1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

A Credal Approach to Naive Classification

Marco Zaffalon
IDSIA - Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Corso Elvezia 36, 6900 Lugano Switzerland
zaffalon@idsia.ch

Abstract

Convex sets of probability distributions are also called credal sets. They generalize probability theory with special regard to the relaxation of the precision requirement about the probability values. Classification, i.e., assigning *class* labels to instances described by a set of *attributes*, is a typical domain of application of Bayesian methods, where the naive Bayesian classifier is considered among the best tools. This paper explores the classification model obtained when the naive Bayesian classifier is extended to credal sets. Exact and effective solution procedures for classification are derived, and the related dominance criteria are discussed. A methodology to induce the classifier from data is proposed.

Keywords. Imprecise probabilities, credal sets, classication, naive Bayesian classification, Bayesian networks.

1 Introduction

The relationship between polytopic credal sets (credal sets, for short) of discrete distributions, Bayesian networks (BNs) and classification is studied in this paper. Credal sets [13] are polytopes of distributions and can also be seen as the convex hull of a finite number of distributions. Credal sets provide a framework to formalize uncertain domains in a more flexible way as compared to probability theory. In fact, although credal sets keep their roots in the axioms of probability, their contemporary treatment of many distributions allows the precision requirement of probability theory to be relaxed. The precision requirement is related to the need of precisely defining the probability values in a probabilistic model. Very often, for a number of reasons (ignorance about the phenomenon, economic or temporal constraints, etc.), the latter is not a realistic assumption. Therefore, many authors consider precision the most critical point for the application of probability theory.

An application of probability theory is the task of classification. Classification is a typical machine learning task, where *class* labels must be assigned to instances described by a set of *attributes*. For example, when the problem is the recognition of hand-written characters, the class labels are the alphabetical symbols, while the attributes can be the boolean variables corresponding to the pixels of a b/w digital image. In this case, doing classification is equivalent to output the character that *best* matches the image described by the joint state of the pixels. In general, the classification task is of basic importance in the fields of data analysis and pattern recognition. Bayesian methods are very effective in the field of classification. In particular, it is recognized that the *naive Bayesian classifier* (NBC, section 2.2) is competitive with the state-of-the-art classifiers [4, 6]. The naive Bayesian classifier is a particular case of Bayesian network. A Bayesian network is a graphical model for the decomposition of a joint probability distribution, which is currently the leading model for probabilistic reasoning [10]. Classifiers based on the more general paradigm of Bayesian networks have also been proposed [7].

At the same time, Bayesian networks have been investigated in relationship with credal sets [2, 3, 5]. In this case, the graphical model is extended to treat *sets* of distributions instead of a single joint distribution and can be called *credal net*. Credal nets realize the bridge between credal sets and probabilistic reasoning in the form of Bayesian networks.

This paper analyzes the task of classification when it is realized by a naive Bayesian classifier extended to credal sets. To this purpose, first, the main concepts are introduced in section 2; here a survey of credal sets, Bayesian networks and their relationship is given. Furthermore, classification is described with regard to its implementation with a naive Bayesian classifier. Then, the extension of naive Bayesian classification to credal sets is developed (section 3) and two algorithms are derived; the first algorithm (section 3.1) computes posterior probability intervals for the classification variable; such intervals allow the states of the classification variable to be (partially) ranked through the stochastic dominance criterion (section 3.3). However, since not all dominances are captured by stochastic dominance when credal sets are used (section 3.4), the more general criterion of *credal dominance* is proposed. The second algorithm allows credal dominance to be tested. Section 4 describes an example of naive credal classification. Section 5 shows that the proposed method for naive classifiers can be directly extended to the more general Bayesian networks classifiers, if the patterns to classify do not present missing values. Finally, an inducer for naive credal classifiers (NCCs) is proposed in section 6, where also the results of some experiments on real datasets are reported.

2 Basic Concepts

This section briefly introduces the main concepts discussed in the paper, that is, credal sets, Bayesian networks and the NCC.

2.1 Credal Sets and Bayesian Networks

In the present paper, the term *credal set* is used for a closed and bounded geometric region described by linear constraints, i.e., a polytope, where every point is a probability distribution. Credal set theory generalizes probability to uncertainty about the distribution: more precisely, it assumes that probability is a proper way of dealing with uncertain domains, but criticizes the requirement of precision in the probability values required by probability theory. This brings to the more general view of sets of distributions. The coherence with the axioms of probability is maintained for every distribution in the set; however, the *joint* behavior of the credal set determines many new characteristics of the theory, such that credal sets cannot be seen simply as an extension of classical (point) probability.

As far as graphical models are concerned, Bayesian networks are the current leading model for probabilistic reasoning [10]. A Bayesian network is a couple $\langle G, P\rangle$ where $G = \langle N, A\rangle$ is a directed acyclic graph, whose nodes (N) are interpreted as variables and whose arcs (A) express the direct dependences between them. Each variable (node) $X \in \Omega_X$ has a conditional distribution, $P[X\,|Pa(X)]$, for every possible state $Pa(X)$ of its direct predecessor nodes (parents). It is possible to show that the joint distribution over the variables of the graph is obtained by multiplying the conditional distributions of the nodes: $P[X_1,\ldots,X_n] = \prod_{i\in N} P[X_i\,|Pa(X_i)]$.

Credal sets can be put over Bayesian networks by allowing a credal set to be used in the place of each conditional distribution of the net. That is, a credal network is a couple $\langle G, \wp\rangle$, where G is like above with the difference that each node X now maintains as many (credal) sets $\wp_X^{Pa(X)}$ of conditional distributions as many joint states $Pa(X)$ of the nodes' parents exist. $\wp$ is the set of all the possible joint distributions P over the variables of the net, which is obtained by making any possible choice of the conditional distributions in the credal sets local to the nodes,

$$\wp = \left\{ \begin{array}{c} P[X_1,\ldots,X_n] = \prod_{i\in N} P[X_i\,|Pa(X_i)], \\ s.t.\ P[X_j\,|Pa(X_j)] \in \wp_{X_j}^{Pa(X_j)}, j = 1\ldots n \end{array} \right\}.$$

Remark 1 *Notice that there are other possible ways of extending Bayesian networks to sets of distributions. For example, instead of defining a polytope for each joint state of a node's parents, it would be possible to use a single, higher-dimensional, polytope for each node. In particular, consider the naive classifier; according to the first definition of credal nets, node* A_i *has a collection of* $|\Omega_C|$ *credal sets:* $\wp_{A_i}^C$, $C \in \Omega_C$. *For a given state* c *of* C, *the generic vector in the credal set is* $[P[A_i\,|c]]_{A_i\in\Omega_{A_i}}$. *Instead, with the second definition of credal nets, there would be a single set, where the generic element would be a vector like* $[P[A_i\,|C]]_{A_i\in\Omega_{A_i}, C\in\Omega_C}$. *The latter definition is more general, since it includes the former case and gives the further chance of using constraints between different conditional distributions of the same node. However, the effects of such a greater generality are not clear and, moreover, the implications on computational complexity are strong. This is discussed in section 3.2.3. For the above reasons, this paper adopts the initial definition of credal nets.*

Making inference in a credal net is similar to the Bayesian net case, that is, computing the posterior probability of a variable given a set of instantiated nodes, called the *evidence* nodes. More precisely, if E is the set of evidence nodes, such that $E = e$ is their known state, the inference (or updating) is defined as the computation of $P[X\,|E = e]$, for a node X in the graph. But notice that whereas in a Bayesian network $P[X\,|E = e]$ is a number, it is an interval for a credal net. In fact, it is: $\underline{P}[X\,|E = e] \le P[X\,|E = e] \le \overline{P}[X\,|E = e]$, where the extremes of the interval are the solutions respectively to the minimization and maximization problems below,

$$\underline{P}[X\,|E = e] = \min_{P[X_1,\ldots,X_n]\in\wp} P[X\,|E = e], \quad (1)$$

$$\overline{P}[X\,|E = e] = \max_{P[X_1,\ldots,X_n]\in\wp} P[X\,|E = e]. \quad (2)$$

The interval expresses the ignorance about the posterior probability that logically follows from the ignorance about the joint distribution that is part of the credal net model. Exactly solving problems of type (1) and (2) is a difficult task in the general case. The only known algorithm [2] has an exponential worst case complexity also for singly-connected nets. Restricting the attention to a subset of credal nets can help to develop more effective algorithms; for example, singly-connected credal nets with binary variables admit a solution algorithm which is linear to the size of the graph [5]. The present paper emphasizes that classification is a field that constitutes another example in which an efficient inference algorithm can be realized. This is due to the particular type of computation to be realized in the network.

With special regard to solution algorithms, a basic property of credal networks is that problems (1), (2) can be transformed to combinatorial optimization problems. In other words, it is possible to compute the minimum and the maximum by examining a *finite* number of points. In particular, for each credal set in the network, it is sufficient to consider the (finite) subset of extreme points. The extreme points correspond to the vertices of the polytope represented by the credal set. For this reason, the distributions corresponding to vertices of the polytopes are called *extreme distributions*.

2.2 Naive Bayesian Classification

The naive Bayesian classifier is the particular case of Bayesian network depicted in figure 1.

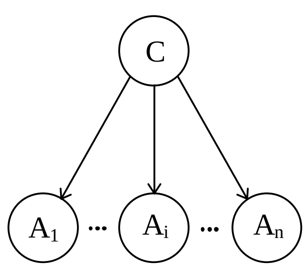

Figure 1: The Naive Bayesian Classifier

The naive classifier has a single root node (C) which corresponds to the classification variable and its children nodes are the attribute variables. Each node X in the network has the usual conditional distributions $P\left[X\left|Pa\left(X\right)\right.\right]$ of the node given the state of the parent(s), $Pa\left(X\right)$, which are estimated from data or from other types of knowledge. Given a particular instance, $A_1 = a_1, \ldots, A_n = a_n$, of the attribute variables, the classification is made by computing the posterior probability $P\left[C\left|a_1, \ldots, a_n\right.\right]$ for each value of C and by picking up the value with maximum probability, $c^* = \arg\max_C P\left[C\left|a_1, \ldots, a_n\right.\right]$.

The naive classifier is quite a simple model and this is due to strong independency assumptions; the attributes must be conditionally independent given the state of the classification variable (this corresponds to the absence of an arc between any couple of attributes). This requirement is met very rarely in applications. However, such a classifier is used much, since it is empirically well-known — and some recent theoretical insights tend to justify such evidence [4, 6] — that the naive Bayesian classifier predictive performance is competitive with the state-of-the-art classifiers (literature also tends to relax the above independence requirements, by allowing the dependences between the attribute variables to be stated using the more general framework of Bayesian networks [7]).

3 Naive Credal Classification

As with any other BN, credal sets can be applied to the naive classifier by substituting the conditional distributions $P\left[C\right], P\left[A_i\left|C\right.\right]$ $(i = 1, \ldots, n, C \in \Omega_C)$ with the credal sets $\wp_C, \wp_{A_i}^C$ $(i = 1, \ldots, n, C \in \Omega_C)$. Then, the focus is on the way the classification should be realized. In the sequel, first, it is analyzed the straight extension of the point probability procedure to credal sets. This implies to compute $P\left[C\left|A_1, \ldots, A_n\right.\right]$ $\forall C \in \Omega_C$ (section 3.1) and to compare them (section 3.3). Recall that the above posterior probabilities are generally intervals; therefore their comparison is addressed with the stochastic dominance criterion. The latter is discussed in section 3.4 where it is shown that it cannot identify all the dominances expressed by credal sets. For this reason, a more general criterion, called *credal dominance*, is proposed. At the light of credal dominance, an alternative procedure for classification is derived in section 3.4.

3.1 Interval Computation

Suppose that the current instance of the attribute variables is $A_1 = a_1, \ldots, A_n = a_n$. As explained in section 2.1, two problems must be solved:

$$\min_{P[C, A_1, \ldots, A_n] \in \wp} P\left[C\left|A_1 = a_1, \ldots, A_n = a_n\right.\right], \quad (3)$$

$$\max_{P[C, A_1, \ldots, A_n] \in \wp} P\left[C\left|A_1 = a_1, \ldots, A_n = a_n\right.\right]. \quad (4)$$

Consider the minimization, the maximization is analogous. Furthermore, consider the computation of the posterior probability for a fixed value c of C. The objective function can be rewritten applying the definition of conditional probability and using the marginalization as follows,

$$P[c|a_1,\ldots,a_n] = \frac{P[c,a_1,\ldots,a_n]}{\sum_{C\in\Omega_C} P[C,a_1,\ldots,a_n]} = \tag{5}$$

$$\left(\frac{\sum_{C\in\Omega_C} P[C,a_1,\ldots,a_n]}{P[c,a_1,\ldots,a_n]}\right)^{-1} = \tag{6}$$

$$\left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C,a_1,\ldots,a_n]}{P[c,a_1,\ldots,a_n]}\right)^{-1}, \tag{7}$$

where it is assumed $P[c,a_1,\ldots,a_n] \neq 0$ for the passage from Eq. (5) to Eq. (6); notice that if $P[c,a_1,\ldots,a_n] = 0$, it is clear from Eq. (5) that $P[c|a_1,\ldots,a_n] = 0$ (the case when also $P[a_1,\ldots,a_n] = 0$ is not considered, because it would imply an undefined conditional probability $P[c|a_1,\ldots,a_n]$).

Finally, Eq. (7) is written according to the topology of the graph,

$$P[c|a_1,\ldots,a_n] = \left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C]\prod_{i=1}^n P[a_i|C]}{P[c]\prod_{i=1}^n P[a_i|c]}\right)^{-1}. \tag{8}$$

Now, the minimization problem (3) is written by substituting its objective function with the right member of Eq. (8),

$$\min_{P[C]\in\wp_C} \min_{P[A_i|C]\in\wp^C_{A_i}, C\in\Omega_C, i=1,\ldots,n} \left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C]\prod_{i=1}^n P[a_i|C]}{P[c]\prod_{i=1}^n P[a_i|c]}\right)^{-1}. \tag{9}$$

Notice that since the objective is now expressed by means of the local conditional distributions, also the minimization is taken over the possible conditional distributions in the credal sets local to the nodes. Let us now focus on the inner minimization problem only: the goal is the *maximization* of the fractional function inside the parentheses, since this is equivalent to minimizing the reciprocal. First of all, notice that it is possible to minimize the denominator and to maximize the numerator separately, in fact they do not share any term. Second, notice that the quantities $P[a_i|C]$ $(C\in\Omega_C, i=1,\ldots,n)$ are independent one another[1] because they are defined by means of disjoint sets of constraints. This means that the choice of a conditional distribution $P[a_i|C]$ in $\wp^C_{A_i}$ can be made without taking into account the choice of any other conditional distribution. The observations above allow the inner optimization to be solved. Consider the denominator case. $P[c]$ is a non-negative number; therefore, the denominator is minimized when the product $\prod_{i=1}^n P[a_i|c]$ is minimized. This is obtained by setting any $P[a_i|c]$ to its minimum, i.e. $\prod_{i=1}^n \underline{P}[a_i|c]$. An analogous argument holds for the numerator; $P[C]$ is non-negative $\forall C\in\Omega_C$, and the sum is made by terms that can be optimized separately. Hence, the numerator is maximized when the product of the conditional distributions is set to $\prod_{i=1}^n \overline{P}[a_i|C]$ $\forall C\in\Omega_{C\setminus\{c\}}$.

Problem (9) becomes,

$$\min_{P[C]\in\wp_C} \left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C]\prod_{i=1}^n \overline{P}[a_i|C]}{P[c]\prod_{i=1}^n \underline{P}[a_i|c]}\right)^{-1}. \tag{10}$$

The next step is the application of the combinatorial property of the problem, cited in section 2.1. For any given node X, denote with $\overline{\wp}^{Pa(X)}_X$ the finite subset of $\wp^{Pa(X)}_X$ made by its extreme points. Problem (10) is equivalent to the combinatorial problem (11),

$$\min_{P[C]\in\overline{\wp}_C} \left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C]\prod_{i=1}^n \overline{P}[a_i|C]}{P[c]\prod_{i=1}^n \underline{P}[a_i|c]}\right)^{-1}, \tag{11}$$

which is simply solved by enumerating the extreme distributions in $\overline{\wp}_C$. Of course, it is also necessary to know the extremes of $P[a_i|C]$ $(C\in\Omega_C, i=1,\ldots,n)$ in order to apply formula (11). This is obtained by solving problems $\min_{P[A_i|c]\in\overline{\wp}^c_{A_i}} P[a_i|c]$ and $\max_{P[A_i|C]\in\overline{\wp}^C_{A_i}} P[a_i|C]$ $\forall C\in\Omega_C\setminus\{c\}, \forall i = 1,\ldots,n$, which can again be done enumerating the extreme distributions in the respective feasible sets.

Following the same argument, it is straightforward to obtain the formula solving problem (4), which is reported below,

$$\max_{P[C]\in\overline{\wp}_C} \left(1+\frac{\sum_{C\in\Omega_C\setminus\{c\}} P[C]\prod_{i=1}^n \underline{P}[a_i|C]}{P[c]\prod_{i=1}^n \overline{P}[a_i|c]}\right)^{-1}. \tag{12}$$

[1]Observe that this is a different concept as compared to probabilistic independence.

3.2 Complexity

The purpose of this section is to analyze the complexity of doing classification with credal nets. In particular, the complexity of the combinatorial solution of problems (11) and (12) is discussed; furthermore, a different exact solution procedure is proposed for the cases when the combinatorial approach is too expensive. Finally, a brief analysis is carried out with regard to the classification complexity with the more general model described in the remark 1.

3.2.1 Combinatorial Procedure

In order to apply formulas (11) and (12), it is necessary, first, to compute the extreme values of the conditional distributions of the model. Suppose that the extreme distributions of the credal sets are already known[2], and that the maximum number of extreme distributions in a set is k. Hence, the combinatorial computation of the extremes of $P\left[a_i\left|C\right.\right]$ takes $O\left(k\right)$ time. This must be repeated $\forall C \in \Omega_C$ and $\forall i = 1,\ldots,n$, yielding to $O\left(nk\left|\Omega_C\right|\right)$.

With regard to formulas (11) and (12), when the values of the conditional probabilities are fixed, the expression in parentheses is computed in time $O\left(\left|\Omega_C\right|\right)$. The latter must be repeated for each extreme distribution in $\overline{\wp}_C$, in order to evaluate the optimum, yielding to $O\left(k\left|\Omega_C\right|\right)$. The total complexity is the sum of the two expressions above, i.e.,

$$O\left(nk\left|\Omega_C\right|\right). \tag{13}$$

Observe that the complexity is linear with every independent quantity making up expression (13), i.e., it is linear to n, k and $\left|\Omega_C\right|$.

3.2.2 Linear Programming Procedure

Among the three quantities in formula (13), n and $\left|\Omega_C\right|$ are directly under the control of the model builder; instead, k can be any integer, depending on the constraints that define the credal sets[3]. If k is not *too* large, the combinatorial approach is efficient; but there are cases when k can be, like when a credal set is defined via interval constraints [1]. In this case, k can be formally bounded on the basis of the size of a variable's domain. For example, if the variable has 10 possible states, k can be 1260 at most; for 11 states, $k \leq 2722$. As the number of states grow, k can be quite large and the combinatorial approach leading to complexity (13) may be significantly slowed down. This should not be perceived as a limitation to the applicability of credal classification. In fact, there is no need to use explicitly the extreme distributions if more sophisticated approaches based on linear programs (LPs) are adopted. The latter directly use the constraints representation of the polytopes and can solve problem (9) exactly and efficiently (the maximization is analogous): first, the computation of an extreme of $P\left[a_i\left|C\right.\right]$, $\forall i = 1,\ldots,n$ and $\forall C \in \Omega_C$, is an LP. A linear program is usually solved efficiently also for very large domains with the simplex method; furthermore it is a task polynomially bounded when an interior point method is used. Second, observe that problem (10) is a fractional linear program, because the values of the conditional probabilities are fixed and only the probabilities of the states of C are variables. Fractional linear problems can be turned to LPs by changing variables ([11], section 2.2.2).

[2] The extreme distributions can be computed given the contraints (and vice versa).

[3] Except for the case when a variable is binary; in this case k can be 2 at most.

3.2.3 About More General Models

In the derivation of the solution procedure, in section 3.1, the passage from Eq. (9) to Eq. (10) is a significant point both for complexity and for the nature of problem itself. In fact, it states that the optimal values of the conditional distributions $P\left[A_i\left|C\right.\right]$ can be fixed to $\overline{P}\left[a_i\left|C\right.\right]$ and $\underline{P}\left[a_i\left|c\right.\right]$, $\forall C \in \Omega_C \setminus \{c\}$ and $\forall i = 1\ldots n$, independently on each other and on the chosen distribution $P\left[C\right] \in \wp_C$. On one hand, this makes the combinatorial computation viable, and on the other, it allows the problem to be described within the framework of fractional linear programming.

Such nice properties are lost, for example, by passing to the more general model described in the remark 1. In fact, in that case, $\forall i = 1\ldots n$, the quantities $P\left[A_i\left|C\right.\right], C \in \Omega_C$, are *not* mutually independent (because there can be constraints between them). Therefore, the argument that justifies the passage from Eq. (9) to Eq. (10) cannot be applied. The first consequence is that a combinatorial approach can still work but with much higher complexity that, in facts, renders the method not viable. The combinatorial approach can be realized by doing all the combinations of the extreme points of the attribute variables' convex sets; but this implies a fast combinatorial explosion of the number of points to treat. The second consequence is a seeming failure of the non-linear approach, too; in fact, since the values $P\left[A_i\left|C\right.\right]$ are not mutually independent, they cannot be fixed, i.e., they are *variables* in problem (9). Hence, the latter cannot be reduced to a fractional linear program, being the ratio of two polynomials over a linear set. This moves the problem to a difficult side of the global optimization field. In the author's knowledge, literature does not present any general effective procedure to globally optimize such problems [11, 12].

3.3 Stochastic Dominance

Section 3.1 provides the means to compute the probability interval where $P[C|a_1,\ldots,a_n]$ lies, $\forall C \in \Omega_C$. The second step of classification is the choice of the value $c^* \in \Omega_C$ that maximizes $P[C|a_1,\ldots,a_n]$. As opposed to ordinary classification, credal classification also faces the problem of doing such a choice when the probability is defined as an interval. Optimality is not always defined when intervals are present. This is related to the observation that the set of intervals $\{P[C|a_1,\ldots,a_n] \mid C \in \Omega_C\}$ cannot always be totally ordered. An objective criterion for comparing probability intervals is the so-called *stochastic dominance*: given two intervals, $I_1 = [a_1, b_1]$ and $I_2 = [a_2, b_2]$, I_2 dominates I_1 iff $a_2 \geq b_1$. The rationale behind the criterion is that since each probability in I_2 is greater or equal to any probability in I_1, the event associated to I_2 is, at least, as probable as the event related to I_1. Unfortunately, stochastic dominance generally implies only a partial order; if I_1 and I_2 overlap, they cannot be compared. Therefore, there does not always exist an interval that dominates all the others, which is necessary for optimality; the only objective output of the system can be the set of the undominated states of C, which in *some* cases can be a singe state, though not in general. In a sense, credal classification can be interpreted as a more cautious procedure as compared to Bayesian classification. In fact, it does not provide a single answer, when there are not the conditions for doing that.

3.4 Credal Dominance

So far, the analysis has focused on the computation of the ignorance intervals for the states of C, with the aim of comparing them using stochastic dominance. It is useful to observe that the potentially available information with credal sets is greater than that provided by intervals; the credal set for $P[C|a_1,\ldots,a_n]$, say $\wp_C^{a_1,\ldots,a_n}$, generally conveys greater knowledge than that given by the separate intervals for the posterior probabilities of C. In fact, the credal set can also represent possible constraints between the above probabilities, which disappear with the interval view. Therefore, it is natural to wonder if a comparison criterion different from stochastic dominance can better exploit the knowledge that is proper of credal sets.

Consider the following example. Suppose that $\Omega_C = \{c_1, c_2, c_3\}$ and that $\wp_C^{a_1,\ldots,a_n}$ has only three extreme distributions: $[.4, .4, .2]$, $[.5, .4, .1]$ and $[.5, .5, 0]$, where the j-th elements of the vectors represent $P[C = c_j|a_1,\ldots,a_n]$. The intervals for the posterior probability of c_1, c_2 and c_3 are, respectively, $[.4, .5]$, $[.4, .5]$ and $[0, .2]$. Then, stochastic dominance allows the state c_3 to be discarded, but not the other two states to be compared, i.e., stochastic dominance produces two undominated states, c_1 and c_2. But it is easy to see that for any distribution in the credal set, it is $P[C = c_1|a_1,\ldots,a_n] \geq P[C = c_2|a_1,\ldots,a_n]$, i.e., c_2 is dominated. In other words, for the credal set at hand, there is only one dominant state, but this fact is hidden when intervals and stochastic dominance are used. Thus, we are lead to the definition of a dominance criterion for credal sets.

Definition 1 *Let X be a discrete variable and x_1, x_2 two states in the domain of X. Consider the distribution $P[X|E] \in \wp_X^E$, where E represents what is known and $\wp_X^E$ is a non-empty set of distributions. The state x_1 is said to be* credal-dominant *as compared to x_2, $x_1 \succeq x_2$, if for any distribution $P[X|E] \in \wp_X^E$, $P[X = x_1|E] \geq P[X = x_2|E]$.*

Credal dominance is a straightforward generalization of stochastic dominance to sets of distributions. Notice that stochastic dominance implies credal dominance, whereas the converse is not true, as the example above shows. That is, not all dominances are discovered by stochastic dominance; hence, the latter can be seen as an approximation to credal dominance. Now, the point is if the computation of credal dominance can be realized in an effective way. Before addressing the naive classification case, it is useful to observe that, in general, credal dominance can be checked in a combinatorial way; when $\wp_X^E$ is a polytope, $x_1 \succeq x_2 \iff P[X = x_1|E] \geq P[X = x_2|E]$ $\forall P[X|E] \in \overline{\wp}_X^E$. In one sense ($\Rightarrow$) the proof is trivial; the other sense ($\Leftarrow$) is easily proved by using the fact that any distribution in $\wp_X^E$ is a convex combination of the extreme distributions.

Let us develop the particular case of the NCC. Consider two states of C, namely c_1 and c_2. We want to check if

$$P[C = c_1|a_1,\ldots,a_n] \geq P[C = c_2|a_1,\ldots,a_n] \quad (14)$$

holds for all the joint distributions in $\wp$. The question is equivalent to solving the following problem,

$$\min_{P[C,A_1,\ldots,A_n]\in\wp} \frac{P[C = c_1|a_1,\ldots,a_n]}{P[C = c_2|a_1,\ldots,a_n]}, \quad (15)$$

where, for the moment, $P[C = c_2|a_1,\ldots,a_n]$ is assumed positive. If the optimum of problem (15) is greater of equal to 1, the answer to the credal dominance question is affirmative; in fact, it means that the inequality (14) is always true. Whenever the optimum is strictly lower than 1, the inequality (14) is false, at least for the joint distribution in $\wp$ where the optimum is attained, and $c_1 \succeq c_2$ is not verified (but it might still be $c_2 \succeq c_1$).

Problem (15) can be solved according to an argument completely analogous to that used for problem (3). In particular, problem (15) is initially rewritten as $\min_{P[C,A_1,\ldots,A_n]\in\wp} P\left[c_1, a_1, \ldots, a_n\right] / P\left[c_2, a_1, \ldots, a_n\right]$, by deleting the probability of the considered instance; then, it is represented according to the topology of the graph and the values of the conditional probabilities are fixed, as follows,

$$\min_{P[C]\in\wp_C} \frac{P\left[c_1\right]\prod_{i=1}^{n} \underline{P}\left[a_i\left|c_1\right.\right]}{P\left[c_2\right]\prod_{i=1}^{n} \overline{P}\left[a_i\left|c_2\right.\right]}. \tag{16}$$

Finally, if the combinatorial approach is chosen, the solution of problem (16) is obtained enumerating the extreme distributions of $\wp_C$, after computing the extremes of the conditional distributions; otherwise, the LP-based solution procedure can be used, by computing the extremes of the conditional distributions by LPs and then solving the remaining fractional linear program (16). Notice that problem (16) allows only $c_1 \succeq c_2$ to be checked; testing $c_2 \succeq c_1$ requires the minimization of the reciprocal objective to be solved.

A simple extension also solves the case when for some distributions in the set $\wp$, $P\left[c_2\right]\prod_{i=1}^{n} \overline{P}\left[a_i\left|c_2\right.\right] = 0$. The latter happens if either $P\left[c_2\right] = 0$ or $\overline{P}\left[a_i\left|c_2\right.\right] = 0$ for some $i = 1, \ldots, n$. Consider the case $P\left[c_2\right] = 0$, the other case is analogous. If $P\left[c_2\right] = 0$ *for all* the distributions $P\left[C\right] \in \wp_C$, $P\left[C = c_2\left|a_1, \ldots, a_n\right.\right]$ is always zero, therefore the inequality (14) always holds; otherwise the distributions corresponding to $P\left[c_2\right] = 0$ can be discarded, because the problem is a minimization.

From the above analysis it also follows that credal dominance between two states (i.e., checking both the directions of the inequality) can be computed even more quickly than stochastic dominance. In fact, two optimization problems must be solved, and the overall time required is lower than for problems (3) and (4) together. Compare, for instance, problem (16) with problem (10); problem (16) is solved more quickly, because it only has $O(n)$ terms at the numerator versus the $O(n\left|\Omega_C\right|)$ terms of problem (10). For symmetry, the same is true for the remaining couple of problems.

Conversely, it seems possible that computing the set of undominated states of C is more expensive using credal dominance than using stochastic dominance. With stochastic dominance, $2\left|\Omega_C\right|$ optimizations like problems (10) allow all the intervals for C to be computed; then, $O(\left|\Omega_C\right|^2)$ interval comparisons select the stochastic-undominated states. With credal dominance, $O(\left|\Omega_C\right|^2)$ optimizations like problem (16) are required, which is clearly more expensive than $O(\left|\Omega_C\right|^2)$ interval comparisons; however, each optimization is simpler than the corresponding one with stochastic dominance, as outlined above. The precise definition of the data structures to use and a more careful organization of the procedures can lead to a deeper comparison of the relative computational weights of the two different approaches. For the moment, let us investigate the complexity of another computation, i.e., how to compute the credal-dominant state of C, if it exists. This is useful when a single dominant state is required. Such task is carried out solving only $O\left(\left|\Omega_C\right|\right)$ instances of problem (16).

The solution is obtained with a procedure in two steps. In the first step, two states of C, c_1 and c_2, are considered; $c_1 \succeq c_2$ is tested by solving problem (16). If $c_1 \succeq c_2$ holds, c_2 is discarded because it cannot be dominant; for the same reason, if $c_1 \succeq c_2$ is false, c_1 is discarded. Therefore the set of states to analyze is decreased of 1 state at least. The procedure is repeated on the new set of states until its cardinality is lower or equal to 1, which happens after $\left|\Omega_C\right|-1$ optimizations. Afterwards, if the set is empty there is no dominant state, for definition; otherwise, if the set contains a single state, the latter has verified the necessary condition to be credal-dominant. The sufficiency (second step) is verified by testing its credal-dominance over all the states of C not already compared with it in the first step; this is obtained with $\left|\Omega_C\right| - 1$ optimizations at most. The overall number of optimizations (step 1 + step 2) is then $O(\left|\Omega_C\right|)$.

4 An Example

This section introduces an easy example to show the principles of credal classification with regard to the solution procedure described in section 3.1. The example is made simple for clarity; the probabilities are arbitrary.

An insurance company wants to assess the risk it incurs about the car insurance for a *new* customer. The risk (R) is classified as *low*, *medium*, *high*, and is related to the number and type of car accidents that must be expected by such a customer. The company decides to infer the risk on the basis of two attributes of the customer: the *age* ($A \in$ {young, middle-aged, old}) and the *city* ($T \in$ {Venezia (VE), Treviso (TV), Milano (MI)}) where the customer lives.

The credal sets of the NCC are defined with the following purpose. About the customer's age, it is supposed that the middle-aged persons have better behavior, as compared both to young and old people; about the risk of the cited Italian cities, the ranking is, Venezia < Treviso < Milano. The credal sets, defined by means of (proper and reacheable, see [1]) interval constraints, for ease of treatment, are reported in tables 1, 2 and 3.

R	Intervals
low	[.77,.85]
medium	[.10,.15]
high	[.05,.08]

Table 1: $P[R]$

	R		
A	low	medium	high
young	[.15,.22]	[.27,.32]	[.60,.70]
middle-aged	[.50,.55]	[.33,.38]	[.05,.15]
old	[.28,.34]	[.34,.38]	[.20,.30]

Table 2: $P[A|R]$

	R		
T	low	medium	high
VE	[.70,.72]	[.15,.20]	[.02,.06]
TV	[.18,.20]	[.60,.65]	[.22,.28]
MI	[.08,.10]	[.20,.25]	[.66,.72]

Table 3: $P[T|R]$

For example, table 1 expresses the fact that the greatest part of customers are known to be low-risk, in a percentage that varies from 77% to 85%; a minority are medium-risk people (in the range 10% - 15%); and that few people (5% - 8%) are high-risk. Table 2 mainly expresses that when the risk is low, the most probable age is middle-age; when it is medium, the three states of A have similar probability, whereas when the risk is high, people are most probably young, otherwise old.

Now, formulas (11) and (12) are applied; they require the extremes of the conditional probabilities $P[A_i|C]$. Such extremes are readily available when using reacheable intervals. The above formulas also require the extreme distributions of $P[C]$ ($P[R]$ in the present example). Such distributions are computed on the basis of the intervals with simple procedures [1].

Let us consider the case of an old person, living in Venezia. Formulas (11) and (12) yield the three probability intervals in Table 4. Table 4 shows that, in this case, stochastic dominance implies a total order on the states of R, and the customer is classified low-risk. A different situation happens when the case of a young person, in Milano, is considered. Intuitively, the subject should be high-risk, because each attribute is in the worst condition. Formulas (11) and (12) produce the intervals in table 5.

R	Intervals
low	[.922,.975]
medium	[.024,.070]
high	[.001,.009]

Table 4: $P[R|A = old, T = VE]$

R	Intervals
low	[.150,.426]
medium	[.085,.290]
high	[.435,.693]

Table 5: $P[R|A = young, T = MI]$

Notice that there in only a partial order between the states of R, however, it is still possible to obtain a single dominant state, i.e., *state high*. The last case is about a young person who leaves in Treviso. This is slightly more difficult to classify, intuitively, since there are opposite tendencies in the attributes: being a young person makes risk higher, but the risk should be lowered down by living in a city with moderate traffic. Applying Eq. (11) and (12), the intervals in table 6 are obtained. This time a single dominant state is not available, however, the high-risk state can be disregarded, because it is dominated by the low-risk one.

R	Intervals
low	[.307,.621]
medium	[.238,.525]
high	[.100,.267]

Table 6: $P[R|A = young, T = TV]$

5 Observations

The first important observation is that the extension of the NBC to credal sets implies the extension of the more general Bayesian networks classifiers, too. A Bayesian network classifier (BNC) [7] is a Bayesian network with nodes $C, A_1, \ldots, A_n$, such that the portion of graph related to the classification variable is like in figure 1, and the part related to nodes $\{A_1, \ldots, A_n\}$ is a generic BN. It is well known that any BN with some evidence nodes (E) is equivalent to the network obtained by removing the arcs departing from E [10]. In a BNC, $E \equiv \{A_1, \ldots, A_n\}$, if the considered instance has not missing data, and the removal of the arcs departing from the attributes turns the model to a NBC. Hence, the classification of an instance on a BNC with credal sets (credal networks classifier, CNC) reduces to the computation on a NCC, which is solved by the procedures in the above sections. The modelling capability obtained in this way seems quite general.

The second observation is about the potential uses of credal classification, which seems to embrace a wide area of different applications. For example, it can be used to do sensitivity analysis; building classifiers on the basis of experts' opinions fits naturally in the credal classification paradigm; credal classifiers can be used when the probabilities are estimated from databases with missing data; in general, ignorance about the domain is allowed to be included in the model, in order not to be forced to improperly reduce to the case of a single distribution.

6 An Inducer for Credal Classifiers

The present section discusses the construction of an *inducer* for the NCC. An inducer is an algorithm that builds a classifier from a dataset [8]; this is a relevant topic, since classification is a typical data mining task. The literature of imprecise probabilities can directly be exploited to this extent: Walley shows how to compute lower and upper probabilities for an event, given a dataset, on the basis of an *imprecise* Dirichlet model [14]. For a generic state x_j of a discrete variable X, the interval is,

$$\left[\frac{n_{x_j}}{N+s}, \frac{n_{x_j}+s}{N+s}\right], \tag{17}$$

where n_{x_j} is the number of occurrences of the state x_j, N is the number of observations and $s > 0$ is a hyperparameter (observe that the interval contains the observed frequency of x_j $\forall s > 0$). The hyperparameter s determines how quickly the lower and upper probabilities converge as more data become available; larger values of s produce more cautious inferences. Walley suggests two candidate values for s, i.e. $s = 1$ or $s = 2$, also if no definite statement about s is claimed.

An inducer was implemented on the basis of the above result, i.e., each NCC model probability was defined as an interval, according to (17). It is worth observing that, whenever the dominant state exists, the classification of the induced NCC agrees with the NBC's. To see this, consider that: the interval (17) always contains the observed frequency (i.e., the probability used by the NBC), therefore each credal set of the NCC contains the related distribution used by the NBC, and hence the NBC posterior distribution for C is contained in the NCC posterior credal set of C; when a state is credal-dominant, it is dominant for all the distributions in the posterior credal set of C.

For the above observation, the difference between the induced NCC and NBC lies in the patterns that the NCC cannot uniquely classify. In the experiments, the induced NCC was tested on the "breast", "corral" and "german" datasets from UCI repository [9]. All three datasets have 2 levels of the classification variable; therefore for a given pattern of the attributes there are only two cases, either a single dominant state exists or no dominant state exists. The chosen policy about the latter cases was to reject them. Using such a procedure, the comparison of the NCC and the NBC must be made on the basis of the rejected patterns. It seems reasonable to require that the NBC prediction performance on the rejected patterns is 50% or lower (recall that there are only 2 levels for C), in order to judge the NCC superior over the NBC.

Experimental methodology. The datasets were preprocessed by removing the observations with missing values, and by discretizing them, when needed. Then, in order to evaluate the predictive performance of the classifier, each dataset was randomly split into a training set for supervised learning and a test set. This procedure was repeated a number of times to improve the accuracy of the estimates. More precisely, the experimental scheme was the so-called k-folds cross validation [8] with $k = 5$; k-folds cross validation was repeated up to obtaining that the standard deviations of the sample means were lower than 1/3.

		s					
		1	2	1	2	1	2
Dataset	NBC	NCC	NCC	R	R	N(R)	N(R)
Breast	96.65	97.83	97.89	1.65	1.91	25.80	32.74
Corral	86.41	88.85	90.39	6.98	14.11	53.87	62.78
German	75.21	76.57	77.82	5.94	11.60	53.62	55.11

Table 7: The measured experimental percentages

Table 7 shows the percentages of the measured quantities. Column NBC reports the predictive performance of the NBC; column NCC displays the predictive performance of the NCC on the set of unrejected patterns; the percentage of rejected patterns is in column R; finally, N(R) is the performance of the NBC on the rejected set. All the experiments were made both for $s = 1$ and $s = 2$.

Discussion. When C has 2 levels, predicting at 50% is equivalent to guessing. Table 7 shows that the choice $s = 1$ makes N(R) to be about 50% or lower for all the datasets. This means that Walley's intervals allowed the NCC to reject patterns about which the NBC just guesses or is wrong most of times. Clearly, this also allowed the NCC performances to be greater than the corresponding NBC ones. Using $s = 2$, the NCC again improves its performance, but, discarding more patterns (being more cautious), can reject also patterns on which the NBC performs significantly (a little) better than 50%, as in the "corral" case. The behavior of the NCC for $s = 1$ seems quite appealing; if larger experimental studies confirmed it, useless or bad prediction performances might be avoided.

We can do two final remarks. First, the behavior of credal classification should be also more evident and useful when CNCs were used; in fact, given a dataset, the CNCs model probabilities are more variables as compared to the NCC case (hence Walley's intervals would be larger). Second, experimental analyses might profit from using artificial databases generated according to a known BNC; this would allow the effect of partial knowledge of the probability values to be distinguished from the effects of the partial knowledge about the dependency structure.

7 Conclusions

This paper proposes an approach to classification based on the naive Bayesian classifier and on credal sets. The application of credal sets to naive classification is simple to realize and, moreover, allows uncertainty about the probability values to be included in the model. The approach seems a proper way of doing classification, which preserves the advantages of the classic Bayesian approach, while adding flexibility. The paper presents the basic tools to classify with the NCC. Future research lines might address the topics that the gained flexibility allows to be tackled, like the combination of imprecise subjective and objective knowledge, the treatment of missing data, robustness analysis, etc., in order to develop an integrated platform for credal classification.

Acknowledgements

The author is grateful to Peter Walley and to the anonymous reviewers for their useful comments on the first version of the paper.

References

[1] L. Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.

[2] J.E. Cano, S. Moral, and J.F. Verdegay-Lòpez. Propagation of convex sets of probabilities in directed acyclic networks. In B. Bouchon-Meunier, L. Valverde, and R.R. Yager, editors, *Uncertainty in Intelligent Systems*, pages 85–96. Amsterdam: Elsevier, 1993.

[3] F. Cozman. Robustness analysis of bayesian networks with local convex sets of distributions. In D. Geiger and P.P. Shenoy, editors, *UAI-97*, pages 108–115. San Francisco: Morgan Kaufmann, 1997.

[4] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, 1997.

[5] E. Fagiuoli and M. Zaffalon. 2u: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[6] J. Friedman. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[7] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29(2/3):131–163, 1997.

[8] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI-95*, pages 1137–1143. San Mateo: Morgan Kaufmann, 1995.

[9] P.M. Murphy and D.W. Aha. Uci repository of machine learning databases, 1995. http://www.sgi.com/Technology/mlc/db/.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann, 1988.

[11] S. Schaible. Fractional programming. In R. Horst and P.M. Pardalos, editors, *Handbook of Global Optimization*, pages 495–608. The Netherlands: Kluwer, 1995.

[12] S.A. Vavasis. Complexity issues in global optimization: a survey. In R. Horst and P.M. Pardalos, editors, *Handbook of Global Optimization*, pages 27–41. The Netherlands: Kluwer, 1995.

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

[14] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *J.R. Statist. Soc. B*, 58(1):3–57, 1996.