0} is the branching variable we add to this node a child for each one of its possible values. All the children are open nodes.

The simple method does not need β, Abellán and Moral [4]. It only considers α and it carries out a branching if this value is less than or equal to the uncertainty of the actual node ($TU(\mathcal{P}_C^\sigma)$). As above, the branching variable is the one for which the value α is attained. In the double method, we demand that the uncertainty is reduced. However, the double method looks for relationships of two variables with C at the same time. The simple method only considers the information of a single variable about C. In some cases, some multidimensional relationships do not give rise to pairwise relationships between the implied variables, and then they will not be detected by the simple method.

4.1 Decision in the Leaves

In order to classify a new case with observations of all the variables except in the variable to be classified C, we start at the root of the tree and follow the path corresponding to the observed values of the variables in the interior nodes of the tree, i.e. if we are at a node with variable X_i and this variable takes the value x_i^r in this particular case, then we choose the child corresponding to this value. This process is followed until we arrive at a leaf node. We then use the associated credal set about C, $\mathcal{P}_C^\sigma$, to obtain a value for this variable.

We will use a **strong dominance criterion** on C. This criterion generally implies only a partial order, and in some situations, no possible precise classification can be done. We will choose an attribute of the variable $C = c^h$ if $\forall i \neq h$

$$\overline{P}(c^i|\sigma) < \underline{P}(c^h|\sigma)$$

When there is no value dominating all other possible values of C, the output is the set of non-dominated cases (cases c^i for which there is no other case c^h verifying inequality). In this way, we obtain what Zaffalon [26] calls a *credal* classifier, in which, for a set of observations, we obtain a set of possible values for the variable to classify, non-dominated cases, instead of unique prediction. In the experiments, when there is no dominant value, we simply do not classify, without calculating the set of non-dominated attributes. This implies a loss of some valuable information in certain situations.

We want to compare our methods with existing classification methods. These methods classify all the records of the training and test sets, without rejecting any of the cases. In order to carry out a fair comparison with such complete procedures, we also use the **maximum frequency criterion** based on frequency of the data, i.e. we will choose the case with maximum frequency in $\mathcal{D}[\sigma]$ as the attribute of the variable to be classified.

5 Experimentation

We have applied this method to some known data sets, obtained from the *UCI repository of machine learning databases*, which can be found on the following website: http://www.sgi.com/Technology/mlc/db. We use the less conservative parameter $s = 1$, since with $s > 1$, we obtained a high degree of non-classified data in some databases (although with a greater percentage of correct classifications).

We plan to compare the behavior of the two total uncertainty measures we have previously defined:

- $TU1 = G^* + IG$
- $TU2 = G^*$

The data sets are: *Breast*, *Breast Cancer*, *Heart*, *Hepatitis*, *Cleveland*, *Cleveland nominal* and *Pima*(medical); *Australian* (banking); *Monks1* (artificial) and *Soybean-small* (botanical).

These databases were used by Acid [7]. Some of the original data sets have observations with missing values and in some cases, some of the variables are not discrete. The cases with missing values were removed and the continuous variables have been discretized using MLC++ software, available at the website http://www.sgi.com/Technology/mlc. The measure used to discretize them is the entropy. The number of intervals is not fixed and it is obtained following the Fayyad and Irani procedure [13]. Only the training part of the database is used to determine the discretization procedure. In Table 1 there is a brief description of these databases.

In general, when there is no case dominating all the other possible values of the variable to be classified, we simply do not classify this individual.

Data set	N. Tr	N. Ts	N. variables	N. classes
Breast Cancer	184	93	9	2
Breast	457	226	10	2
Heart	180	90	13	2
Hepatitis	59	21	19	2
Cleveland nominal	202	99	7	5
Cleveland	200	97	13	5
Pima	512	256	8	2
Vote1	300	135	15	2
Australian	460	230	14	2
Monks1	124	432	6	2
Soybean-small	31	16	21	4

Table 1: Description of the databases. The column *N. Tr* contains the number of cases of the training set, the column *N. Ts* is the number of cases of the test set, the column *N. variables* is the number of variables in the database and the column *N. classes* is the number of different values of the variable to be classified

Algorithms have been implemented using Java language version 1.1.8. In order to obtain the value of G^* for probability intervals we have used the algorithm proposed in Abellán and Moral [3].

The percentages obtained of correct classifications with the simple model and $TU1$ can be seen in Table 2.

In Table 2, the training column is the percentage of correct classifications in the data set that was used for learning. The $UC(Tr)$ column shows the percentage of rejected cases, i.e. the observations that were not classified by the method due to the fact that no value verifies the strong dominance criterion, and the $UC(Ts)$ column shows the rejected cases in the test set.

In the results presented in Table 2 (Abellán and Moral [4]) there is no overfitting (one of the most common problems of learning procedures): the success of the training set and the test set are very similar.

Only the *Cleveland* database has a high rate of non-classified data. This is the case with the highest number of cases of the variable to be classified and then it is more difficult to obtain a class dominating all the other classes. In this case, we would have obtained more information by changing the output to a set of non-dominated cases. In most of the other databases, the variable to be classified has two possible states and in this situation our classification is equivalent to the set of non-dominated values.

In Table 3, we see the success of other known methods on the same databases, Acid [7]. The NB-columns correspond to the results of the Naive Bayesian classifier on the training set and the test set. Similarly, the C4.5-columns correspond to Quinlan's method [19], based on the ID3 algorithm [18], where a classification

Data set	Training	UC(Tr)	Test	UC(Ts)
Breast Cancer	75.5	0.0	81.7	0.0
Breast	98.0	1.3	96.9	0.9
Heart	92.2	7.2	95.2	6.7
Hepatitis	96.4	5.0	94.7	9.5
Cleveland nominal	62.7	4.4	66.0	5.0
Cleveland	72.8	21.0	69.9	24.7
Pima	79.7	0.2	80.5	0.0
Australian	92.3	3.4	91.0	3.4
Vote1	96.1	6.6	96.9	5.9
Soybean-small	100.0	0.0	100.0	0.0

Table 2: The measured experimental percentages of the simple method and $TU1$. The columns *UC(Tr)* and *UC(Ts)* are the percentages of the rejected cases obtained with the training and the test set respectively.

tree with classical precise probabilities is used. We report the results obtained by Acid [7]. We can see that there is overfitting in these methods, principally in C4.5, being especially notable in certain data sets (*Cleveland nominal*, *Cleveland*, *Hepatitis*).

In Table 4 we can see the results of the simple method with $TU2$ and strong dominance. We have a higher percentage of success and a higher percentage of unclassified cases. This total uncertainty measure obtains larger trees as we can observe for the number of leaves presented in Table 5.

The success of the simple method with all cases classified (0% of rejected cases) with the frequency criterion are presented in Table 6 for the test set, to compare it with the models C4.5 and Naive Bayes. Table 7 shows the results of similar experiments with the double method. We can see the high percentages of correct classifications with $TU2$. These are a little higher than those obtained with $TU1$ and notably higher than the other methods (C4.5 and Naive Bayes).

The results of the simple and double methods are similar (slightly better in the double method). In order to see the potential of the double method we use an artificial database: *Monks1*.

Monks1 is a database with six variables. The variable to be classified has two possible states: $a0$ and $a1$, being $a1$ when the first and the second variables are equal or the fourth variable has the first of its possible four states. This type of dependency is very difficult to find for some classification methods, as this is a deterministic relationship involving more than two variables. The double method should be much better than the simple one.

Table 8 shows the success of the methods C4.5 and Naive Bayes. Table 9 shows the success of the simple and double method with all cases classified.

Data set	NB(Tr)	NB(Ts)	C4.5(Tr)	C4.5(Ts)
Breast Cancer	78.2	74.2	81.5	75.3
Breast	97.8	97.3	97.6	95.1
Cleveland nominal	63.9	57.6	69.3	51.5
Cleveland	78.0	50.5	73.5	54.6
Pima	76.4	74.6	79.9	75.0
Heart	87.8	82.2	83.3	75.6
Hepatitis	96.2	81.5	96.2	85.2
Australian	87.6	86.1	89.3	83.0
Vote1	87.6	88.9	94.5	88.3
Soybean-small	100	93.8	100	100

Table 3: Percentages of another methods

Data set	Training	UC(Tr)	Test	UC(Ts)
Breast Cancer	89.0	16.3	93.5	17.2
Breast	99.1	2.6	98.6	2.6
Cleveland nominal	73.6	21.2	74.4	13.1
Cleveland	82.6	34.0	80.3	31.9
Pima	86.6	15.6	86.2	15.2
Heart	93.9	8.8	93.8	10.0
Hepatitis	96.4	5.0	94.7	9.5
Australian	95.3	6.5	94.4	6.5
Vote1	98.2	5.3	98.4	4.4
Soybean-small	100.0	0.0	100.0	0.0

Table 4: Simple method with TU2 and strong dominance

Data set	TU1	TU2	N of possible leaves
Breast	10	17	512
Cleveland	17	112	635904

Table 5: Number of leaves of the trees obtained with the simple method and $TU1$ and $TU2$

Data set	TU1(Ts)	TU2(Ts)	NB(Ts)	C4.5(Ts)
Breast Cancer	81.7	90.3	74.2	75.3
Breast	96.9	97.8	97.3	95.1
Cleveland nominal	65.7	75.8	57.6	51.5
Cleveland	67.0	80.4	50.5	54.6
Pima	80.5	80.9	74.6	75.0
Heart	93.3	92.2	82.2	75.6
Hepatitis	95.2	95.2	81.5	85.2
Australian	90.9	93.5	86.1	83.0
Vote1	94.8	97.8	88.9	88.3
Soybean-small	100	100	93.8	100

Table 6: Success of the simple method with TU1 and TU2 with the frequency criterion on the test set

Database	TU1(Ts)	TU2(Ts)	NB(Ts)	C4.5(Ts)
Breast Cancer	81.7	91.4	74.2	75.3
Breast	96.9	98.7	97.3	95.1
Cleveland nominal	68.7	74.7	57.6	51.5
Cleveland	67.0	80.4	50.5	54.6
Pima	80.5	82.4	74.6	75.0
Heart	93.3	94.4	82.2	75.6
Hepatitis	95.2	95.2	81.5	85.2
Australian	89.1	91.7	86.1	83.0
Vote1	94.8	98.5	88.9	88.3
Soybean-small	100	100	93.8	100

Table 7: Success of the double method with TU1 and TU2 with the frequency criterion on the test set

Data set	NB(Tr)	NB(Ts)	C4.5(Tr)	C4.5(Ts)
Monks1	79.8	71.3	83.9	75.7

Table 8: C4.5 and Naive Bayes on Monks1

	Simple method		Double method	
Function	Tr	Ts	Tr	Ts
TU1	81.5	80.6	94.4	91.7
TU2	89.5	80.6	96.7	94.4

Table 9: Percentages on *Monks1* of the methods with TU1 and TU2 and all cases classified

We can see some interesting things. There is an appreciable overfitting in C4.5 and Naive Bayes but not in our methods. The percentage obtained with the test set is better in the extended method than in the simple method and there is a difference of 23.1% of the extended method and $TU2$ with respect to Naive Bayes success.

6 Conclusions

In this paper, we have discussed the role of maximum entropy as a total uncertainty measure in credal sets. First, we have revised some decision theoretic justification based on the logarithmic scoring rule. We have carried out a series of experiments in which we compare this measure with the one we had previously used in our experiments. The main conclusion is that, in general, the results are always the same or better when only the maximum entropy is used than when a non-specificity value is added to it (the other total uncertainty measure). And in some cases, the percentages of success are notably better.

Other conclusions from the experiments can be summarized in the following points:

- Imprecise probability methods are outstandingly better than classical probabilistic methods, and also have the option of not classifying difficult cases.
- In general, the double method produces slightly better results than the single one, but in some particular cases the differences can be remarkable.
- Maximum entropy (TU_2) produces larger trees than the other uncertainty measure (TU_1), but even this classifier does not suffer from overfitting.

Acknowledgements

We are very grateful to the anonymous referees for their valuable comments and suggestions.

References

[1] J. Abellán and S. Moral. Completing a Total Uncertainty Measure in Dempster-Shafer Theory. *Int. J. General Systems*, 28:299–314, 1999.

[2] J. Abellán and S. Moral. A Non-specificity Measure for Convex Sets of Probability Distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8:357–367, 2000.

[3] J. Abellán and S. Moral. Maximum entropy for credal sets. To appear in *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2003.

[4] J. Abellán and S. Moral. Using the Total Uncertainty Criterion for Building Classification Trees. *Proceeding of the International Symposium of Imprecise Probabilities and Their Applications*, 1-8, 2001.

[5] J. Abellán and S. Moral. Construcción de árboles de clasificación con probabilidades imprecisas. *Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial*, 2:1035-1044, 2001.

[6] J. Abellán. *Medidas de entropía y distancia en conjuntos convexos de probabilidad: definiciones y aplicaciones*. PhD thesis, Universidad de Granada, 2003.

[7] S. Acid. *Métodos de aprendizaje de Redes de Creencia. Aplicación a la Clasificación*. PhD thesis, Universidad de Granada, 1999.

[8] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth Statistics, Probability Series, Belmont, 1984.

[9] G. Choquet. Théorie des Capacités. *Ann. Inst. Fourier*, 5:131–292, 1953/54.

[10] A.P. Dempster. Upper and Lower Probabilities Induced by a Multivaluated Mapping, *Ann. Math. Statistic*, 38:325–339, 1967.

[11] D. Dubois and H. Prade. A Note on Measure of Specificity for Fuzzy Sets. *BUSEFAL*, 19:83–89, 1984.

[12] D. Dubois and H. Prade. Properties and Measures of Information in Evidence and Possibility Theories. *Fuzzy Sets and Systems*, 24:183–196, 1987.

[13] U.M. Fayyad and K.B. Irani. Multi-valued Interval Discretization of Continuous-valued Attributes for Classification Learning. *Proceeding of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, 1022-1027, 1993.

[14] D. Harmanec and G.J. Klir. Measuring Total Uncertainty in Dempster-Shafer Theory: a Novel Approach, *Int. J. General System*, 22:405–419, 1994.

[15] G.J. Klir and M.J. Wierman. *Uncertainty-Based Information*, Phisica-Verlag, 1998.

[16] S. Kullback. *Information Theory and Statistics*, Dover, 1968.

[17] Y. Maeda and H. Ichihashi. A Uncertainty Measure with Monotonicity under the Random Set Inclusion, *Int. J. General Systems* 21:379–392, 1993.

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

[19] J.R. Quinlan. *Programs for Machine Learning*. Morgan Kaufmann series in Machine Learning, 1993.

[20] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[21] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[23] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J.R. Statist. Soc. B*, 58:3–57, 1996.

[24] R.R. Yager. Entropy and Specificity in a Mathematical Theory of Evidence. *Int. J. General Systems*, 9:249–260, 1983.

[25] M. Zaffalon. A Credal Approach to Naive Classification. *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*, 405-414, 1999.

[26] M. Zaffalon. The Naive Credal Classifier. *Journal of Statistical Planning and Inference*, 105:5–21, 2002.

J. Abellán is with the Universidad de Granada, Spain.

S. Moral is with the Universidad de Granada, Spain.