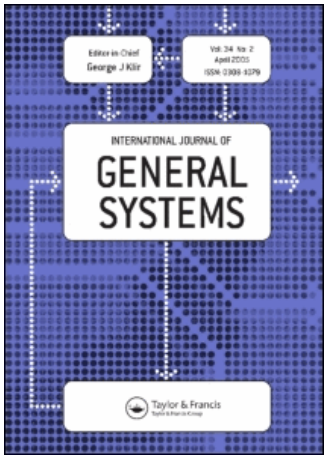

International Journal of General Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713642931

Application of uncertainty measures on credal sets on the naive Bayesian classifier

Joaquín Abellán [a]
[a] Department of Computer Science and Artificial Intelligence, University of Granada. Granada. Spain

To cite this Article: Joaquín Abellán , 'Application of uncertainty measures on credal sets on the naive Bayesian classifier', International Journal of General Systems, 35:6, 675 - 686
To link to this article: DOI: 10.1080/03081070600867039
URL: http://dx.doi.org/10.1080/03081070600867039

International Journal of General Systems, Vol. 35, No. 6, December 2006, 675–686

Taylor & Francis
Taylor & Francis Group

Application of uncertainty measures on credal sets on the naive Bayesian classifier

JOAQUÍN ABELLÁN*

Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

(Received 6 April 2006; in final form 31 May 2006)

The naive Bayes classifier is known to obtain good results with a simple procedure. The method is based on the independence of the attribute variables given the variable to be classified. In real databases, where this hypothesis is not verified, this classifier continues to give good results. In order to improve the accuracy of the method, various works have been carried out in an attempt to reconstruct the set of the attributes and to join them so that there is independence between the new sets although the elements within each set are dependent. These methods are included in the ones known as semi-naive Bayes classifiers. In this article, we present an application of uncertainty measures on closed and convex sets of probability distributions, also called credal sets, in classification. We represent the information obtained from a database by a set of probability intervals (a credal set) via the imprecise Dirichlet model and we use uncertainty measures on credal sets in order to reconstruct the set of attributes, such as those mentioned, which shall enable us to improve the result of the naive Bayes classifier in a satisfactory way.

Keywords: Imprecise probabilities; Imprecise Dirichlet model; Uncertainty measures; Maximum entropy; Classification; Naive Bayesian classifier

1. Introduction

Classification is an important problem in the field of machine learning where the classic theory of probability has traditionally been used. The problem can be summarized in the following way: we have a set of observations, called the training set and we want to obtain a set of laws from it so that each new observation may be assigned a value of the variable to be classified. The set used to verify the quality of this set of laws is called the test set. This classification has considerable applications in medicine, physics, character recognition, astronomy, economy, etc. With objectives such as disease recognition, weather forecasts, loan concessions, etc. rules are applied that allow us to associate a possible value of a variable to a new observation with a specific set of values of other variables. The analyzed objects (i.e. patients, meteorological situations, fuzzy characters, stars, or bank customers) have certain variable values that can be appreciated, observations or attribute variables and our aim is to

*Email: jabellan@decsai.ugr.es

International Journal of General Systems
ISSN 0308-1079 print/ISSN 1563-5104 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/03081070600867039

predict another value of a variable that we shall call the variable to be classified or the class variable. We shall focus on the problem where both the class variable and the attributes variables are discrete.

For this type of problem, Duda and Hart (1973) introduced the naive Bayes classifier which is based on the consideration that the attributes are independent given the class variable. This supposition of independence allows them to obtain the probability of a joint event as the product of the individual ones. In literature, this method has demonstrated that this simple and efficient classification model obtains good results on real data sets even when the independence condition is weak.

Looking for a condition or pseudo-condition of independence between the attribute variables in order to improve the accuracy of the naive Bayes classifier, different procedures have been introduced on the set of attribute variables. These procedures can basically be divided into two groups (Zheng and Webb 2005): those that join and eliminate attributes (Kittler 1986, Kononenko 1991, Langley and Sage 1994, Pazzani 1996) and those that establish dependency connections between attributes (Friedman *et al.* 1997, Keogh and Pazzani 1999, Zheng and Webb 2000, Webb *et al.* 2005).

In this article, we shall present an application of the uncertainty measures on imprecise probabilities and shall propose a model to group attribute variables in a similar way to Pazzani (1996). We shall join the variables that have some relation of dependency given the class variable. In order to represent the information which each attribute variable expresses on the class variable, we shall use the imprecise Dirichlet model (IDM) (Walley 1996) obtaining a set of probability intervals that also can be expressed by a belief function (Abellán 2006). In order to determine the degree of dependency between variables, we shall consider measures of uncertainty/information[†] on closed and convex sets of probability distributions (Abellán and Moral 2000, 2003, Klir 2006, Abellán *et al.* 2006a) (otherwise known as credal sets) which are obtained in the IDM application. We shall use the maximum entropy function to measure total uncertainty on the credal sets obtained. This measure includes both types of uncertainty that all measures of total or global uncertainty must quantify (Klir and Wierman 1998): conflict and non-specificity. Some authors have proposed the maximum of entropy as the best measure of total uncertainty on credal sets (Klir and Smith 2001, Abellán and Moral 2005b), but it has only been recently that this measure has been justified and coherently separated into the parts of conflict and non-specificity (Abellán *et al.* 2006a).

In order to check our proposed model, we shall apply it on a series of well-known databases where the accuracy of the naive Bayes classifier is not totally satisfactory and shall compare it with one of the models included within the so-called semi-naive Bayes classifiers that obtain the best average result according to work by Zheng and Webb (2005): averaged one-dependence estimators (AODE) of Webb *et al.* (2005). We will see that with our method, the results of the naive Bayes classifier are notably improved for this sets of databases.

In Section 2 of this paper, we shall detail the necessary previous knowledge for the development of the work and we shall briefly define the naive Bayes classifier and the

[†]We consider the concept of "information based on uncertainty" (Klir 2006) relating to information deficiency (incomplete, vague, fuzzy, contradictory, deficient, etc.) that can appear from different types of uncertainty. We shall always refer to the term "information" in the context of reduction of uncertainty, unlike its use in logic or in computability theory.

imprecise Dirichlet model (used to represent information from a sample) and shall outline the development of uncertainty measures on credal sets. In Section 3, we shall present our method for joining variables based on the IDM model and uncertainty measures. In Section 4, we shall check our procedure on well-known databases and shall compare the results with those obtained with the AODE model. Finally, Section 5 shall summarize our conclusions and future lines of work.

2. Previous knowledge

2.1 Naive Bayes

The success of Duda and Hart's (1973) naive Bayes classifier is mainly due to its simplicity, efficiency and effectiveness in classification problems. In order to present the method, we shall previously define the problem of supervised classification on which the work focuses. We considered a database $\mathcal{D}$ of values of a set $\mathcal{L}$ of discrete or discretized attribute variables $\{X_i | i = 1, \ldots, r\}$, where each attribute variable has a set of possible states or cases $\Omega_{X_i} = \{x_1^i, x_2^i \ldots, x_{|\Omega_{X_i}|}^i\}$ and a class variable C, with states in the set $\Omega_C = \{c_1, c_2 \ldots, c_k\}$. The objective is to obtain information from the database so that given a new observation (a set of values of all the attribute variables), we are able to associate it with a value of the class variable.

If we denote a new observation as $\mathbf{x}$, when $\mathbf{x} = \{x_{h_1}^1 \ldots, x_{h_r}^r\}$, with $h_i \in \{1, \ldots, |\Omega_{X_i}|\}$. The naive Bayes classifier selects the value c_i in the following way:

$$\text{argmax}_{c_i}(P(c_i|\mathbf{x})),$$

and supposing that the attribute variables are independent given the class variable, this can be expressed as:

$$\text{argmax}_{c_i}\left(P(c_i)\prod_{j=1}^{r} P(x_{h_j}^j|c_i)\right).$$

2.2 Credal sets

Various mathematical models can be used to represent the information available in a certain situation. None of these is generally more justifiable than another, but each is more useful than the others in specific situations. Walley (1991) compiles most of the mathematical models for representing the absence of information through imprecise probabilities. In this section, we shall introduce the model based on imprecise probabilities that we will use: reachable sets of probability intervals.

2.2.1 Reachable sets of probability intervals. As an important reference on this type of credal set, we should mention the work by Campos *et al.* (1994), where we can find an excellent account of the basic operations for working with probability intervals, as well as their relation with other models such as those of upper and lower probabilities, capacities of order 2 and belief functions.

The main characteristic of this model is that there are many interesting operations between sets of probability intervals without having to leave the model, i.e. providing us with another set of probability intervals.

They can be described as follows: let X be a variable that takes values in $\Omega_X = \{x_1, x_2, \ldots, x_{|\Omega_X|}\}$. A system of probability intervals is a family of intervals $L = \{[l_i, u_i] : i \in \{1, 2, \ldots, |\Omega_X|\}\}$ verifying that $0 \leq l_i \leq u_i \leq 1$. The credal set associated to a set of intervals L on X can be defined as:

$$K_L^X = \{p \in \mathcal{P}(\Omega_X) | \, l_i \leq p_i \leq u_i, \, \forall i\},$$

expressing p_i as $p(\{x_i\})$ and $\mathcal{P}(\Omega_X)$ as the set of all probability distributions on Ω_X.

One condition so that this set is nonempty is that

$$\sum_i l_i \leq 1 \leq \sum_i u_i.$$

Any element in the set $\{l_i, u_j | \, i, j \in \{1 \ldots |\Omega_X|\}\}$ therefore belongs to at least one probability distribution of K_L^X (which is why the set of intervals is defined as reachable) and the following conditions must be verified:

$$\sum_{j \neq i} l_j + u_i \leq 1, \quad \sum_{j \neq i} u_j + l_i \geq 1,$$

for each i. If this set of conditions is not verified, it is possible to obtain the reachable set of intervals from the following property:

PROPOSITION 1. Given a set of probability intervals $L = \{[l_i, u_i] : i \in \{1, \ldots, |\Omega_X|\}\}$, the set $L' = \{[l_i', u_i'] : i \in \{1, \ldots, |\Omega_X|\}\}$ where

$$l_i' = \max_i \{l_i, 1 - \sum_{i \neq j} u_j\}, \quad u_i' = \min_i \{u_i, 1 - \sum_{i \neq j} l_j\},$$

give us the same set of probability distributions, $K_L^X = K_{L'}^X$, where this last set is a reachable set of probability intervals.

2.3 Imprecise Dirichlet model

The IDM was introduced by Walley (1996) to infer about the probability distribution of a categorical variable. Let us assume that Z is a variable taking values on a finite set Z and that we have a sample of size N of independent and identically distributed outcomes of Z. If we want to estimate the probabilities, $\theta_z = p(z)$, with which Z takes its values, a common Bayesian procedure consists in assuming a prior Dirichlet distribution for the parameter vector $(\theta_z)_{z \in Z}$ and then taking the posterior expectation of the parameters given the sample. The Dirichlet distribution depends on the parameters s, a positive real value and $\mathbf{t}$, a vector of positive real numbers $\mathbf{t} = (t_z)_{z \in Z}$, verifying $\sum_{z \in Z} t_z = 1$. The density takes the form

$$f((\theta_z)_{z \in Z}) = \frac{\Gamma(s)}{\prod_{z \in Z} \Gamma(s \cdot t_z)} \prod_{z \in Z} \theta_z^{s \cdot t_z - 1},$$

where Γ is the gamma function.

If n_z is the number of occurrences of value z in the sample, the expected posterior value of parameter θ_z is $(n_z + s \cdot t_z)/(N + s)$, which is also the Bayesian estimate of θ_z (under quadratic loss).

The imprecise Dirichlet model only depends on parameter s and assumes all the possible values of $\mathbf{t}$. This defines a convex set of prior distributions. It represents a much weaker assumption than a precise prior model, but it is possible to make useful inferences. In our particular case, where the IDM is applied to a single variable, we obtain a credal set for this variable Z that can be represented by a system of probability intervals. For each parameter θ_z, we obtain a probability interval given by the lower and upper posterior expected values of the parameter given the sample. These intervals can be easily computed and are given by $[n_z/(N + s), (n_z + s)/(N + s)]$. The associated credal set on X is given by all the probability distributions p' on Z, such that $p'(z) \in [n_z/(N + s), (n_z + s)/(N + s)]$, $\forall z$. The intervals are coherent in the sense that if they are computed by taking infimum and supremum in the credal set, then the same set of intervals is again obtained. The associate credal set can be obtained in the same way as in the previous subsection,

$$K^Z_{\text{Lidm}} = \{p \in \mathcal{P}(\Omega_Z) | l_i \leq p_i \leq u_i,\ l_i = \frac{n_{z_i}}{N+s},\ u_i = \frac{n_{z_i}+s}{N+s},\ \forall i\},$$

and represents a credal set from a reachable set of probability intervals.

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 (1996) does not provide a definitive recommendation, but he advocates values between $s = 1$ and $s = 2$. In Bernard (2005), we found reasons for using values greater than 1 for s and in Abellán *et al.* (2006b), the value $s = 1.5$ is used for classification applications.

2.4 Uncertainty measures on credal sets

The study of uncertainty measures in the Dempster–Shafer theory of evidence (Dempster 1967, Shafer 1976) has been the starting point for the development of these measures on more general theories (a study of the most important measures proposed in literature can be seen in Klir 2006). As a reference for the definition of an uncertainty measure on credal sets, Shannon's entropy (Shannon 1948) has been used due to its operation on probabilities. In any theory which is more general than the probability theory, it is essential that a measure be able to quantify the uncertainty that a credal set represents: the parts of conflict and non-specificity (Klir 2006).

In recent years, Klir and Smith (2001) and Abellán and Moral (2005b) justified the use of the maximum of entropy on credal sets as a good measure of total uncertainty that verifies a set of needed properties (Klir and Wierman 1998). The problem lies in separating this function into others, which really do measure the parts of conflict and non-specificity, respectively and this entails the use of a credal set to represent the information. More recently, Abellán *et al.* (2006) presented a separation of the maximum of entropy into functions which are capable of coherently measuring the conflict and non-specificity of a credal set K on a finite variable X, as well as algorithms for facilitating its calculation in capacities of order 2 (Abellán and Moral 2005a, 2006) and this may be expressed in the following way:

$$S^*(K) = S_*(K) + (S^* - S_*)(K),$$

where S^* represents the maximum of entropy and S_* represents the entropy minimum on the credal set K:

$$S^*(K) = \max_{p \in K} \sum_x p_x \log(p_x), \quad S_*(K) = \min_{p \in K} \sum_x p_x \log(p_x).$$

where $S_*(K)$ coherently quantifies the conflict part of the credal set K and $(S^* - S_*)(K)$ represents the non-specificity part of K (Abellán *et al.* 2006a).

In order to obtain the maximum of entropy on a set of probability intervals in the application of the IDM model from a sample, we can use the algorithm presented in Abellán and Moral (2003) for probability intervals. When using values between 1 and 2 for the parameter s, we can use a simpler procedure of Abellán (2006), which is a simplification of the one presented in Abellán and Moral (2003). Given a credal set K^X_{Lidm} defined as in the previous section, we must first determine the set $H = \{x_j | n_{x_j} = \min_i\{n_{x_i}\}\}$. Let $|H|$ be the cardinal of the set H. If we use $\hat{p}$ to denote the distribution where the maximum of entropy will be reached, the procedure of Abellán (2006) can be expressed in the following way:

Case 1. $|H| > 1$ or $s = 1$

$$\hat{p}(x_i) = \begin{cases} \frac{n_{x_i}}{N+s} & x_i \notin H \\ \frac{n_{x_i} + s/|H|}{N+s} & x_i \in H. \end{cases}$$

Case 2. $|H| = 1$ and $s > 1$.

Assign:
$n_{x_j} \leftarrow n_{x_j} + 1$ (where $H = \{x_j\}$),
$s \leftarrow s - 1$.
Obtain new H.
Obtain $\hat{p}$ as in Case 1.

3. Presentation of the method for combine variables

In a similar way to Pazzani (1996), we shall introduce a method that obtains Cartesian products of attribute variables as a prior step to the application of the naive Bayes classifier. We shall use the maximum entropy as the total measure of uncertainty or information[‡] on the obtained probability intervals of the IDM application. The general concept is very simple: we shall join the attribute variables that are more informative jointly than separately. In order to prevent the set of new attribute variables being too complex (i.e. it does not produce an excessive number of variables), we shall limit the number of variables that can contain a new variable as Huang *et al.* (2002) did, establishing an informative threshold (which is no more than a value that the joint information to the sum of the individual information should not

[‡]Considering an uncertainty measurement U we can express the measurement of information associated as $-U$.

exceed). We call this value as **u**. Consequently, the number of attribute variables in each new attribute variable will not surpass a value that will depend on the chosen threshold ($\eta(\mathbf{u})$).

Formally, we have a set of attribute variables $\{X_1, \ldots, X_r\}$ and we want to obtain another set $\{W_1, \ldots, W_r\}$ with $v \leq r$, such that:

$$W_j = \bigcup_{u=1}^{d_j} X_{j_t}, \quad d_j \leq \eta(\mathbf{u}), \quad \forall j,$$

$$W_j \cap W_i = \emptyset, \quad \forall i \neq j, \quad i,j \in \{1,\ldots,v\},$$

$$\bigcup_{j=1}^{v} W_j = \{X_1,\ldots,X_r\},$$

therefore applying the naive Bayes as usual

$$\text{argmax}_{c_i}\left(P(c_i)\prod_{j=1}^{v} P(w_{h_j}^j|c_i)\right),$$

with $w_{h_j}^j$ being the elements of the Cartesian product of the variables X_{j_u} that comprise the new variable W_j.

In order to describe the procedure for obtaining the set of the new variables $\mathbf{W} = \{W_1,\ldots,W_v\}$ from a data set $\mathcal{D}$, we shall use $\text{Inf}(A, B|C)$ to denote the value of the information gain that gives the attribute variables A, B on the class variable C and we shall use $\text{Inf}(A|C)$ and $\text{Inf}(B|C)$ to denote the equivalent information for A and B, respectively. Starting from $W_i = X_i$, $\forall i$, the procedure can be expressed of the following form:

Procedure New-Variables (**W**,**u**)

1. For all $j \neq k$ obtain $\text{Inf}(W_j, W_k|C)$, $\text{Inf}(W_j|C)$, $\text{Inf}(W_k|C)$.
2. For all $j > k$ obtain
   $F(W_j, W_k) = \text{Inf}(W_j, W_k|C) - \mathbf{u} - \text{Inf}(W_j|C) - \text{Inf}(W_k|C)$
3. If $\max_{j>k}\{F(W_j, W_k)\} \leq 0$ Exit
4. Else
   5. Let $\{W_\alpha, W_\beta\} = \arg\max_{j>k}\{F(W_j, W_k)\}$
   6. Assign $W_\alpha \leftarrow W_\alpha \cup W_\beta$
   7. $\mathbf{W} \leftarrow \mathbf{W} - W_\beta$
8. Call New-Variables (**W**,**u**)

It is now necessary to describe how we shall obtain the functions $\text{Inf}(W_k|C)$ and $\text{Inf}(W_j, W_k|C)$, or equivalently the functions $F(W_j, W_k)$, for all the attribute variables W_j,W_k.

3.1 Obtaining the information gain measure

As can be seen from its description, the previous procedure for acquiring new variables allows us to use different measures of uncertainty/information of the attribute variables on

the class variable. Our proposal is to use the IDM model to represent the information by means of a credal set and to obtain the value of the uncertainty by applying the maximum entropy function on the obtained credal set. For the formal description, the following configuration concept was necessary:

DEFINITION 1. A configuration, σ, on a set of variables $\mathbf{W}$ is an assignment of values for a subset of variables: $\mathbf{Y} = \mathbf{y}$, where $\mathbf{Y} \subseteq \mathbf{W}$.

In order to obtain the function $\mathrm{Inf}(A|C)$ for an attribute variable A with values in $\Omega_A = \{a_1 \ldots, a_{|\Omega_A|}\}$, we will use $n_{c_j}^{\{A=a_i\}}$ to denote the frequency of the configuration $\{A = a_i, C = c_j\}$ in the data set $\mathcal{D}$. From the set of frequencies obtained using the variables A and C, we can have a credal set for each value $A = a_j$ by applying the IDM model:

$$K^C_{\mathrm{Lidm}}(\{A = a_j\}) = \{p \in \mathcal{P}(\Omega_C) | \frac{n_{c_i}^{\{A=a_j\}}}{N+s} \leq p_i \leq \frac{n_{c_i}^{\{A=a_j\}} + s}{N+s}, \quad \forall i\}.$$

By considering only the values n_{c_i}, $\forall j$, we can obtain the set $K^C_{\mathrm{Lidm}}(\emptyset)$ in the same way. We therefore defined the information gain function of the attribute variable A on the class variable C as:

$$\mathrm{Inf}(A|C) = \sum_j \rho_{a_j} S^*(K^C_{\mathrm{Lidm}}(\{A = a_j\})) - S^*(K^C_{\mathrm{Lidm}}(\emptyset)),$$

where $\rho_{\{A=a_j\}}$ is the relative frequency of the configuration $\{A = a_j\}$ in the data set $\mathcal{D}$. Similarly, the values $\mathrm{Inf}(A, B|C)$ can be obtained for all the pairs of attribute variables A, B.

Using the same notation, we would obtain a credal set for each configuration $\sigma_{j,k} = \{A = a_j, B = b_k\}$:

$$K^C_{\mathrm{Lidm}}(\sigma_{j,k}) = \{p \in \mathcal{P}(\Omega_C) | \frac{n_{c_i}^{\sigma_{j,k}}}{N+s} \leq p_i \leq \frac{n_{c_i}^{\sigma_{j,k}} + s}{N+s}, \forall i\}.$$

Hence:

$$\mathrm{Inf}(A, B|C) = \sum_j \rho_{\sigma_{j,k}} S^*(K^C_{\mathrm{Lidm}}(\sigma_{j,k})) - S^*(K^C_{\mathrm{Lidm}}(\emptyset)),$$

with $\rho_{\sigma_{j,k}}$ being the relative frequency of the configuration $\sigma_{j,k}$ in the data set $\mathcal{D}$. Finally, the function $F(A, B)$, on the attribute variables A, B, used in the new variables procedure would be:

$$F(A, B) = \sum_{j,k} \rho_{\sigma_{j,k}} S^*(K^C_{\mathrm{Lidm}}(\sigma_{j,k})) + S^*(K^C_{\mathrm{Lidm}}(\emptyset)) - \mathbf{u} - \sum_j \rho_{\{A=a_j\}} S^*(K^C_{\mathrm{Lidm}}(\{A = a_j\})) - \sum_k \rho_{\{B=b_k\}} S^*(K^C_{\mathrm{Lidm}}(\{B = b_k\}))$$

4. Experimentation

As we can see in the extensive work of Webb *et al.* (2005) on 37 databases, the naive Bayes classifier obtains better results than J48 method of classification, that corresponds to an

Table 1. Description of the databases.

Databases	*N*	*r*	*k*
Car	1728	6	4
Monks1	556	6	2
Tic-tac-toe	958	9	2
Corral	160	6	2

improved version of the C4.5 method of Quinlan (1993), based on the ID3 method (Quinlan 1986), which uses a classification tree with classic probabilities[¶]. Also, in this work, we can see that the AODE of Webb *et al.* (2005) obtains the best average result compared with others similar methods.

In order to check the improvement obtained with naive Bayes classifier using the method presented above, we have used 4 well-known databases where J48 obtains notable better results than the naive Bayes classifier motivated by the relations between the attribute variables. These databases can be found in the UCI repository of machine learning databases which can be obtained directly from ftp:ftp.ics.uci.edu/pub/machine-learning-databases. In order not to benefit any classification method with a discretization procedure, we have used only databases with discrete variables. In some cases, some variables have missing values that were eliminated. In table 1, there is a brief description of these databases. We can see the number of cases in the database (N), the number of attribute variables in the database (r) and the number of different states of the class variable (k).

The algorithms for the process of creating new variables were implemented using the Java language version 1.5. We have used the parameter $s = 1.5$ for the IDM model for the reasons stated in Abellán *et al.* (2006b). The $\mathbf{u}$ value of 0.001 has been taken so that the number of variables to be joined into a new variable is not higher than 4. There is in fact only one database (Tic-tac-toe) which joins 4 variables and in all the remaining ones we have unions of only 2 attribute variables.

In order to obtain the results we have used weka software. The following methods have been applied on the databases: naive Bayes before (NB) and after (NB-NV) the creation of the new attribute variables; AODE, also available at weka and the J48 method (J48). In order to evaluate the predictive performance of the classifiers, we have used the experimental scheme of $k - 10$ folds cross validation for all the databases except for the artificial database Monks1. For this database, we have used (as usually Monks1 is used) a fixed little training set (20% of the total database) and the remaining for the test set. The results can be seen in table 2.

Observing the displayed table 2, we can say that for these databases the application of the procedure of new variables (NB-NV) allow us to improve notably the naive Bayes classifier results. This difference is more important seeing the results obtained with AODE method.

For those databases where J48 is notably superior to the naive Bayes classifier, we can see that the results obtained by NB-NV are generally better than those of the J48 method, with the average obtained with the NB-NV method being greater than that of the J48 method, mainly as a result of the value obtained with the Monks1 database.

[¶]J48 method can be obtained via weka software, available in http://www.cs.waikato.ac.nz/ml/weka/

Table 2. Percentage of good results with $k - 10$ folds cross validation on the databases.

Databases	*NB*	*NB-NV*	*AODE*	*J48*
Car	85.6	90.0	91.6	92.2
Monks1	72.4	100	82.0	81.6
Tic-tac-toe	69.8	82.8	74.0	86.4
Corral	86.9	100	90.6	100
Average	78.7	93.1	84.2	88.4

5. Conclusions and future work

In this paper, we have presented a new method which combines variables in the form of a Cartesian product as a prior step to the application of the naive Bayes classifier. The main idea behind our method is to obtain a new set of attribute variables by combining subsets of variables so that the new attribute variables are independent given the class variable, although there may be dependences in each subset. With this procedure, we obtain better results with the application of the naive Bayes classifier. We have used a concept of independence based on uncertainty/information functions: two attribute variables are independent given the class variable if they provide more information about the class variable individually than jointly.

In this method, combination is carried out using the IDM model to represent the information that an attribute variable provides about the class variable from a database, thereby obtaining a credal set, really a set of probability intervals that represents a belief function. On this set, we applied a well-established total uncertainty function such as Shannon's maximum of entropy for credal sets which is simple to calculate due to the existence of fast algorithms for calculating it on the type of probability interval obtained with the IDM model. The result is an information gain function which enables us to join the variables that are more informative jointly than individually.

One important characteristic of the proposed method is that the difference between the joint information and the sum of the individual pieces of information can be limited in order to prevent unions being made between an excessive number of attribute variables. In our experimental work, we have seen that with the established limitation, there are only 1 union of 4 attribute variable, with the remainder being unions of 2 attribute variables. A few unions have been sufficient in order to obtain good results with respect to one of the best known semi-naive model (AODE) when we use databases where naive Bayes classifier is in disadvantage with respect to others classification methods motivated by the relations between the attribute variables.

Analyzing this first work on the application of credal sets and uncertainty measures for the improvement of the naive Bayes classifier, we considered that perhaps this model needs a greater experimental study of the used parameters in order to obtain an improvement. This will be one of our future lines of research.

In the line of this future work, we shall also consider the use of credal sets and information measures to complete data preprocessing prior to the application of a classification method that can help improve the results of this method. This preprocessing shall include operations such as the elimination of attribute variables with little or no information about the class variable, union of the attribute variables (such as in this article) and grouping of equally

informative cases of each attribute variable. All of this work is encapsulated within what Kononenko and Zupan (1999) have called attribute mining.

Acknowledgements

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

References

J. Abellán, "Uncertainty measures on probability intervals from imprecise Dirichlet model", To appear in *Int. J. Gen. Syst.*, 35(3), 2006.

J. Abellán and S. Moral, "A non-specificity measure for convex sets of probability distributions", *Int. J. Unc. Fuzz. Knowl. Based Syst.*, 8, pp. 357–367, 2000.

J. Abellán and S. Moral, "Maximum of entropy for credal sets", *Int. J. Unc. Fuzz. Knowl. Based Syst.*, 11(5), pp. 587–597, 2003.

J. Abellán and S. Moral, "Maximum difference of entropies as a non-specificty measure for credal sets", *Int. J. Gen. Syst.*, 34(3), pp. 201–214, 2005a.

J. Abellán and S. Moral, "Upper entropy of credal sets. Applications to credal classification", *Int. J. Approx. Reasoning*, 39, pp. 235–255, 2005b.

J. Abellán and S. Moral, "An algorithm that computes the upper entropy for order-2 capacities", *Int. J. Unc. Fuzz. Knowl. Based Syst.*, 14(2), pp. 141–154, 2006.

J. Abellán, G.J. Klir and S. Moral, "Disaggregated total uncertainty measure for credal sets", *Int. J. Gen. Syst.*, 35(1), pp. 29–44, 2006a.

J. Abellán, S. Moral, M. Gómez and A. Masegosa, "Varying parameter in classification based on imprecise probabilities" Accepted in *Third international conference on Soft Methods in Probability and Statistics 2006*, 2006b.

J.M. Bernard, "An introduction to the imprecise Dirichlet model for multinomial data", *Int. J. Approx. Reasoning*, 39, pp. 123–150, 2005.

L.M. Campos, J.F de Huete and S. Moral, "Probability intervals: a tool for uncertainty reasoning", *Int. J. Unc. Fuzz. Knowl. Based Syst.*, 2, pp. 167–196, 1994.

A.P. Dempster, "Upper and lower probabilities induced by a multivaluated mapping", *Ann. Math. Statist.*, 38, pp. 325–339, 1967.

R.O. Duda and P.E. Hart, *Pattern Classification And Scene Analysis*, New York: John Wiley and Sons, 1973.

N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian networks classifiers", *Mach. Learn.*, 29, pp. 131–163, 1997.

K. Huang, I. King and M. Lyu, "Learning maximum likelihood semi-naive Bayesian network classifier". *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Hammamet, Tunisia, Vol. 3. 2002.

E.J. Keogh and M.J. Pazzani, "Learning augmented Bayesian classifiers: a comparizon of distribution-based and classification-based approaches", *Proc. Int. Workshop on Artificial Intelligence and Statistics*, 1999, pp. 225–230.

J. Kittler, "Feature selection and extraction", in *Handbook of Pattern Recognition and Image Processing*, T.Y. Young and K.S. Fu, Eds., New York: Academic Press, 1986.

G.J. Klir, *Uncertainty and Information: Foundations of Generalize Information theory*, Hoboken, NJ: John Wiley, 2006.

G.J. Klir and R.M. Smith, "On measuring uncertainty and uncertainty-based information: recent developments", *Ann. Math. Art. Intell.*, 32, pp. 5–33, 2001.

G.J. Klir and M.J. Wierman, *Uncertainty-Based Information*, Heidelberg, Germany: Phisica-Verlag, 1998.

I. Kononenko, "Semi-naive Bayesian classifier", *Proceedings of the 6th European Working Session on Machine Learning*, Berlin: Springer-Verlag, 1991, pp. 206–219.

I. Kononenko and B. Zupan, "Attribute mining: evaluation, discretization, subset selection and constructive induction", *Proceedings of the ICML-99 Workshop on From Machine Learning to Knowledge Discovery in Databases*, Bled, Slovenia, 1999, pp. 1–15.

P. Langley and S. Sage, "Induction of selective Bayesian classifiers", *Proceedings of the Tenth Conference Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann, 1994, pp. 399–406.

M.J. Pazzani, "Constructive induction of Cartesian product attribute", *Inform. Stat. Induction Sci.*, pp. 66–77, 1996.

J.R. Quinlan, "Induction of decision trees", *Mach. Learn.*, 1, pp. 81–106, 1986.

J.R. Quinlan, *Programs for Machine Learning*, San Francisco: Morgan Kaufmann, 1993, series in Machine Learning.

G. Shafer, *A Mathematical Theory of Evidence*, Princeton: Princeton University Press, 1976.

C.E. Shannon, "A mathematical theory of communication", *Bell Syst. Tech. J.*, 27, pp. 379–423, 623–656 1948.
P. Walley, *Statistical Reasoning with Imprecise Probabilities*, New York: Chapman and Hall, 1991.
P. Walley, "Inferences from multinomial data: learning about a bag of marbles", *J. Roy. Statist. Soc. B*, 58, pp. 3–57, 1996.
G.I. Webb, J. Boughton and Z. Wang, "Not so naive Bayes: aggregating one-dependence estimators", *Mach. Learn.*, 58, pp. 5–24, 2005.
Z. Zheng and G.I. Webb, "Lazy learning of Bayesian rules", *Mach. Learn.*, 41, pp. 53–84, 2000.
Z. Zheng and G.I. Webb, "A comparative study of semi-naive Bayes methods in classification learning", *Proc. Fourth Australasian Conference on Knowledge Discovery and Data Mining*, 2005, pp. 141–156.