The Naive Credal Classifier

Marco Zaffalon
IDSIA—Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Galleria 2, CH-6928 Manno, Switzerland
zaffalon@idsia.ch

Abstract

Convex sets of probability distributions are also called credal sets. They generalize probability theory by relaxing the requirement that probability values be precise. Classification, i.e. assigning *class* labels to instances described by a set of *attributes*, is an important domain of application of Bayesian methods, where the naive Bayes classifier has a surprisingly good performance. This paper proposes a new method of classification which involves extending the naive Bayes classifier to credal sets. Exact and effective solution procedures for naive credal classification are derived, and the related dominance criteria are discussed. Credal classification appears as a new method, based on more realistic assumptions and in the direction of more reliable inferences.

AMS Subject Classification: Primary 62H30, 68T10; secondary: 68T37, 90C05.

Keywords: Credal Sets; Classification; Pattern Recognition; Naive Bayes Classifier; Imprecise Probabilities

1 Introduction

Classification (also known as pattern recognition, identification, or selection) is a multivariate technique concerned with allocating new objects to previously defined groups on the basis of observations on several characteristics of the objects; see Giri (1996), Huberty (1994) and McLachlan (1992). Formally, a classifier is a function that maps instances of a set of variables, called attributes or features, to a state of a categorical class variable. Such a paradigm is very general; problems in very different fields can be represented as classification problems. For instance, the recognition of hand-written characters or faces, and the problem of diagnosing a disease from symptoms, are classification problems. For that reason, classification is important in many fields of research. Publications concerning classification can be found in the statistical literature (e.g., in the *Journal of Classification*) as well as in the literature of artificial intelligence (Fayyad et al., 1996).

Learning a classifier from a random sample is an important type of statistical inference. After observing the attribute values and category (class) for each unit in the sample, the problem is to define an association rule that predicts the correct category from future instances of the attributes. An instance of the attributes is also referred to as *pattern*.

Formally, let use denote the classification variable by C, taking values in the finite set $\mathcal{C}$, where the possible classes are denoted by lower-case letters. We measure n features $A_1, \ldots, A_n$ taking generic values $a_1, \ldots, a_n$ from the sets $\mathcal{A}_1, \ldots, \mathcal{A}_n$, which are assumed to be finite. We can define a classifier by means of the discrete joint probability distribution $P[C, A_1, \ldots, A_n]$. The classification of a new pattern

$(a_1, \ldots, a_n)$ is realized by selecting a class $c \in \mathcal{C}$ that maximizes $P[c\,|a_1, \ldots, a_n]$. When misclassification costs are equal, this classification rule is optimal in the sense that it minimizes the expected cost of misclassification (Johnson and Wichern, 1988).

Unfortunately, without further assumptions, this theoretical approach is not effective in practice. The number of probabilities which define the joint distribution grows exponentially with the number of attributes. Thus their estimates from the sample are generally poor and so is the prediction accuracy of the classifier.

Duda and Hart (1973) have proposed to assume independence of the attributes conditional on the class,

$$P[A_1, \ldots, A_n\,|C] = \prod_{i=1}^{n} P[A_i\,|C]\,. \tag{1.1}$$

The resulting classification model is referred to as the *naive Bayes classifier* (NBC). Now the joint distribution $P[C, A_1, \ldots, A_n]$ requires the specification of a smaller number of probabilities, which can be estimated from the sample in a more robust way. Notice that these probabilities are generally biased because assumption (1.1) is strong and is not likely to hold in many domains. However, it is well known that assumption (1.1) is *not critical for classification*, as shown by Domingos and Pazzani (1997) and by Friedman (1997): the NBC is very often accurate even when assumption (1.1) is violated substantially. In fact, the NBC is competitive with the state-of-the-art classifiers, while often being faster and simpler; and, despite its relatively long history in pattern recognition, the NBC is still the subject of active research.

This paper is concerned with extending the NBC to a new and more general method of classification which we call *credal classification*. A credal classifier is a function that maps instances of attributes to a *set* of categories. A standard classifier is a credal classifier that always outputs singletons.

The definition of credal classification is intimately related to the passage from

standard probability theory to *imprecise probabilities* (Walley, 1991). Relaxing the requirement of a unique output class is analogous to relaxing the requirement of a single probability measure, in favor of a set of distributions or *credal set* (Levi, 1980). A credal set is defined to be the convex hull of a non-empty and finite family of probability measures. Credal sets have great expressive power, encompassing a number of other models for uncertainty (e.g., possibility measures, belief functions, Choquet capacities, coherent lower probabilities), and they are equivalent to *coherent lower previsions* as defined in Walley (1991). The axioms of probability are maintained for every distribution in the set; however, the *joint* behavior of the credal set determines many new characteristics of the theory, so that credal sets cannot be seen simply as an extension of classical probability.

This paper realizes a credal classifier by extending the NBC to credal sets, thus defining the *naive credal classifier* (NCC). The NCC enables imprecision to be taken into account, as possibly generated by unobserved or rare events, small sample sizes and missing data; and it enables this to be done *efficiently*. As a consequence, for a given pattern of the attributes, imprecision in the input may prevent a single output class from being obtained; then the result of the NCC classification is a set of classes, all of which are candidates to be the correct category. In other words, the NCC recognizes that the available knowledge may not suffice to isolate a single class and thus gives rise to a set of alternatives. This seems a very natural process of learning under partial information, but it is currently denied to classifiers that are always required to satisfy the chimera of precision, and to output a single class even when there is very little information.

We define the naive credal classifier in Section 2, where we discuss the model and introduce some terminology. Then we present two procedures for classification. Section 3 introduces the first procedure as an extension of the NBC classification procedure. This is based on the posterior probability intervals for the class and on a criterion of *interval dominance* to select the set of output classes. Section 3.1

formalizes the optimization problems that define the posterior probability intervals. Section 3.2 provides two alternative algorithms to solve such problems and analyzes their computational complexity.

The second procedure for classification is based on a strengthening of the criterion of interval dominance. We discuss the motivations to refine interval dominance and define the new criterion of *credal dominance* in Section 4. This section also provides a classification procedure based on credal dominance and analyzes its complexity. The procedure is specialized to the credal sets produced by probability intervals in Section 4.1. Next, in Section 5 we present an example of naive credal classification, and in Section 6 we provide a method to compute the posterior probability of the proposed set of classes.

Finally, Section 7 summarizes the results of the paper and highlights the issues to be addressed in future research, including statistical inference of the NCC from a random sample. We present motivations supporting the choice of the *imprecise Dirichlet model* in Walley (1996b) for this purpose.

2 The naive credal classifier

Let $\mathcal{P}_C$ denote a set of distributions $P[C]$. For a generic attribute A_i and for each $c \in \mathcal{C}$, let $\mathcal{P}^c_{A_i}$ denote a credal set of the conditional distributions $P[A_i|c]$. We refer to these sets also as to *local* credal sets. Note that we do not address the way the local credal sets must be provided. This can be done by using statistical inference (e.g., as discussed in the concluding section) or by subjective judgements. In the following we assume that the local credal sets are given.

Definition 2.1 The naive credal classifier is the model characterized by the set $\mathcal{P}$ of joint distributions $P[C, A_1, \ldots, A_n]$ that are obtained by assuming (1.1) and

making every possible combination of the distributions in the local credal sets,

$$\mathcal{P} = \left\{ P[C] \prod_{i=1}^{n} P[A_i | C] \,\middle|\, P[C] \in \mathcal{P}_C, P[A_j | c] \in \mathcal{P}_{A_j}^c, j = 1, \ldots, n, c \in \mathcal{C} \right\}.$$

The definition of the NCC emphasizes that the availability of the local credal sets is all that is needed to build the classifier. Broadly speaking, we can specify a credal set in two different yet formally equivalent ways. In the first case, we provide a set of distributions and then we take its convex hull. This can always be represented by the set of its extreme points because the convex hull of the extreme points is the original credal set. An extreme point or extreme distribution of a credal set is an element that cannot be expressed as convex combination of other points in the set.

The second view characterizes a credal set by means of *linear constraints*. In fact, the convex hull of a non-empty and finite number of points is by definition a polytope and the extreme points are the vertices of the polytope. A polytope is also a closed and bounded geometric region described by linear constraints. For a credal set, the constraints can be imposed on the unknown probabilities of the elementary events. For example, the polytope $\mathcal{P}_C$ may be specified by imposing linear constraints on $P[c]$, $c \in \mathcal{C}$.

Thus extreme points and constraints are equivalent representation of a credal set; in the following we speak either of extreme points or of constraints according to our convenience. In either case, observe that the probability of a generic event lies in an interval whose extremes are the minimum and the maximum of the probability when the distribution varies in the credal set. Such extremes are also referred to as the *lower* and *upper probability* of the event and are denoted by $\underline{P}$ and $\overline{P}$, respectively.

In the next sections we derive the procedures for classification. These hold for local credal sets which satisfy the following assumptions,

$$\begin{cases} \overline{P}[c] > 0 \\ \overline{P}[a_i|c] > 0 \end{cases} \quad \text{for all } c \in \mathcal{C}, a_i \in \mathcal{A}_i, i = 1 \dots n. \tag{2.1}$$

Finally, let us emphasize that the definition of the NCC assumes that the local credal sets can be specified separately (Walley, 1991), meaning that they are logically independent. All the given procedures rely on this characteristic of the NCC.

3 Interval-dominance classification

This section develops a procedure for naive credal classification by a straight extension of the procedure for the NBC. The NBC classifies a generic pattern $(a_1, \dots, a_n)$ by computing the probability $P[c|a_1, \dots, a_n]$ for each $c \in \mathcal{C}$ and comparing them to select the class of maximum posterior probability. In the credal case such probabilities are intervals and for this reason we must provide procedures to compute intervals and define the way two intervals should be compared.

We address the comparison of two generic intervals by the following dominance criterion (which is called *strong dominance* in Luce and Raiffa, 1957).

Definition 3.1 Let X be a discrete random variable defined over $\mathcal{X}$ and let $\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}$ be two generic events. Let E represent what is known, and let the probabilities $P[\mathcal{X}'|E]$ and $P[\mathcal{X}''|E]$ be respectively represented by the intervals $I' = \left[\underline{P}[\mathcal{X}'|E], \overline{P}[\mathcal{X}'|E]\right]$ and $I'' = \left[\underline{P}[\mathcal{X}''|E], \overline{P}[\mathcal{X}''|E]\right]$. The interval I' is said to *dominate* I'' if $\underline{P}[\mathcal{X}'|E] > \overline{P}[\mathcal{X}''|E]$; in this case $\mathcal{X}'$ is said to *interval dominate* $\mathcal{X}''$.

The rationale behind the criterion is that since each probability in I' is greater than each probability in I'', $\mathcal{X}'$ is certainly more probable than $\mathcal{X}''$. Notice also

that interval dominance generally implies only a partial order, because if I' and I'' overlap, they cannot be compared.

The definition of interval dominance raises two questions. First, in order to apply interval dominance we only need the extreme points of the intervals. With special regard to the classification procedure, we only need to compute $\underline{P}\,[c\,|a_1,\ldots,a_n]$ and $\overline{P}\,[c\,|a_1,\ldots,a_n]$ for each $c \in \mathcal{C}$. These are defined by two optimization problems,

$$\underline{P}\,[c\,|a_1,\ldots,a_n] = \min_{P[C,A_1,\ldots,A_n]\in\mathcal{P}} P\,[c\,|a_1,\ldots,a_n]\,, \tag{3.1}$$

$$\overline{P}\,[c\,|a_1,\ldots,a_n] = \max_{P[C,A_1,\ldots,A_n]\in\mathcal{P}} P\,[c\,|a_1,\ldots,a_n]\,. \tag{3.2}$$

Section 3.1 solves such problems by an efficient organization of the calculations, as shown by the computational complexity analysis given in Section 3.2.

The second point is more closely concerned with the nature of credal classification. Since interval dominance only provides us with a partial order of the intervals, there may be no unique optimal choice for the class. This behavior is a fundamental characteristic of a credal classifier. Credal classification does not provide a single class unless the conditions justify one.

3.1 Computation of the interval

In this section we assume that the local credal sets satisfy the following,

$$\begin{cases} \underline{P}\,[c] > 0 \\ \underline{P}\,[a_i\,|c] > 0 \end{cases} \quad \text{for all } c \in \mathcal{C}, a_i \in \mathcal{A}_i, i = 1\ldots n, \tag{3.3}$$

for ease of presentation. The derivation extends to the case (2.1) by Theorem 8.1 of Walley (1981), as also discussed at the end of this section.

Consider the computation of the lower probability (the case of the upper probability is analogous), $\underline{P}\,[c\,|a_1,\ldots,a_n] = \min_{P[C,A_1,\ldots,A_n]\in\mathcal{P}} P\,[c\,|a_1,\ldots,a_n]$. The

objective function (i.e. the function to optimize) can be rewritten by applying the definition of conditional probability and using marginalization, as follows:

$$P\left[c\,|a_1,\ldots,a_n\right] = \frac{P\left[c,a_1,\ldots,a_n\right]}{\sum_{c'} P\left[c',a_1,\ldots,a_n\right]} \tag{3.4}$$

$$= \left(1+\frac{\sum_{c'\neq c} P\left[c',a_1,\ldots,a_n\right]}{P\left[c,a_1,\ldots,a_n\right]}\right)^{-1}, \tag{3.5}$$

where the passage from Eq. (3.4) to Eq. (3.5) is possible since $P\left[c,a_1,\ldots,a_n\right]>0$ by assumption (3.3). According to (1.1), expression (3.5) is also

$$P\left[c\,|a_1,\ldots,a_n\right] = \left(1+\frac{\sum_{c'\neq c} P\left[c'\right]\prod_{i=1}^{n} P\left[a_i\,|c'\right]}{P\left[c\right]\prod_{i=1}^{n} P\left[a_i\,|c\right]}\right)^{-1}. \tag{3.6}$$

Now the minimization problem is written by replacing its objective function with the right side of (3.6) and by doing the minimization over the local credal sets,

$$\min_{P[C]\in\mathcal{P}_C}\ \min_{P[A_i|c']\in\mathcal{P}^{c'}_{A_i},c'\in\mathcal{C},i=1\ldots n}\left(1+\frac{\sum_{c'\neq c} P\left[c'\right]\prod_{i=1}^{n} P\left[a_i\,|c'\right]}{P\left[c\right]\prod_{i=1}^{n} P\left[a_i\,|c\right]}\right)^{-1}. \tag{3.7}$$

Let us focus on the inner minimization problem: the goal is the *maximization* of the fractional function in parentheses, since this is equivalent to minimizing the reciprocal. Notice that it is possible to minimize the denominator and to maximize the numerator separately, since they do not share any term. This observation allows the inner optimization to be solved. Consider the denominator. $P\left[c\right]$ is non-negative, therefore the denominator is minimized when the product $\prod_{i=1}^{n} P\left[a_i\,|c\right]$ is minimized. This is done by setting each $P\left[a_i\,|c\right]$ to its lower probability, giving $\prod_{i=1}^{n} \underline{P}\left[a_i\,|c\right]$. An analogous argument holds for the numerator; $P\left[c'\right]$ is non-negative ($\forall c'\neq c$), and the sum consists of terms that can be optimized separately. Hence, the numerator is maximized when the product of the conditional probabilities is set to $\prod_{i=1}^{n} \overline{P}\left[a_i\,|c'\right]$ ($\forall c'\neq c$). Problem (3.7) becomes

$$\min_{P[C]\in\mathcal{P}_C}\left(1+\frac{\sum_{c'\neq c} P\left[c'\right]\prod_{i=1}^{n} \overline{P}\left[a_i\,|c'\right]}{P\left[c\right]\prod_{i=1}^{n} \underline{P}\left[a_i\,|c\right]}\right)^{-1}. \tag{3.8}$$

Following a similar argument, it is straightforward to obtain the formula for the upper probability, which is

$$\max_{P[C]\in\mathcal{P}_C}\left(1+\frac{\sum_{c'\neq c}P[c']\prod_{i=1}^{n}\underline{P}[a_i\,|c']}{P[c]\prod_{i=1}^{n}\overline{P}[a_i\,|c]}\right)^{-1}. \tag{3.9}$$

Observe that we can think of $P[c]$ as prior probabilities and $\prod_{i=1}^{n}\overline{P}[a_i\,|c]=\overline{P}[a_1,\ldots,a_n\,|c]$, $\prod_{i=1}^{n}\underline{P}[a_i\,|c]=\underline{P}[a_1,\ldots,a_n\,|c]$ as upper and lower likelihood functions (regarding $a_1,\ldots,a_n$ as data and c as possible parameter values). From this point of view, problems (3.8) and (3.9) are special cases of formulae given in Walley (1991, Section 8.5.4). (See also Walley, 1996a and Tessem, 1992, for the case when $\mathcal{P}_C$ is defined by specifying the upper and lower probabilities of each possible class.) These previously appeared in Theorem 8.1 of Walley (1981) that also allows (3.8) and (3.9) to be extended to the case (2.1) by rewriting them in order to remove the reciprocal.

3.2 Solution procedures and complexity

The following sections describe two different solution methods for problems (3.8) and (3.9), which are based on a combinatorial and a linear programming approach, respectively. The methods are also analyzed with respect to their computational complexity. We express the complexity by the notation $O(\cdot)$ as in Graham et al. (1989). Given the real-valued functions $f,g:\mathbb{N}^{+}\to\mathbb{R}$, we write $f(x)=O(g(x))$ if there exists a constant H such that $|f(x)|\leq H\,|g(x)|$ for each x.

3.2.1 Combinatorial procedure

We can solve problems (3.8) and (3.9) by a combinatorial approach. In fact, it is well-known that the optimal distributions can be found in the set of extreme distributions (Walley, 1991); in other words, it is possible to compute the optima by examining a *finite* number of points.

Let $ext\mathcal{P}^c_{A_i}$ denote the finite subset of $\mathcal{P}^c_{A_i}$ consisting of its extreme points ($\forall c \in \mathcal{C}, i = 1 \ldots n$). Analogously, let $ext\mathcal{P}_C$ be the set of extreme points of $\mathcal{P}_C$. Problems (3.8) and (3.9) are respectively equivalent to the following ones,

$$\min_{P[C] \in ext\mathcal{P}_C} \left(1 + \frac{\sum_{c' \neq c} P[c'] \prod_{i=1}^{n} \overline{P}[a_i | c']}{P[c] \prod_{i=1}^{n} \underline{P}[a_i | c]} \right)^{-1}, \tag{3.10}$$

$$\max_{P[C] \in ext\mathcal{P}_C} \left(1 + \frac{\sum_{c' \neq c} P[c'] \prod_{i=1}^{n} \underline{P}[a_i | c']}{P[c] \prod_{i=1}^{n} \overline{P}[a_i | c]} \right)^{-1}. \tag{3.11}$$

These can be solved by enumerating the extreme distributions in $ext\mathcal{P}_C$.

Of course, we also need the lower and upper probabilities of $P[a_i | c']$ ($\forall c' \in \mathcal{C}, i = 1 \ldots n$); these can be found by solving the problems $\min_{P[A_i|c'] \in ext\mathcal{P}^{c'}_{A_i}} P[a_i | c']$ and $\max_{P[A_i|c'] \in ext\mathcal{P}^{c'}_{A_i}} P[a_i | c']$ ($\forall c' \in \mathcal{C}, i = 1 \ldots n$), which can again be done by enumerating the extreme distributions in the respective feasible set.

In order to analyze the computational complexity of the procedure, we assume that the extreme distributions of the credal sets are already available. Notice that if the credal sets were specified by constraints, the procedure in Section 3.2.2 would be more appropriate. The present procedure might be applied as well, *after* computing the extreme distributions from the constraints, but computational problems might arise (refer to Section 3.2.2).

We denote by K the maximum of the number of extreme distributions, taken over all the local credal sets. Hence, the combinatorial computation of the extremes of $P[a_i | c']$ takes $O(K)$ time in the worst case. This must be repeated for every $c' \in \mathcal{C}$ and all $i = 1 \ldots n$, yielding $O(nK |\mathcal{C}|)$. We can regard this expression as the time required to setup the classifier, because the extrema of the conditional probabilities are computed once for all the subsequent classifications.

With regard to formulae (3.10) and (3.11), when the values of the conditional upper and lower probabilities are known, the expression inside parentheses is computed in time $O(n |\mathcal{C}|)$. The latter must be repeated for each extreme distribution in $ext\mathcal{P}_C$, in order to evaluate the optimum, yielding $O(nK |\mathcal{C}|)$. By adding this

term to the setup time we have the overall worst-case complexity, i.e.

$$O\left(nK\left|\mathcal{C}\right|\right). \tag{3.12}$$

3.2.2 Linear programming procedure

With reference to formula (3.12), we see that the classification of a pattern is a well-solvable task, if K is *not large*. In fact, the complexity grows only linearly with either the number of attributes or the number of *classes*, and these quantities are directly under the control of the model builder.

The case of K is different. Although the complexity grows linearly with K too, K is a potentially weak point for computational time. In fact, the number of vertices of a polytope of distributions can be *very large* and this number can be hidden in the definition of a credal set, in a way that it may not be immediately clear to the model builder: the latter might define the credal sets by providing constraints on the probability of some events; but, in this way, K *can grow exponentially* with the number of constraints.

As an example, consider a set of probability intervals $I_X = \{[l_i, u_i] \mid 0 \leq l_i \leq u_i \leq 1, i = 1 \ldots t, t > 1\}$ defining a credal set, $\mathcal{P}_X$, for a generic variable X defined over $\{x_1, \ldots, x_t\}$. $\mathcal{P}_X$ is the set of distributions $P[X]$ subject to the constraints $l_i \leq P[x_i] \leq u_i$, $i = 1 \ldots t$. Tessem (1992) has shown that the worst-case number of vertices of $\mathcal{P}_X$, say k_X, depends on t according to the following formula,

$$k_X(t) = \begin{cases} \binom{t+1}{(t+1)/2} \frac{t+1}{4}, \text{ if } t \text{ is odd} \\ \binom{t}{t/2} \frac{t}{2}, \text{ if } t \text{ is even.} \end{cases}$$

For instance: $k_X(10) \simeq 1.2 \times 10^3$, $k_X(20) \simeq 1.8 \times 10^6$ and $k_X(30) \simeq 2.3 \times 10^9$.

Such an exponential growth is a strong reason to search for an alternative solution procedure which can guarantee a better worst-case complexity. Linear programming is a basic tool for this purpose.

Let us reconsider the solution of problem (3.8), where we assume that each local credal set of the NCC is specified by a set of linear constraints. Let L denote

the worst-case complexity to solve a linear program (taken over all the local credal sets).

Setting up the classifier requires solving the problems $\min_{P[A_i|c'] \in \mathcal{P}^{c'}_{A_i}} P[a_i|c']$ and $\max_{P[A_i|c'] \in \mathcal{P}^{c'}_{A_i}} P[a_i|c']$ $(\forall c' \in \mathcal{C}, i = 1 \ldots n)$. These are optimizations of linear functions over linear domains (in particular, the objective function is represented by the single *optimization* variable $P[a_i|c']$); therefore the overall setup complexity is $O(nL|\mathcal{C}|)$.

Then, we compute the lower probability by maximizing the fractional function inside parentheses in (3.8). Observe that once the conditional upper and lower probabilities are available, such a problem is a fractional linear program. In fact, the objective function is a linear combination of the optimization variables $P[c']$ $(\forall c' \neq c)$ divided by a term proportional to the optimization variable $P[c]$; the feasible set of the problem is the polytope $\mathcal{P}_C$. Concerning complexity, we can compute all the products that define the objective, namely $\prod_{i=1}^{n} \overline{P}[a_i|c']$ $(c' \in \mathcal{C})$, in time $O(n|\mathcal{C}|)$. Solving the problem itself requires $O(L)$ time, because fractional linear problems can be turned into linear problems (by a result in Charnes and Cooper, 1962, also reported in Schaible, 1995, Section 2.2.2). Therefore, the overall time is $O(nL|\mathcal{C}| + n|\mathcal{C}| + L)$ which is also

$$O(nL|\mathcal{C}|). \tag{3.13}$$

Comparing expression (3.13) with (3.12), we see that now L replaces K. Formally, this means that the complexity of the linear programming-based solution procedure is polynomial; this is because linear programming can be solved in polynomial time (Khachian, 1979). In other words, L only grows as a polynomial function of the size of the linear problem (which depends on the number of variables and on the number of constraints), even when K grows exponentially.

Thus the new solution procedure provides the user of the system with a *good* theoretical bound on computational time. Observe that the procedure is also practically effective. In fact, the simplex method is well-known to efficiently solve very

large linear problems too. Finally, note that specialized methods are available that solve linear problems on particular credal sets (e.g., those generated by probability intervals, as in Walley, 1996a, p. 18) in even a quicker way .

4 Credal-dominance classification

So far, our analysis has focused on computation of the uncertainty intervals for the states $c \in \mathcal{C}$, with the aim of comparing them using interval dominance. It is important to observe that the information provided by credal sets is greater than that provided by intervals. The credal set can also represent constraints between probabilities, which disappear with the interval view. In particular, the credal set for $P\left[C \left| a_1, \ldots, a_n \right.\right]$, say $\mathcal{P}_C^{a_1,\ldots,a_n}$, generally conveys more information than that given by $\left[\underline{P}\left[c \left| a_1, \ldots, a_n \right.\right], \overline{P}\left[c \left| a_1, \ldots, a_n \right.\right]\right]$ $(c \in \mathcal{C})$. Therefore, it is natural to wonder if a comparison criterion different from interval dominance can better exploit such information.

Consider the following example. Suppose that $\mathcal{C} = \{c', c'', c'''\}$ and that $\mathcal{P}_C^{a_1,\ldots,a_n}$ has four extreme distributions: $(0.40, 0.35, 0.25)$, $(0.40, 0.25, 0.35)$, $(0.50, 0.35, 0.15)$ and $(0.50, 0.45, 0.05)$, where the elements of the vectors are respectively $P[c'|a_1, \ldots, a_n]$, $P\left[c'' \left| a_1, \ldots, a_n \right.\right]$ and $P\left[c''' \left| a_1, \ldots, a_n \right.\right]$. The intervals that these posterior probabilities belong to are $[0.40, 0.50]$, $[0.25, 0.45]$ and $[0.05, 0.35]$, respectively. We have that c' interval-dominates c''' (so that it can be discarded) and this is the only interval dominance, because the intervals of the remaining states overlap; interval dominance produces two undominated states, c' and c''. But it is easy to see that for any distribution in the credal set, $P\left[c' \left| a_1, \ldots, a_n \right.\right] > P\left[c'' \left| a_1, \ldots, a_n \right.\right]$ (because this is true for all the extreme distributions), i.e. c'' is dominated by c'. In other words, for the credal set at hand there is only a single dominant state, but this fact is hidden when interval dominance is used. Thus, we are led to the definition of a dominance criterion for sets of distributions.

Definition 4.1 Let X be a discrete random variable defined over $\mathcal{X}$ and let $\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}$ be two generic events. Consider the distribution $P[X|E] \in \mathcal{P}_X^E$, where E represents what is known, and $\mathcal{P}_X^E$ is a non-empty set of distributions. $\mathcal{X}'$ is said to be *credal dominant* as compared to $\mathcal{X}''$, $\mathcal{X}' \succ \mathcal{X}''$, if for every distribution $P[X|E] \in \mathcal{P}_X^E$, $P[\mathcal{X}'|E] > P[\mathcal{X}''|E]$.

Credal dominance is a strengthening of interval dominance to sets of distributions; it is also is a special case of strict preference as defined in Walley (1991, Section 3.7.7). Notice that interval dominance implies credal dominance, whereas the converse is not true as the example above shows. That is, not all credal dominances are captured by interval dominance.

Now, we examine whether 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 by combinatorial methods. When $\mathcal{P}_X^E$ is a polytope, $\mathcal{X}' \succ \mathcal{X}'' \iff P[\mathcal{X}'|E] > P[\mathcal{X}''|E]$ for all $P[X|E] \in ext\mathcal{P}_X^E$.

Let us develop the particular case of the NCC. Consider two states of C, namely c' and c''. We want to check whether

$$P[c'|a_1, \ldots, a_n] > P[c''|a_1, \ldots, a_n] \tag{4.1}$$

holds for all the joint distributions in $\mathcal{P}$. The question is equivalent to solving the following problem (Walley, 1991),

$$\min_{P[C, A_1, \ldots, A_n] \in \mathcal{P}} \left(P[c', a_1, \ldots, a_n] - P[c'', a_1, \ldots, a_n] \right). \tag{4.2}$$

If the optimum of problem (4.2) is positive, the answer to the credal dominance question is affirmative. Whenever the optimum is non-positive, the inequality (4.1) is false and $c' \succ c''$ is not verified. Notice that problem (4.2) only allows $c' \succ c''$ to be checked; testing $c'' \succ c'$ requires the minimization of the negated objective to be solved.

Problem (4.2) can be rewritten following an argument completely analogous to that used for problem (3.8), thus obtaining

$$\min_{P[C]\in\mathcal{P}_C} \left(P\left[c'\right] \prod_{i=1}^{n} \underline{P}\left[a_i \left| c'\right.\right] - P\left[c''\right] \prod_{i=1}^{n} \overline{P}\left[a_i \left| c''\right.\right] \right). \tag{4.3}$$

Also the solution of problem (4.3) is similar to the solution of problems (3.8) and (3.9). In particular, we can solve problem (4.3) either by a combinatorial approach or by linear programming. In the former case, we obtain the optimum by enumerating the extreme distributions of $\mathcal{P}_C$, after similarly computing the extremes of the conditional distributions. In the latter, we compute the extremes of the conditional distributions by linear programming and then solve the remaining linear problem (4.3).

From the analysis above, it also follows that checking credal dominance between two states (i.e. checking both directions of the inequality) can be performed more quickly than interval dominance. Two optimization problems must be solved, and the overall time required is lower than for problems (3.8) and (3.9) together. Compare, for instance, problem (4.3) with problem (3.8). Problem (4.3) is solved more quickly, because it only has $O(n)$ terms in the numerator versus the $O(n\left|\mathcal{C}\right|)$ terms in problem (3.8). By symmetry, the same is true for the remaining couple of problems.

However, it is possible that computing the set of undominated states of C is more expensive using credal dominance than using interval dominance. With interval dominance, $2\left|\mathcal{C}\right|$ optimizations like problem (3.8) allow all the intervals for C to be computed. Then, $O(\left|\mathcal{C}\right|^2)$ interval comparisons select the interval-undominated states. With credal dominance, $O(\left|\mathcal{C}\right|^2)$ optimizations like problem (4.3) are required, which are generally more expensive than $O(\left|\mathcal{C}\right|^2)$ interval comparisons. But notice that these complexities are the same when we use probability intervals, as Section 4.1 shows.

4.1 The case of probability intervals

Problem (4.3) admits a very simple and efficient solution when the credal sets of the NCC are defined using probability intervals, which are an important special case.

Let $I_X = \{[l_i, u_i] \mid 0 \leq l_i \leq u_i \leq 1, i = 1 \dots t\}$ be a set of probability intervals for the variable X defined over $\mathcal{X} = \{x_1, \dots, x_t\}$ (as in Section 3.2.2). We start by defining proper and reachable probability intervals as in Campos et al. (1994).

Definition 4.2 *I_X is* proper *if* $\sum_{i=1}^{t} l_i \leq 1 \leq \sum_{i=1}^{t} u_i$.

Definition 4.3 *I_X is* reachable *if* $u_i + \sum_{j=1, j\neq i}^{t} l_j \leq 1 \leq l_i + \sum_{j=1, j\neq i}^{t} u_j$, $i = 1 \dots t$.

These definitions are simply coherence conditions. It is possible to show (Campos et al., 1994) that I_X is proper iff $\mathcal{P}_X$ is not empty and that I_X is reachable iff the intervals are tight, i.e. for each lower or upper bound in I_X there is a distribution in $\mathcal{P}_X$ at which the bound is attained (notice that a set of reachable intervals is also proper). In the following we always assume that the intervals satisfy Definition 4.3; this requirement is not restrictive: each set of proper probability intervals can be transformed, in time $O(t)$, into a reachable set without altering $\mathcal{P}_X$.

Reachable intervals can also be regarded as a special case of upper and lower probabilities; they are Choquet capacities of order two. In this case it is well-known (Walley and Fine, 1982) that, given two mutually exclusive events $\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}$, there always exists a distribution $P \in \mathcal{P}_X$ such that $P[\mathcal{X}'] = \underline{P}[\mathcal{X}']$ and $P[\mathcal{X}''] = \overline{P}[\mathcal{X}'']$. When $\mathcal{P}_C$ in problem (4.3) is defined via reachable intervals, such a property applies and the optimal value of (4.3) is

$$\underline{P}[c'] \prod_{i=1}^{n} \underline{P}[a_i | c'] - \overline{P}[c''] \prod_{i=1}^{n} \overline{P}[a_i | c'']. \tag{4.4}$$

The latter has a positive impact on computational complexity. In fact, if all the credal sets of the NCC are defined with reachable intervals, the extremes of the conditional probabilities in (4.4) are, by definition, readily available and consequently the value (4.4) is computed in time $O(n)$. (Strictly speaking, we should also consider the time required to turn the intervals into reachable ones; but there are cases when the intervals, being naturally reachable, do not need such a treatment. For instance, this is the case for the intervals produced by the imprecise Dirichlet model in Walley, 1996b.) This also means that the computation of the credal-undominated states of C takes $O\left(n\left|\mathcal{C}\right|^2\right)$. Comparing the latter with the time required by the NBC to classify a pattern, namely $O\left(n\left|\mathcal{C}\right|\right)$, we see that the advantages of naive credal classification can be achieved with only a minor increase in computational complexity.

To find a simpler condition that is equivalent to credal dominance in the case of intervals, observe that (4.4) is simply the difference $\underline{P}\left[c', a_1, \ldots, a_n\right] - \overline{P}\left[c'', a_1, \ldots, a_n\right]$ and hence c' credal-dominates c'' if $\underline{P}\left[c', a_1, \ldots, a_n\right] > \overline{P}\left[c'', a_1, \ldots, a_n\right]$; but this is exactly the definition of interval dominance applied to the intervals $\left[\underline{P}\left[c', a_1, \ldots, a_n\right], \overline{P}\left[c', a_1, \ldots, a_n\right]\right]$ and $\left[\underline{P}\left[c'', a_1, \ldots, a_n\right], \overline{P}\left[c'', a_1, \ldots, a_n\right]\right]$. It follows that, when the NCC is defined by means of interval probabilities, we can regard the tests of credal dominance as tests of interval dominance on the joint prior probabilities of the class and the attributes. We use this observation in Section 5.

5 An example

This section presents a very simple example to show the principles of credal classification and in particular credal dominance. The example is simplified for clarity; the probabilities are artificial.

An insurance company wants to assess the risk it incurs in selling car insurance to a new customer. The risk (R) is classified as *low*, *medium*, or *high*, and is related

to the number and type of car accidents that are expected for such a customer. The company decides to model the risk in terms of two attributes of the customer: the *age* (A, defined over {young,middle-aged,old}) and the *city* where the customer lives (T, over {Venezia (VE),Treviso (TV),Milano (MI)}).

The credal sets of the NCC are based on the following assumptions. Concerning the customer's age, it is supposed that middle-aged persons have better behavior, as compared to both young and old people; concerning the risk of the Italian cities, the ranking is Venezia $<$ Treviso $<$ Milano. The credal sets, defined by means of reachable probability intervals (see Section 4.1), are reported in Tables 1, 2 and 3.

*** TABLE 1 ABOUT HERE ***
*** TABLE 2 ABOUT HERE ***
*** TABLE 3 ABOUT HERE ***

For example, Table 1 expresses the fact that most 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 few people (5% – 8%) are high-risk. Table 2 expresses the fact that when the risk is low, the most probable age is middle-age; when it is medium, the three states of A have similar probabilities; whereas when the risk is high, people are most probably young, otherwise old.

*** FIGURE 1 ABOUT HERE ***

Let us consider the case of an old person living in Venezia. Following Section 4.1, we compute the three intervals for $P[R, A = old, T = VE]$, recalling that $\underline{P}[R, A = old, T = VE] = \underline{P}[R]\,\underline{P}[A = old\,|R]\,\underline{P}[T = VE\,|R]$ and $\overline{P}[R, A = old, T = VE] = \overline{P}[R]\,\overline{P}[A = old\,|R]\,\overline{P}[T = VE\,|R]$. The intervals are shown in Figure 1 by means of line segments; for instance, $P[R = low, A = old, T = VE]$ belongs to the interval $[0.151, 0.208]$ as represented by the lowest segment. By applying interval dominance to such intervals, we obtain a total order on the states

of R because the intervals do not overlap. In this case the customer is classified as low-risk.

We may be interested in the posterior probability of this risk category. We use (3.10) and (3.11) to compute $\underline{P}[R = low|A = old, T = VE]$ and $\overline{P}[R = low|A = old, T = VE]$, respectively. Such formulae require the extremes of the two conditional probabilities $P\,[A = old\,|R = low\,]$ and $P\,[T = VE\,|R = low\,]$. These extremes are readily available from Tables 2 and 3. The formulae also require the extreme distributions of $P\,[R]$. These can be computed from the intervals in Table 1 by simple procedures, as in Campos et al. (1994). In this way we find that the probability $P[R = low|A = old, T = VE]$ lies in the interval $[0.922, 0.975]$.

*** FIGURE 2 ABOUT HERE ***

Next, let us compute the risk category of a young person in Milano. Intuitively, the subject should be high-risk, because each attribute is in the worst state. The intervals for $P[R, A = young, T = MI]$ are represented in Figure 2. Now interval dominance only implies a partial order of the states of R because the intervals for the states low and medium overlap, but it is still possible to obtain a single credal-dominant state, i.e., *high risk*. As in the preceding case, we can compute the posterior lower and upper probabilities of the dominant class. The probability $P[R = high|A = young, T = MI]$ lies in the interval $[0.435, 0.693]$.

*** FIGURE 3 ABOUT HERE ***

The last case concerns a young person who lives in Treviso. This is slightly more difficult to classify, intuitively, since there are opposite tendencies in the attributes: being a young person makes the risk higher, but the risk should be lowered by living in a city with moderate traffic. The intervals for the present case are in Figure 3. This time a single credal-dominant state is not available because state high is credal-dominated and the intervals for the other two states overlap. The result of the classification is the set $\{low, medium\}$.

In this case, the posterior probability of the classification is $1-P[R = high|A = young, T = TV]$. The extremes of $P[R = high|A = young, T = TV]$ can be computed as in the preceding examples, obtaining the interval $[0.100, 0.267]$. It follows that $P[R \neq high|A = young, T = TV]$ lies in $[0.733, 0.900]$. In general we cannot apply this simple method to compute the posterior lower and upper probabilities of a set of classes, if $\mathcal{C}$ has more than 3 states. Section 6 describes a general method.

6 Posterior lower and upper probabilities of a set of classes

This section addresses the problem of computing the probability of a set of classes $\mathcal{C}' \subseteq \mathcal{C}$, conditional on the observed state of the attributes. As usual, such a probability is an interval, obtained by solving two optimization problems,

$$\underset{P[C,A_1,\dots,A_n]\in\mathcal{P}}{opt} P\left[\mathcal{C}'\,|a_1,\dots,a_n\right], \tag{6.1}$$

where $opt \in \{\min, \max\}$. (We derive the formulae for the optimum by assuming (3.3); similar considerations to those presented in Section 3.1 apply.) Consider the computation of $\underline{P}\left[\mathcal{C}'\,|a_1,\dots,a_n\right]$. Observe that $P\left[\mathcal{C}'\,|a_1,\dots,a_n\right] = \sum_{c\in\mathcal{C}'} P\left[c\,|a_1,\dots,a_n\right]$ since the events are mutually exclusive. The problem becomes

$$\begin{aligned} \underline{P}\left[\mathcal{C}'\,|a_1,\dots,a_n\right] &= \min_{P[C,A_1,\dots,A_n]\in\mathcal{P}} \sum_{c\in\mathcal{C}'} P\left[c\,|a_1,\dots,a_n\right] && (6.2)\\ &= \min_{P[C,A_1,\dots,A_n]\in\mathcal{P}} \frac{\sum_{c\in\mathcal{C}'} P\left[c,a_1,\dots,a_n\right]}{\sum_{c\in\mathcal{C}} P\left[c,a_1,\dots,a_n\right]}. && (6.3) \end{aligned}$$

The arguments used in previous sections allow problem (6.3) to be written as

$$\min_{P[C]\in\mathcal{P}_C} \left(1 + \frac{\sum_{c\in\mathcal{C}\cap\neg\mathcal{C}'} P\left[c\right]\prod_{i=1}^{n}\overline{P}\left[a_i\,|c\right]}{\sum_{c\in\mathcal{C}'} P\left[c\right]\prod_{i=1}^{n}\underline{P}\left[a_i\,|c\right]}\right)^{-1}, \tag{6.4}$$

where $\neg\mathcal{C}'$ denotes the complement of $\mathcal{C}'$.

The problem related to the upper probability is derived in analogous way, obtaining

$$\max_{P[C]\in\mathcal{P}_C} \left(1 + \frac{\sum_{c\in\mathcal{C}\cap\neg\mathcal{C}'} P[c] \prod_{i=1}^{n} \underline{P}[a_i|c]}{\sum_{c\in\mathcal{C}'} P[c] \prod_{i=1}^{n} \overline{P}[a_i|c]}\right)^{-1}. \tag{6.5}$$

Finally, observe that, as usual, problems (6.4) and (6.5) can be solved by optimizing the fractional linear functions in parentheses; this is achieved either by linear programming or via a combinatorial approach.

7 Conclusions

This paper proposes credal classification as a generalization of standard classification and realizes it by extending the naive Bayes classifier to credal sets. It derives the related procedures for classification and for the computation of posterior lower and upper probabilities. By analyzing the computational complexity of the procedures, it shows that naive credal classification is a well-solvable task. In other words, it shows that the application of credal sets to naive classification is simple to realize and allows imprecision about probability values to be included in the model.

Naive credal classification preserves the advantages of the NBC approach while adding flexibility and realism. For that reason it would be useful to extend credal classification to more general models. That might be achieved, for instance, by exploiting known patterns of dependence between attributes, conditional on the class, in the direction of relaxing assumption (1.1).

To make the NCC applicable to common classification problems, the fundamental issue that needs to be addressed is how to learn the local credal sets from a random sample of data. Of course, that is a problem of statistical inference.

We can regard the data as a random sample from a multinomial distribution. We distinguish two cases according to the absence or presence of missing values.

First consider complete data. For this case, Walley (1996b) has proposed the *imprecise Dirichlet model*, which models prior ignorance through a set of Dirichlet distributions and makes posterior inferences by combining it with the observed likelihood function. There are a number of important properties of the model that make it a natural candidate to infer the local credal sets; for example, inferences are independent of the definition of the sample space. Combining the imprecise Dirichlet model with assumption (1.1), which is specific to the NCC, seems a well-founded and promising way to infer the local credal sets.

On the other hand, the NCC should be able to deal also with missing data, which are a pervasive problem in applied statistical inference. That is possible because, when no assumptions are made concerning the missingness mechanism, incomplete data can be regarded as another source of imprecision, as in Zaffalon (2000): the imprecision induced by missing data can be directly represented in the local credal sets of the NCC. That paper also shows that, in some cases, it is possible to combine the imprecision due to missing data with Walley's imprecise Dirichlet model. It appears that this approach will enable the NCC to be inferred from incomplete data sets in a simple and sound way.

Acknowledgements

I am indebted to Peter Walley for his support during all the development of the work. He contributed with very insightful comments, by suggesting relevant references and also by helping me to improve the English style of the paper. This paper has benefited greatly from his suggestions. Thanks also to Jean-Marc Bernard and to two anonymous referees for helpful comments, and to Luca Maria Gambardella and Carlo Lepori for their kind attention and support. This work

was supported in part by SUPSI DIE, http://www.supsi.ch, under CTI grant # KTI 4217.1.

References

Campos, L., Huete, J., Moral, S., 1994. Probability intervals: a tool for uncertain reasoning. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* **2**, 167–196.

Charnes, A., Cooper, W.W., 1962. Programming with linear fractional functionals. *Naval Res. Logist. Quarterly* **9**, 181–186.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* **29**, 103–130.

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

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), 1996. *Advances in Knowledge Discovery and Data Mining*. MIT Press, London.

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

Giri, N.C., 1996. *Multivariate Statistical Analysis*. Marcel Dekker, New York.

Graham, R.L., Knuth, D.E., Patashnik, O., 1989. *Concrete Mathematics: a Foundation for Computer Science*. Addison-Wesley, Reading, MA.

Huberty, C.J., 1994. *Applied Discriminant Analysis*. Wiley, New York.

Johnson, R.A., Wichern, D.W., 1988. *Applied Multivariate Statistical Analysis*, 2nd edn. Prentice Hall, New York.

Khachian, L.G., 1979. A polynomial algorithm for linear programming. *Doklady Akad. Nauk. USSR* **244**, 1093–1096. Translated in *Soviet Math. Doklady* **20**, 191–194.

Levi, I., 1980. *The Enterprise of Knowledge*. MIT press, London.

Luce, R.D., Raiffa, H., 1957. *Games and Decisions*. Wiley, New York.

McLachlan, G.J., 1992. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Schaible, S., 1995. Fractional programming. In: Horst, R., Pardalos, P.M. (Eds.), *Handbook of Global Optimization*. Kluwer, The Netherlands, pp. 495–608.

Tessem, B., 1992. Interval probability propagation. *Internat. J. Approx. Reason.* **7**, 95–120.

Walley, P., 1981. Coherent lower (and upper) probabilities. *Statistics Research Report No. 22*. University of Warwick, Coventry, UK.

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

Walley, P., 1996a. Measures of uncertainty in expert systems. *Artificial Intelligence* **83**, 1–58.

Walley, P., 1996b. Inferences from multinomial data: learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58**, 3–57.

Walley, P., Fine, T.L., 1982. Towards a frequentist theory of upper and lower probability. *Ann. Statist.* **10**, 741–761.

Zaffalon, M., 2000. Exact credal treatment of missing data. *J. Statist. Plann. Inference*. Accepted for publication.

Table 1: Prior probability intervals, $\left[\underline{P}\left[R\right], \overline{P}\left[R\right]\right]$, of risk classes

	low	[0.77,0.85]
R	medium	[0.10,0.15]
	high	[0.05,0.08]

Table 2: Conditional probability intervals for age given *risk category* $\left[\underline{P}\left[A\,|R\right],\overline{P}\left[A\,|R\right]\right]$

		R		
		low	medium	high
A	young	[0.15,0.22]	[0.27,0.32]	[0.60,0.70]
	middle-aged	[0.50,0.55]	[0.33,0.38]	[0.05,0.15]
	old	[0.28,0.34]	[0.34,0.38]	[0.20,0.30]

Table 3: Conditional probability intervals for city given *risk category* $\left[\underline{P}\left[T\left|R\right.\right], \overline{P}\left[T\left|R\right.\right]\right]$

		R		
		low	medium	high
T	VE	[0.70,0.72]	[0.15,0.20]	[0.02,0.06]
	TV	[0.18,0.20]	[0.60,0.65]	[0.22,0.28]
	MI	[0.08,0.10]	[0.20,0.25]	[0.66,0.72]

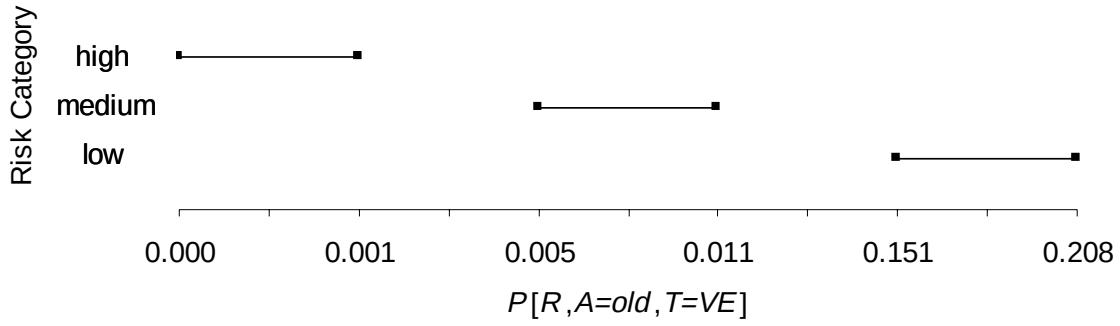

Figure 1: Joint probability intervals for risk category, age and city $[\underline{P}[R, A = old, T = VE], \overline{P}[R, A = old, T = VE]]$

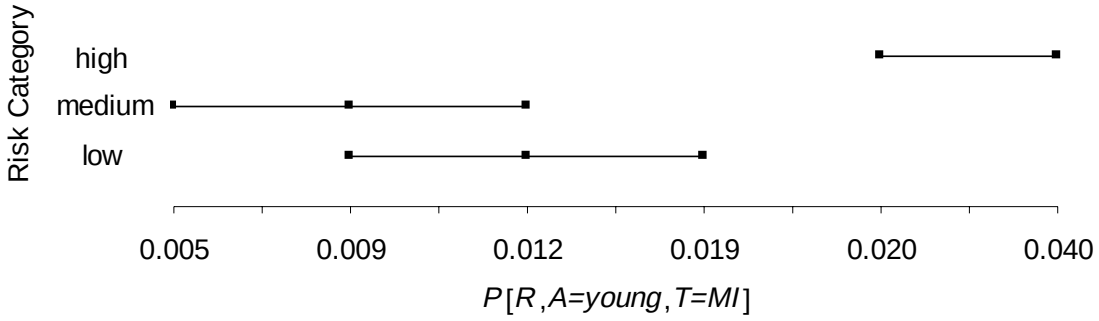

Figure 2: Joint probability intervals $[\underline{P}[R, A = young, T = MI], \overline{P}[R, A = young, T = MI]]$

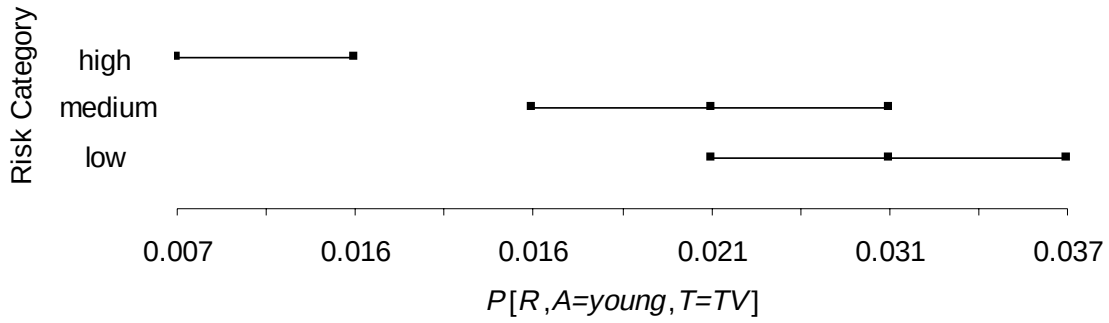

Figure 3: Joint probability intervals $[\underline{P}[R, A = young, T = TV], \overline{P}[R, A = young, T = TV]]$