Mathware & Soft Computing 12 (2005) 83-95

Completing an Uncertainty Criterion of Classification

J. Abellán
Dpto. Ciencias de la Computación e I.A. Univ. Granada,
ETSI Informática, 18071 Granada - Spain
jabellan@decsai.ugr.es

Abstract

We present a variation of a method of classification based in uncertainty on credal set. Similarly to its origin it use the imprecise Dirichlet model to create the credal set and the same uncertainty measures. It take into account sets of two variables to reduce the uncertainty and to seek the direct relations between the variables in the data base and the variable to be classified. The success are equivalent to the success of the first method except in those where there are a direct relations between some variables that decide the value of the variable to be classified where we have a notable improvement.

Keywords. Imprecise probabilities, uncertainty, imprecision, non-specificity, classification, classification trees, credal sets.

1 Introduction

A classic application of the theory of probability is the task of classification, a typical machine learning task, where we have an incoming set of observations, called the training set, and generally we want to obtain a set of rules to assign to any new set of observations one value of the variable to classify. The set used to assess the quality of this set of rules is also called the test set. It 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... Normally, these methods use the theory of Probability to estimate the parameters with a stopping criterion to limit the complexity of the classifier and to avoid overfitting.

In this paper, we will use the theory of imprecise probabilities to build a classification tree. A classification tree is a structure easy to understand and is an efficient classifier. It has its origin in ID3 algorithm by Quinlan [16]. A basic reference is the book by Breiman et al. [6]. Here, we also apply decision trees for classification, but as in Zaffalon [23], the imprecise Dirichlet model is used to estimate the probabilities of belonging to the respective classes defined by the variable to be classified.

In Abellán and Moral [2, 3], we have studied how to measure the uncertainty of a credal set generalizing the measures used in the theory of evidence [8, 18]. We consider two main sources of uncertainty: entropy and non-specificity. We have proved that the proposed functions verify the most basic properties of this type of measures (Abellán and Moral [3], Dubois and Prade [10], Klir and Wierman [13]).

In the first method [5] we started with an empty tree and selecting, in each step, a node and a variable to branch with a greater decreasing in the final entropy of the variable to be classified. In classical probability a branching always implies a decreasing of the entropy. So, it is necessary to include an additional criterion not to create too complex models with over-fitting to the data. With credal sets, a branching will produce a lower entropy but, at the same time, a greater non-specificity. In these conditions, we followed the same procedure as in probability theory, but measuring the total uncertainty of a branching. The stopping criterion was very simple: when every possible branching produces an increment of the total uncertainty (the entropy decrement does not compensate the increment of non-specificity).

Finally to carry out the classification in front of a set of observations, we used a strong dominance criterion to obtain the value of the variable to classify.

The new method quantify the uncertainty of each individual variable in each node of the same way and the uncertainty of the sets of two variables jointly. Going into the node the variable that most reduce the uncertainty belonging to the individual calculations or to the double calculations. So is in the last case go into the variable of the pair that more reduce the uncertainty in an individual way. This method improve one that make go in set of two variables in a node. It allows to see the *future uncertainty* before to branching.

In Section Two we present necessary previous concepts on uncertainty on credal sets. In Section Three we introduce notations and definitions previous to the method. In Section Four we describe in detail our method. In Section Five we will test our procedure with known data sets used in classification and we will make an studio of an artificial database that allows us to know the advantages our new method has.

2 Total Uncertainty on Credal Sets

Theory of evidende is based on the concept of basic probability assignment, and this defines a special type of credal set [8, 18]. In this theory, Yager [22] distinguishes two types of uncertainty. One is associated with cases where the information is focused in sets with empty intersections and the other is associated with cases where the information is focused in sets with cardinality over one. We call them *randomness* and *non-specificity* respectively. Since we consider that a general convex set of probability distributions (a credal set) may contain the same type of uncertainty as a b.p.a., we consider similar randomness and non-specificity measures on it.

In Abellán and Moral [3], we have defined a measure of non-specificity for convex sets that generalizes Dubois and Prade's measure of non-specificity in the theory of evidence [9]. Using the Möbius inverse function for monotonic capacities [7], we

can define:

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

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

where $\wp(X)$ is the power set of X.

Definition 2 *For any mapping $f_{\mathcal{P}} : \wp(X) \to \mathbb{R}$ another mapping $m_{\mathcal{P}} : \wp(X) \to \mathbb{R}$ can be associated by*

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

This correspondence is one-to-one, since conversely, we can obtain

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

as we can see in Shafer [18], who calls the correspondence Möbius inversion.

Definition 3 *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 $m_{\mathcal{P}}$.*

Now, we can define a general function of non-specificity.

Definition 4 *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 on $\mathcal{P}$ the following function of non-specificity:*

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

In Abellán and Moral [4], we have proposed the following measure of randomness for general credal sets

$$GG(\mathcal{P}) = Max \left\{ - \sum_{x \in X} p_x \ln p_x \right\}$$

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 [19], for Dempster-Shafer's theory verifying similar properties. It can be used as one of the components of measure of total uncertainty, Harmanec and Klir [12]. We have proved that this function is also a good randomness measure for credal sets and verifies all the basic properties that were verified in Dempster-Shafer's theory [4].

We define a measure of total uncertainty as $TU(\mathcal{P}) = IG(\mathcal{P}) + GG(\mathcal{P})$. This measure could be modified by the factor introduced in Abellán and Moral [2], 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 [3, 4] and these are similar to the properties verified by total uncertainty measures in Dempster-Shafer's theory [15].

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 variables $\{X_i\}_1^n$ called attribute variables. Each variable will take values or states 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 or attributes in $\Omega_C = \{c^1, c^2, ..., c^{|\Omega_C|}\}$.

Definition 5 *Let $\{X_i\}_1^n$ be a set of discrete variables with values in the finite sets Ω_{X_i}, respectively. We call a configuration of $\{X_i\}_1^n$ 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 $\{X_i\}_1^n$.

Definition 6 *Given a data set and a configuration σ of set $\{X_i\}_1^n$ 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 $j \in \{1, ..., |\Omega_C|\}$, obtained on the basis of the imprecise Dirichlet model, Walley [21], for a generic state $c^j \in \Omega_C$. Here $n_{c^j}^\sigma$ is the number of occurrences of the configuration $\{C = c^j\} \cup \sigma$ in the data set, N is the number of observations compatible with configuration σ and $s > 0$ is a hyperparameter.

We denote this interval as

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

This parameter s determines how quickly the lower an upper probabilities converge as more data become available, larger values of s produce more cautious inferences. Walley [21] suggests a candidate value for s between $s = 1$ and $s = 2$, but no definitive statement is given.

4 Exposition of the method

4.1 Classification procedure

In a similar way that we build our classification tree in the simple method, Abellan and Moral [5], we will build our new method.

A classification tree is a tree where each interior node is labeled with an attribute 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. This method can be described using the following points:

I. We start with an empty tree. We calculate the minimum of the following values:

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

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

with $\rho_{x_i^r}$ the relative frequency of x_i^r, $\rho_{\{x_i^r, x_j^t\}}$ the relative frequency of $\{x_i^r, x_j^t\}$ and $\mathcal{L}$ the list of attribute variables in the data base. This value should be less than $TU(\mathcal{P}_C^{\emptyset})$. In other case, the classification tree will have an only node with $\mathcal{P}_C^{\emptyset}$ and the classification will take into account only the frequency of the states of the variable in classification, and not the values of the rest of the variables.

If $\alpha \leq \beta$ we choose with root node the variable that attains this minimum, in other case we have a pair of attribute variables and we choose of them the one with minimum value of uncertainty in an individual way as in α value.

II. For each node already generated, we compute the total uncertainty of the credal set associated to the configuration, σ, of the path from the root node to that node: $TU(\mathcal{P}_C^{\sigma})$. Again we calculate the minimum value of Then we find the variable X_{i_0} with the value:

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

$$\beta' = \min_{X_i, X_j \in \mathcal{L}^*} \left(\sum_{r \in \{1,..,|\Omega_{X_i}|\}, t \in \{1,..,|\Omega_{X_j}|\}} \rho_{\sigma \cup \{x_i^r, x_j^t\}} TU(\mathcal{P}_C^{\sigma \cup \{x_i^r, x_j^t\}}) \right),$$

where $\mathcal{L}^*$ is the set of attribute variables of the data set minus those that appear in the way from the actual node to the root node.

That is, in a node with a configuration σ, we compute for each attribute variable the weighted average of the total uncertainty of the leaves associated to the branching for this variable, where the weights are the frequencies of occurrence of the different values of the variable under configuration σ. Then, we take the variable X_{i_0}, with minimum value of total uncertainty after branching in a similar sense as in the choose of root node. If this value is lower than the total uncertainty before branching $TU(\mathcal{P}_C^{\sigma})$, this node is labeled with X_{i_0} and a branch is added for each one of its children. The process will be repeated for each one of them.

III. If there is no attribute variable that reduces the uncertainty or $\mathcal{L}^*$ is empty, then this node will be a leaf and will contain the credal set associated to the configuration with the values of the variables in the path from the root node to this leaf.

The original method [5] need the same points, but only the calculation of α and α' no of β and β'. This variation make the method finds some relations between variables for the first method that the new method finds by its construction. The new method is a natural extension of the first.

This way to introduce an attribute variable in a node can improve the one that introduce two attribute variables in that node because it is posible that for a case of the variable the uncertainty can obtain a reduction major than with its pair in the optimum. This allows us to continue reducing the uncertainty.

4.2 Decision in the leaves

To classify a new case with observations in all the attribute variables except in the variable to be classified C, then we start in 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 in a node with variable X_i and this variable takes the value x_i^r in the set of observations, then we choose the child corresponding to this value. This process is followed till we arrive to a leaf node. Then, we use the associated credal set about C 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 it verifies that $\forall i \neq h$

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

When there is no value dominating all other possible values of C, the output can be the set of non-dominated cases (cases c^i for which there is not another case c^h verifying above inequality). In this way, we obtain what Zaffalon [24] 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 an unique prediction. (In the experiments, when there is not a dominant value, we simply do not classify, without calculating the set of non-dominated attributes. This implies to lose some valuable information in some situations.) This avoid the loss of information that we have if no classify some cases where there is a major frequency in two cases and cero in the others as occur in *Cleveland*, where our method has a high grade of no classified cases.

As we will see, we want to compare our methods with another already known. These methods classify all the record of the training and test sets. They no have a rule of rejected data set. So for comparing with these methods we also use a dominance criterion based on frequency of the data, i.e., we will choose an attribute of the variable to be classified with mayor frequency, but this could produce an overfitting on the data base to training, as we will see in Table 5.

An alternative criterion for classification is credal dominance [24] called also strict preference [20]. This criterion is based in comparing the probability of the two cases for each one of the probabilities of the credal set. Strong dominance implies credal dominance, but the converse is not true: there are situations in which there is credal dominance but not strong dominance. However, in this particular case in which we have credal sets that are defined by reachable intervals for the values of the variable to classify it is easy to prove that both criteria are equivalent.

5 Experimentation

We have applied this method to some known data sets, obtained from the Uci Repository of Machine Learning Databases (we can find them in the direction http://www.sgi.com/Technology/mlc/db) with the parameter less conservative $s = 1$, since with $s > 1$ we obtained a high degree of non-classified data in some databases (though with a greater percentage of correct classifications).

The data sets *Breast*, *Breast Cancer*, *Heart*, *Hepatitis* and *Cleveland* (medical); *Vote1* (political); *Australian* (banking); *Soybean-small* (botanical) and *Monks1* (artificial).

These databases were used by Acid [1]. 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 in http://www.sgi.com/Technology/mlc. The measure used to discretize them have been the entropy. The number of intervals is not fixed, and it is obtained following Fayyad and Irani [11] procedure. Only the training part of the database was used to determine the discretization procedure. In Table 1 there is a brief description of these databases. We can see the number of cases of the training set (N. Tr), of the test set (N. Ts), number of variables in the database (N. variables) and the number of different values of the variable to be classified (N. classes).

In general, when there is not a case dominating all the other possible values of the variable to classify, we simply does not classify this individual.

Algorithms have been implemented using Java language version 1.1.8.

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

The Training column is the percentage of correct classifications in the data set that was used for learning. In $UC(Tr)$ column we have 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 in $UC(Ts)$ column we have the rejected cases in the test set. The success rate can be improved if we had used $s = 2$, though the percentage of rejected observations will be also increased.

As we can see in Table 2 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 database *Cleveland* has a high rate of non-classified data. This is the case with the highest number of cases of the variable to classify and then it is

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

Table 1: Description of the databases

Data base	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
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
Heart	92.2	7.2	95.2	6.7
Hepatitis	96.4	5.0	94.7	9.5
Vote1	96.1	6.6	96.9	5.9
Australian	92.3	3.4	91.0	3.4
Soybean-small	100.0	0.0	100.0	0.0

Table 2: The measured experimental percentages of the simple method

more difficult to obtain a class dominating all the other values. 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 compare the results with another known methods with good behavior on the same databases, Acid [1]. We have used the same sets of training and test that the used for this experiments.

The NB-columns correspond to results of the Naive Bayesian classifier on the Training set and the Test set. This known method is based on the conditional independence of the variables given the variable to be classified. Similarly, the C4.5-columns correspond to the Quinlan's method [17], based on ID3 [16], where a classification tree with classical precise probabilities is used. It is possible to obtain a implementation of this method in http://www.sgi.com/Technology/mlc.

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
Vote1	87.6	88.9	94.5	88.3
Australian	87.6	86.1	89.3	83.0
Soybean-small	100	93.8	100	100

Table 3: Percentages of another methods

Data base	Training	UC2(Tr)	Test	UC2(Ts)
Breast Cancer	75.5	0.0	81.7	0.0
Breast	98.0	1.3	96.9	0.9
Cleveland nominal	64.6	5.0	68.8	6.1
Cleveland	72.8	21.0	69.9	24.7
Pima	79.7	0.2	80.5	0.0
Heart	91.7	6.1	94.1	5.6
Hepatitis	96.4	5.0	94.7	9.5
Vote1	96.1	6.6	96.9	5.9
Australian	90.8	0.6	89.0	0.9
Soybean-small	100.0	0.0	100.0	0.0

Table 4: The measured experimental percentages with the strong dominance criterion with the new model

We report the results obtained by Acid [1].

We can see that there is overfitting in these methods, principally in C4.5 being specially notable in some data sets (*Cleveland nominal*, *Cleveland*, *Hepatitis*.

Now, we can see the success of our extended method in Table 4 and Table 5. In the first we have the success with strong dominance criterion and in the second we have the mayor frequency criterion, i.e., with all the cases classified (0% of rejected cases), to compare it with the models C4.5 and Naive Bayes.

As we can see in Table 4, there is a little variation with respect to to simple method, though we must into account that the simple method branching if the uncertainty is equal to the node actual.

There is a high percentage of no classified in data base *Cleveland* where the variable to be classified have 5 posible states. We can think that our method fails, but we do not have an avoid of information because there is a lot of cases where

Data base	Training	Test
Breast Cancer	75.5	81.7
Breast	97.6	96.9
Cleveland nominal	64.9	68.7
Cleveland	68.0	67.0
Pima	79.7	80.5
Heart	90.0	92.2
Hepatitis	96.6	95.2
Vote1	94.0	94.8
Australian	90.9	89.1
Soybean-small	100.0	100.0

Table 5: The measured experimental percentages with no rejected cases with the new model

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

Table 6: C4.5 and Naive Bayes on Monks1

there are two non-dominated states. It make us to think to introduce the criterion that Zaffalon [24] expound.

In Table 5 the percentages of success with all classified records decrease a little, although there is an high grade of no classified data, as we can see in Table 4.

Another thing to be into account is the increasing of success for data bases *Heart* and *Hepatitis* when we classified all the records. This is no rare because in this data base we obtain a lot of leaf that do no have classification because the frequency are 1 and 0 for the states of the variable to be classified. So when we force the classification we have a 100% of success in these cases having a higher percentage of success, as we can see in Table 4 and Table 5. This could make to increase the overfitting in another data bases.

Now, in these data bases there is no important relations to consider important our variation of the original method. To see the potencial of the new method we use an artificial data base as *Monks1*.

Monks1 is a data base with six variables. The variable to be classified has two posible states, $a0$ and $a1$. Being $a1$ when the first and the second variables are equal or the fourth variable has the first of its posible four states. This type of dependency is very difficult to find for the classification methods and make this type of data base no popular.

In Table 6 we have the success of the methods C4.5 and Naive Bayes.

In Table 7 we have the success of the original method (UC) and of its ampliation (UC2) with all cases classified.

Data set	UC(Tr)	UC(Ts)	UC2(Tr)	UC2(Ts)
Monks1	81.5	80.6	94.4	91.7

Table 7: Uncertainty methods on Monks1

There is an apreciable overfitting in C4.5 and Naive Bayes but no in our methods. The percentage in the test set is major with UC2 that with UC, being of 20.4% with respect to Naive Bayes success.

6 Conclusions

We have present an interesting variation of our original method. The new method satisfies: it not suffer of overfitting as the original method, it reduces the uncertainty before the prime do (it produces an inferior number of leaf nodes), in the experimentation it has a slight better percentages of correct classifications than the prime when there are not relations between the attribute variables and the variable to be classified and in these cases it is able to find the relations that the the prime it does not, as we can see for the artificial data set *Monks1*.

We want improve our method using another measures of total uncertainty and to introduce a mixture of our method and the Naive Bayesian classifier because, as we can see for data set *Breast*, when there is independence between the variables known the variable to be classified, the Naive Bayes method obtain a good percentage of correct classification cases.

In our experiments we have rejected the missing data, however as pointed out in [25], credal sets can be an appropriate tool to deal with missing data and so they can be naturally incorporated to credal classification trees. We plan to investigate this possibility in the future.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Technology under project Algra (TIN2004-06204-C03-02).

References

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

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

[3] 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.

[4] J. Abellán and S. Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(5): 587–597, 2003.

[5] J. Abellán and S. Moral. Building Classification Trees using the Total Uncertainty Criterion. *International Journal of Intelligent Systems*, 18: 1215–1225, 2003.

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

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

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

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

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

[11] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[25] M. Zaffalon. Exact Credal Treatment of Missing Data. *Journal of Statistical Planning and Inference*, 105(1): 105–122, 2002.