NEGATIVE INFORMATION INTEGRATION IN PROBABILISTIC CLASSIFIERS

José Carlos Ferreira da Rocha, Alaine Margarete Guimarães, Thiago Marcomini Caires

Universidade Estadual de Ponta Grossa
LIC/DEINFO – UEPG
Ponta Grossa, Brazil

jrocha@uepg.br, amguimaraes@uepg.br

Abstract*. This study presents a procedure for integrating negative information into robust Bayesian inference. In the proposed procedure, negative information is codified as linear restrictions of the conditional probability intervals that quantify the uncertainty in the relationship between the classifying variables. During the inference, the robust Bayesian classifier is converted into a credal classifier. The classifier topology is the same as the Naive Bayesian classifier, and the optimization problems related to the inferences are solved by multilinear optimization. Since the objective of an inference is to compute the posterior probability interval of each class, by integrating the negative information, the inference procedure might obtain more precise intervals than those obtained by a robust Bayesian classifier. This might favor the use of the decision criterion called interval dominance when selecting plausible labels and defining a course of action for a given object of interest. The effectiveness of the procedure is illustrated with an example.*

Keywords. *Imprecise probabilities, Bayesian classification, Negative probabilistic information, Probabilistic inference*

1. Introduction

A classifier is a function that receives a descriptor (feature set) of a given object as input and returns a label that identifies its category. To carry out such a task, the naive Bayesian classifier (NBC) implements a probabilistic model that codifies the relationship between the descriptor attributes and the class labels. This probabilistic model is employed in the inference of the posterior probabilities of the classification hypotheses and the most probable hypothesis is selected using the Bayesian decision rule.

NBC learning can be summarized in two steps: In the first step, the network structure is defined, that is, the class labels are listed, and the attributes used to describe the objects are enumerated. The second step carries out the parameter training. In this step, learning algorithms estimate the marginal probabilities of each class and the conditional distributions of each attribute from the relative frequencies of a set of observations stored in a training data set.

When classifier training must be executed on an incomplete data set, it is not possible to exactly determine the frequency of each attribute value from the data. Consequently, the parameters obtained for the NBC are permeated by uncertainty and inaccuracy [21]. Considering this factor, Ramoni and Sebastiani [15] proposed the robust Bayesian classifier (RBC). This classifier extends the NBC using probability intervals to quantify the uncertainty in the model parameters. In this formalism, decision making regarding the classification of an object is performed in two steps: First, an optimization procedure computes lower and upper limits for the posterior probability of each

hypothesis; second, every classification hypothesis that is plausible with a decision criterion, called *interval dominance*, is reported as a possible categorization of the instance under analysis.

Since RBC can assign multiple hypotheses to the same object, it is not easy to choose a course of action based on the obtained results. Following Destercke [8], this study assumes that one way of mitigating this problem could be exploring the domain knowledge during the posterior probability intervals calculation of each hypothesis and possibly ruling out some non-plausible hypotheses. In this sense, this study presents a procedure called *ir-*, which uses credal network formalism to incorporate negative information into RBC inferences. A piece of negative information is a probabilistic statement that imposes constraints on the numerical parameters of the model.

The *ir-* procedure calculates the lower and upper limits for the probability of each conjunction between a classification hypothesis and the provided evidence. It converts the RBC into a credal classifier [21] and then performs a simplified version of the inference algorithm proposed by Campos and Cozman [7]. Next, *ir-* updates the multilinear program with negative information and solves it. The negative information is assumed to be supplied by experts and resembles a collection of linear inequations that further constrain the intervals of some model probabilities. The *ir-* effectiveness is illustrated by an example.

This paper is organized as follows: Section 2 presents a review of the NBC, the RBC, and the credal classifier; Section 3 introduces the *ir-* procedure; Section 4 presents an example of the *ir-* application; Section 5 concludes the paper.

2. Background review

Consider a random variable *C*, whose sample space, denoted by Ω, has the values $c_1 \ldots c_r$ that indicate the possible hypotheses for object classification of an interest domain. Consider a set of random variables **X,** whose elements $X_i \ldots X_n$ represent the characteristics used to describe the domain objects. The terms "attribute" and "descriptor" are also used to refer to the variables in **X**. In this study, it is assumed that each X_i is discrete and has a finite number of values. The sample space of X_i, i = 1,..., n, is represented as Ω_i and its elements are denoted by $x_{i,1}, \ldots, x_{i,r_i}$. Here, r_i indicates the cardinality of X_i.

A classification problem concludes the class or category identification of an object from its characteristics. A classifier is a function $F : \mathbf{X} \rightarrow C$ that returns a class label $c \in \Omega$, which would ideally be consistent with a set of observations $\mathbf{E} = \{x_{1,k_1}, \ldots, x_{n,k_n}\}$ [17]. Here, x_{i,k_i} reports the value of X_i for an object of interest *I*.

To infer the class of an object, a Bayesian classifier initially calculates the posterior probabilities $P(c_j|E)$, j = 1, ..., r. Then, it employs the Bayesian decision rule [10] to select the hypothesis $c*$ that maximizes the posterior probability.

$$c* = arg \max_{c_j \in \Omega} P\left(c_j|E\right)$$

The NBC can be described as a Bayesian network [13] B = (**G**,**P**), where **G** is an *n*-ary tree of height one, whose root represents the variable *C* and each external node is an element of **X**. **P** is a collection of conditional probability tables (CPTs) defined for the model variables. Specifically, every node X_i stores a function $p(X_i|C)$ in a CPT of dimension $r \times r_i$ in which the *k*th input of the *j*th line specifies the probability $P(x_{i,k}|c_j)$.

The root node CPT contains the marginal distribution $p(C) = \big(p(c_1), \ldots, p(c_r)\big)$. Figure 1 illustrates the NBC topology.

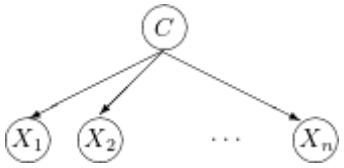

Figure 1: Naive Bayesian classifier topology.

From the Bayes theorem, we obtain P(c_j|**E**) ∝ P(**E**| c_j) P(c_j). In addition, since NBC assumes that each attribute is conditionally independent from the others, given the value of *C*, c* can be obtained by solving the following equation:

$$c* = arg\max_{c_j \in \Omega} P\left(c_j\right) \prod_{i=1}^{n} P\left(x_{i,k_i} \middle| c_j\right). \qquad (1)$$

Given the topology of an NBC and a dataset **D**, with observations about the variables in **X** ∪{*C*}, the maximum likelihood criterion prescribes that the CPTs of the classifier can be estimated as the relative frequencies observed in the data [10] [17]. Thus, if **D** has *m* cases and all the cases are complete, the parameters P($x_{i,k}$|c_j), of the CPT of X_i and P(c_j) of the root node are given, respectively, by the following:

$$P\left(x_{i,j} \middle| c_j\right) = m_{ijk}/m_j$$

$$P\left(c_j\right) = m_j/m$$

where m_j is the absolute frequency of the *j*th class in **D** and m_{ijk} is the number of cases in **D** for which the expression ($X_i = x_k \wedge C = c_j$) is true.

2.1 Robust Bayesian Classifier

A training dataset **D** is said to be incomplete or has missing data, when some of its entries have attributes whose values were not observed [4]. RBC is used to address this situation [15]. This classifier employs imprecise probability theory [20] to represent the uncertainty associated with the numerical parameters of an NBC and express its impact on the inferences. The main argument in favor of the use of this type of model is the following: If it is not possible to determine the probability of an event in a sample, it is possible to work with a probability interval.

RBC is a pair B_r = (**G**,**I**) in which **G** is a tree, arranged in the same way as an NBC, and **I** is a set of interval probabilities. The tree root stores the collection of intervals I_C = ([$\underline{P}(c_1)$, $\overline{P}(c_1)$],…, [$\underline{P}(c_n)$, $\overline{P}(c_n)$]), whereas each external node X_i stores a collection $I_{X_i|c_j}$ = ([$\underline{P}\left(x_{i,1}\middle|c_j\right)$, $\overline{P}\left(x_{i,1}\middle|c_j\right)$],…,[$\underline{P}\left(x_{i,n}\middle|c_j\right)$, $\overline{P}\left(x_{i,n}\middle|c_j\right)$]) for each $c_j \in \Omega_C$. That is, an RBC assumes that P(c_j) belongs to the interval [$\underline{P}(c_j)$,…,$\overline{P}(c_n)$], whose extremes are called lower and upper probabilities. This also applies to P($x_{i,k}$|c_j) in relation to [$\underline{P}\left(x_{i,i}\middle|c_j\right)$, $\overline{P}\left(x_{i,i}\middle|c_j\right)$]. Another CBR assumption is that each collection of intervals is consistent [5]. Additionally, $\sum_{j=1}^{r} P\left(c_j\right) = 1$ and $\sum_{k=1}^{r_i} P\left(x_{i,k}\middle|c_j\right) = 1$.

Figure 2 shows an RBC topology, where C is a propositional variable whose values v and f indicate the class labels. The attributes X_1 and X_2 are two binary variables, which can be + or – (positive or negative) values. In the example, the root node is associated with intervals I_C = ([0.55, 0.6], [0.4, 0.45]). The attributes are related to the following interval collections: $I_{X_1|v}$ = ([0.1, 0.3], [0.7, 0.9]), $I_{X_1|f}$ = ([0.6, 0.8], [0.2, 0.4]), $I_{X_2|v}$ = ([0.9, 0.6], [0.1, 0.4]), and $I_{X_2|f}$ = ([0.5, 0.6], [0.4, 0.5]).

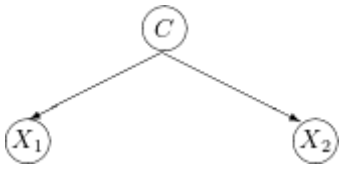

Figure 2 – Robust Bayesian classifier

Given an incomplete dataset, the extremes of the interval $I_{X_i|c_j}$ are calculated according to the expressions:

$$\underline{P}(x_{i,k}|c_j) = m_{ijk}/(m_j + \underline{m}_{ijk}) \quad (2)$$

$$\overline{P}(x_{i,k}|c_j) = (m_{ijk} + \overline{m}_{ijk})/(m_j + \overline{m}_{ijk}) \quad (3)$$

In these equations, m_j and m_{ijk} are defined in the same way as in the NBC training. The term $\underline{m}_{ijk}$ indicates the number of cases in which: (a) X_i is not observed and $C = c_j$ or (b) C is unknown and X_i is known and different from x_{ik} or (c) neither X_i nor C is known. Therefore, $\underline{m}_{ijk}$ accounts for the cases that are, or could be, associated with the category c_j. Additionally, $\overline{m}_{ijk}$ is the sum of all the incomplete cases that could be related to inputs in which $X = x_{ik}$ and $C = c_j$. Thus, if (a) m_{ij*} is the number of cases in which $C = c_j$ but X_i is not observed; (b) m_{i*h} indicates the number of registers in which $X_i = x_{ih}$ while the label of the instance is unknown; and (c) m_{***} is the number of cases in which X_i and C are unknown, the terms $\underline{m}_{ijk}$ and m_{ijk} are obtained by:

$$\underline{m}_{ijk} = m_{ij*+m_*+\sum_{h\neq k} m_{i*h}}$$

$$\overline{m}_{ijk} = m_{ij*+m_*+m_{i*k}}.$$

In these expressions, m_{ij*}, m_{i*h} and m_{***} are named virtual frequencies and indicate the number of incomplete cases that could be related to the event $x_{i,k}|c_j$. Regarding the root node, the extremes of each interval are given by $P(c_j) = m_j/m$ and $P(c_j) = (m_*+m_j)/m$.

Similar to NBC, the classification of an instance I with a RBC is also processed in two phases. The first phase calculates the posterior intervals of each hypothesis using message propagation algorithms in interval Bayesian networks [18] [4]. The collection of intervals is represented as $I_{C|E} = ([\underline{P}(c_j|E]), \overline{P}(c_j|E)]: i = 1, \ldots, r)$. The second phase dismisses hypotheses that are not plausible using an interval dominance criterion [15] [20].

A hypothesis c_l is dominated by the hypothesis c_j, $l \neq j$, if, and only if, $\overline{P}(c_l|E) < \underline{P}(c_j|E)$. If c_j dominates every other hypothesis, the analyzed object should be labeled as a member of c_j. If c_j is dominated by every other hypothesis, then it should be discarded. If c_j is non-dominated, but there are other hypotheses that neither dominate nor are dominated by c_j, the object of interest should be associated with every non-dominated hypothesis.

2.2 Credal Classifier

A credal set on $X \in \mathbf{X}$, denoted by *K(X)*, is a convex set of marginal distributions *p(X)* [12]. In this study, it is assumed that *K(X)* is a polytope, whose vertices are probabilistic distributions on *X*. Thus, if $p_1(X), \ldots, p_t(X)$ denotes the extremes of the credal set and *cc* represents the convex hull operation [19], then $K(X) = cc(p_1(X), \ldots, p_t(X))$. Similarly, a conditional credal set *K* is composed of conditional distributions *P* so that $X_1, X_2 \in X$ and $x_{2,*\in\Omega_2}$. Similar to the marginal case, *K* can be described as the convex hull of *t* extreme distributions.

A credal network is a pair B_C = (**G**,**Q**) in which **G** is a directed acyclic graph whose nodes represent elements of a set of variables **X** and the arcs indicate a direct probabilistic dependence [6]. This study considers networks that present a naive credal classifier (NCC) topology [21, 4]. **G** is a tree of height one whose structure is close to NBC. Each node also represents an attribute that stores a collection of separately specified credal sets [16]. The classifier root stores the collection **Q**(*C*) whose single element is the credal set *K(C)* and each external node maintains the collection **Q**(X_i|*C*) composed of the conditional credal sets $K(X_i|c_1), \ldots, K(X_i|c_r)$.

Figure 3 illustrates an NCC whose structure is of sufficient likeness to RBC (Figure 2). In this example, the root node is associated with the collection **Q**(*C*) = {*K(C)* = *cc*((1⁄3; 2⁄3), (0,4; 0,6))}. The collections of X_1 and X_2 are as follows: **Q**(X_1|*C*) = {$K(X_1|v)$ = *cc*((0,1; 0,9),(0,3; 0,7)), $K(X_1|f)$ = *cc*((0,6; 0,4),(0,8; 0,2))} and **Q**(X_2|*C*) = {$K(X_2|v)$ = *cc*((0,9; 0,1),(0,6; 0,4)), $K(X_2|f)$ =*cc*((0,5; 0,5),(0,4; 0,6))}.

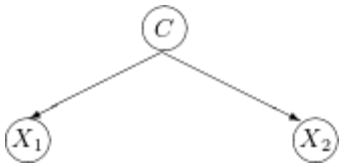

Figure 3: Naive credal classifier topology.

NCC also assumes that each attribute is conditionally dependent on the others, given the value of the variable *C*. However, as different ways of formulating the independence relation in the credal set theory exist [1], it is necessary to choose the one most suitable to the objectives of the application. This study considers NCCs that encode strong independence relations [5]. Given the instantiation $C = c_j$, two attributes X_i and X_l are strongly independent if each vertex of the credal set $K(X_i, X_l|c_j)$, whose elements are distributions of the type, $p(X_i, X_l|c_j)$, can be factored as $p(X_i|c_j).p(X_l|c_j)$. Additionally,

$p(X_i|c_j) \in Ext(K(X_i|c_j))$ and $p(X_l|c_j) \in Ext(K(X_l|c_j))$.

Campos and Cozman [7] describe an inference algorithm in credal networks that can be used to calculate upper and lower bounds for the posterior probability of any event defined on the network variables. The algorithm has two steps: In the first step, the algorithm generates a multilinear optimization program [9], whose objective function is the posterior probability of the interest event and the feasible region is defined by the specification of the credal network sets and the probability theory axioms. In the second step, the algorithm performs a non-linear optimization procedure to solve the multilinear program.

3. Robust classification and negative information

A negative information specifies values that cannot be assigned to the parameters of a model [3], for example, by establishing that a parameter cannot be larger than a given threshold. Regarding imprecise probabilistic models, this type of information usually sets up linear constraints for the probability measures [8]. Given that, this work assumes that any report of negative information concerning the parameters of an RBC encodes inequalities that: (a) set up bounds for the probability of an event, or (b) make a partial order relationship (comparison) between two measures of explicit conditional probabilities.

Thus, as before, let $I_{X_i|c_j} = ([\underline{P}(x_{i,1}|c_j), \overline{P}(x_{i,1}|c_j)]\ldots,[\underline{P}(x_{i,n}|c_j), \overline{P}(x_{i,n}|c_j)])$ be the collection of probability intervals of the variable X_i in a CBR B_r. A negative information that sets a lower bound $\underline{l}_{ikj}$ for $P(x_{i,k}|c_j)$ assumes the format of Equation 4. Similarly, an upper bound $\overline{l}_{ikj}$ for $P(x_{i,k}|c_j)$ is expressed as in Equation 5 and the information that x_{ik_1} is as likely as or is more likely than x_{ik_2} when $C = c_j$, is described in Equation 6.

$$P(x_{i,k}|c_j) \geq \underline{l}_{ikj} \quad (4)$$

$$P(x_{i,k}|c_j) \leq \overline{l}_{ikj} \quad (5)$$

$$P(x_{i,k_1}|c_j) - P(x_{i,k_2}|c_j) \geq 0 \quad (6)$$

Procedure *ir*¯ explores two facts. The possibility of convert each collection $I_{X_i|c_j}$ of a RBC into a credal set $K(X_i|c_j)$ and the possibility of append the constraints declared in the negative information to the definition of $K(X_i|c_j)$. It generates a new credal set $K' \subseteq K(X_i|c_j)$. Notably, K' can reduce the feasible region of the multilinear program of an inference and, by doing so, it can reduce the imprecision of the posterior probability interval of each class label.

Algorithm 1 (Figure 4) describes the *ir*¯ procedure. How it can be seen, *ir*- receives a RBC, B_r, a collection of negative information, L and the evidence **E. In** its first step, B_r is converted into a credal classifier B_C. That is, it builds a credal network with the same topology of the RBC and then, converts each RBC collection into a credal set **Q** [4]. In next, **Q** is associated with a node of the NCC.

Procedure: ir^-

Input: a RBC B_r, the evidence **E** and the negative information L;

1 Convert B_r into a credal classifier B_C;

2 For each $c_j \in \Omega_C$ do

a Execute the algorithm in Algorithm 2 [7] to generate the multilinear programs M_1 and M_2

b Generate the programs M'_1 and M'_2 by inserting the negative information into M_1 and M_2;

c Determine $[\underline{P'}(c_j \wedge E), \overline{P'}(c_j \wedge E)]$ by solving M'_1 and M'_2;

3 Return the intervals for each class in Ω_C.

Figure 4: Algorithm 1 - The *ir-* procedure.

Thus, for the root node, we have $\mathbf{Q}(C) = \{K(C)\}$, with $K(C) = \{P(C)\}$. For an external node, $I_{X_i|c_j}$originates the credal set $K(X_i|c_j)$; here, $K(X_i|c_j)$ is defined as the largest credal set[1] that agrees with the intervals in $I_{X_i|c_j}$:

$$K(X_i|c_j) = \{p(X_i|c_j): \forall_{x_{ik}\in\Omega_i} P(x_{ik}|c_j) \in [\underline{P}(x_{i,k}|c_j), \overline{P}(x_{i,k}|c_j)], \sum_{x_{ik}\in\Omega_i} P(x_{ik}|c_j) = 1\}.$$

The collection $\mathbf{Q}(X_i|C)$ is then defined as the set $\{ K(X_i|c_j): j = 1..t \}$.

The second step of ir^- embeds the inference computation. The step (2a) build the multilinear programs M_1 and M_2. The step (2b) generates M'_1 and M'_2 by appending the negative information into M_1 and M_2. In next, step (2c) runs a solver algorithm that computes the solution of M'_1 and M'_2.

In the last step, the procedure ir^- returns an interval $[\underline{P'}(c_j \wedge E), \overline{P'}(c_j \wedge E)]$ for each classification hypothesis c_j. The limits $\underline{P'}(c_j \wedge E)$ and $\overline{P'}(c_j \wedge E)$ are outer bounds for the desired probabilities, given the negative information and credal sets of the NCC.

Procedure: Multilinear program generator for computing $\underline{P}(c_j \wedge E)$

Input: a NCC B_C, the evidence **E** and an integer *j*;

[1] In this study, the credal sets were implemented using a representation of polytopes based on semi-spaces (representation-H; see [2], [11]).

Definition: **R** is a set of inequalities;

1 Initialize **R** with the constraints that specify the credal sets in B_C;

2 Define de program $\mathcal{M}_1 = min\, P(c_j) \cdot \prod_{i=i}^{n} P(x_{i,k_i}|c_j)$, s.a. **R**;

3 Return $\mathcal{M}_1$.

Figure 5: Algorithm 2 – Multilinear program generator.

Algorithm 2 shows the routine that generates the multilinear program $\mathcal{M}_1$. , adapted from Campos and Cozman [7]. Similarly to the original algorithm, the procedure in Figure 5 processes the measures of the model probabilities in terms of symbolic expressions and generates a multilinear optimization program. In step (1), the algorithm builds up the feasible region **R** as the union of every constraint defining the credal sets of C. In second step, the procedure defines $\mathcal{M}_1$ by setting up its objective function and feasible region **R**. The third step returns $\mathcal{M}_1$ to ir^-.

The routine that generates the multilinear program $\mathcal{M}_2$ is similar to Algorithm 2. Basically, it replaces the minimization operation with a maximization one.

4 An example application

Let $\mathcal{B}_r$ be the RBC of Figure 2 and **D** be the dataset presented in Table 1. The sample spaces of C, X_1 and X_2 are Ω_C = {*n*,*s*}, Ω_1 = {-,+} e Ω_2 = {0,1}, respectively. RBC learning procedure estimates the following probability intervals I_C = ([1⁄3;1⁄3],[2⁄3;2⁄3]), $I_{X1|C=n}$ = ([0,583;1],[0;0,417]), $I_{X1|C=s}$ = ([0,167;0,375],[0,625;0,83]), $I_{X2|C=n}$ = ([0,167;0,5],[0,5;0,83]) and $I_{X2|C=s}$ = ([0,542; 0,75],[0,25;0,458]), from equations 2 and 3, and dataset **D**.

Additionally, let it be the task of computing interval of probabilities for $P(C \wedge \mathbf{E})$ so that **E** = {X_1 = "-", X_2 = 1}. Using the procedure ir^-, the RBC B_R is converted into a credal classifier B_C whose collections **Q**(C), **Q**($X_1|C$) and **Q**($X_2|C$) are listed in Appendix I. In the absence of negative information, the procedure ir^- generates multilinear programs whose feasible region **R,** is defined as the union of the expressions listed in such appendix.

Table 1: Training base used in the example.

C	X_1	X_2		C	X_1	X_2		C	X_1	X_2
s	+	0		s	+	0		s	+	0
s	+	0		s	+	1		s	-	0
n	-	0		n	-			s	-	
s	+	1		n		0		s		
n	-	1		s	+	1		s	-	0
s	+	0		s		1		n		
n		1		n	-	1		n		1
s	+	0		s		0		s	+	1
n	-	1		n	-			s	+	0
s	+			s	+	0		s	-	1
s				n	-	1		s	+	0
s	+			n				s		0

Thus, Equations (8) and (9) describe the optimization problems $\mathcal{M}_1$ and $\mathcal{M}_2$ relatives to $\mathrm{P}(C = n \wedge E)$. Equations (10) and (11) do the same for $\mathrm{P}(C = s \wedge E)$.

$$\mathcal{M}_1: \min \mathrm{P}(\mathrm{C} = \mathrm{n}) \cdot \mathrm{P}(\mathrm{X}_1 = -|\ \mathrm{n}) \cdot \mathrm{P}(\mathrm{X}_2 = 1|\ \mathrm{n}) \text{ s.t. R} \quad (8)$$

$$\mathcal{M}_2: \max \mathrm{P}(C = n) \cdot P(X_1 = -|n) \cdot P(X_2 = 1|n) \text{ s.t. } R \quad (9)$$

$$\mathcal{M}_1: \min \mathrm{P}(C = s) \cdot P(X_1 = -|s) \cdot P(X_2 = 1|s) \text{ s.t. } R \quad (10)$$

$$\mathcal{M}_2: \max \mathrm{P}(C = s) \cdot P(X_1 = -|s) \cdot P(X_2 = 1|s) \text{ s.t. } R \quad (11)$$

After solving the problems above, *ir*- reports the intervals [0.06,0.1] and [0.037,0.083] for $\mathrm{P}(C = n \wedge E)$ and $\mathrm{P}(C = s \wedge E)$, respectively. Since, in this case, none of the classes is dominated, the classifier cannot discard any of them. In this example, the optimization phase was carried out with the Cobyla solver[14].

Now, let R_1 be a negative information report that states the constraint $P(X_1 = +|C = -) \leq 0.05$, the procedure *ir*- defines appends the list L = $\{R_1\}$ to the feasible region of $\mathcal{M}'_1$ and $\mathcal{M}'_2$ and then generates the following multilinear programs:

$$\mathcal{M}'_1: minP(C = n) \cdot P(X_1 = -|n) \cdot P(X_2 = 1|n) \text{ s.t. R} \cup \text{L} \quad (8)$$

$$\mathcal{M}'_2: maxP(C = n) \cdot P(X_1 = -|n) \cdot P(X_2 = 1|n) \text{ s.t. R} \cup \text{L} \quad (9)$$

$$\mathcal{M}'_1: minP(C = s) \cdot P(X_1 = -|s) \cdot P(X_2 = 1|s) \text{ s.t. R} \cup \text{L} \quad (10)$$

$$\mathcal{M}'_2: maxP(C = s) \cdot P(X_1 = -|s) \cdot P(X_2 = 1|s) \text{ s.t. R} \cup \text{L} \quad (11)$$

The solution of these new problems produces the intervals $\mathrm{P}(C = n \wedge \boldsymbol{E})$ = [0.1,0.11] and $\mathrm{P}(C = s \wedge \boldsymbol{E})$ = [0.04, 0.083]. As it could be observed, now, given the negative information, the new results allow the determination of the hypothesis *C* = s as

the most probable.

5 Conclusion

This work introduced the procedure ir^- that explores the uses of negative information when carrying out inferences in robust Bayesian classifiers. The procedure initially converts the RBC into a credal classifier and then integrates negative information, a kind of *a priori* knowledge, into the multilinear programs associated to each inference. As observed in the example, the procedure proposed was able to use negative information to reduce the uncertainty in inferences.

The procedure proposed assumes that negative information is consistent with the classifier probabilistic model. Whenever this is not the case, one alternative would be the use of the approach proposed by Destercke [8] that combines negative information to the classifier credal sets. If, on the one hand, this course of action might increase the inaccuracy of inferences and make the decision making process more difficult; on the other hand, it provides a way to explain the discrepancy between the trained model and information coming from other sources such as experts, technical reports and meta-analysis. This problem shall be approached in future studies.

Acknowldgements

To CAPES and Fundação Araucaria for the financial support.

6 References

[1] T. Augustin, F.P.A., C. G. de Cooman, and M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. Willey, 2014.

[2] D. Avis, K.Fukuda, and S. Picozzi. On canonical representation of convex polyhedra. In A. Cohen, X.-S. Gao, and N. Takayama, editors, *1st International Congress of Mathematical Software*, pages 350–360, 2002.

[3] I. Bloch, A. Petrosino, A. G. B. Tettamanzi, D. Dubois, and H. Prade. Special issue: Fuzzy sets in interdisciplinary perception and intelligence an overview of the asymmetric bipolar representation of positive and negative information in possibility theory. *Fuzzy Sets and Systems*, 160(10):1355 – 1366, 2009.

[4] G. Corani, A. Antonucci, and M. Zaffalon. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, chapter Bayesian Networks with Imprecise Probabilities: Theory and Application to Classification, pages 49–93. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[5] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk, Decision and Policy*, (5):165–185, 2000.

[6] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

[7] C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Second Starting AI Researcher Symposium (STAIRS)*, pages 50–61, Valencia, Spain, 2004.

[8] S. Destercke. Handling bipolar knowledge with imprecise probabilities.

International Journal of Intelligent Systems, 26(5):426–443, 2011.

[9] R. F. Drenick. Multilinear programming: Duality theories. *Journal of Optimization Theory and Applications*, 72(3):459–486, 1992.

[10] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2000.

[11] C. J. Geyer. Using the rcdd package, 2014.

[12] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, 1980.

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

[14] M. J. D. Powell. *A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation*, pages 51–67. Springer Netherlands, Dordrecht, 1994.

[15] M. Ramoni and P. Sebastiani. Robust bayes classifiers. *Artificial Intelligence*, 125(1.2):209 – 226, 2001.

[16] J. C. F. Rocha and F. G. Cozman. Inference with separately specified sets of probabilities in credal networks. In *18th Annual Conference on Uncertainty in Artificial Intelligence Conference*, pages 430–437, San Francisco, 2002. Morgan Kaufmann.

[17] S. Russell and P. Norvig. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, 3a edition, 2010.

[18] B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, (7):95–120, 1992.

[19] C. Toth, J. O'Rourke, and J. Goodman. *Handbook of Discrete and Computational Geometry*. Chapman and Hall/CRC, 2nd edition, 2004.

[20] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1991.

[21] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5 – 21, 2002.

Appendix 1

Collections

- $Q(C):\ \mathrm{K} = \{\mathrm{p(C)}: \mathrm{P(C = n)} = \frac{1}{3},\ \mathrm{P(C = s)} = \frac{2}{3}\ \}$
- $Q(X_1|C)$:
  - $\mathrm{K(X_1|\ C = n)} = \{\mathrm{p(X_1|\ n)}: \mathrm{P(X_1 = +|\ n) + P(X_1 = -|\ n) = 1},\ 0.417 \leq \mathrm{P(X_1 = +|\ n)} \leq 0.625, 0.583 \leq \mathrm{P(X_1 = -|\ n)} \leq 1\}$
  - $\mathbf{K}(X_1|\ C = s) = \{p(X_1|\ s): P(X_1 = +|\ s) + P(X_1 = -|\ s) = 1,\ 0.625 \leq P(X_1 = +|\ s) \leq 1, 0.167 \leq P(X_1 = -|\ s) \leq 0.375\}$
- $Q(X_2|C)$:
  - $\mathrm{K}(X_2|\ C = n) = \{\mathrm{p}(X_2|\ n): \mathrm{P}(X_2 = 1|\ n) + \mathrm{P}(X_2 = 0|\ n) = 1,\ 0.5 \leq P(X_2 = 1|\ n) \leq 0.8, 0.167 \leq P(X_1 = 0|\ n) \leq 0.5\}$
  - $\mathrm{K}(X_2|\ C = s) = \{\mathrm{p}(X_2|\ s): \mathrm{P}(X_2 = 1|\ s) + \mathrm{P}(X_2 = 1|\ s) = 1,\ 0.25 \leq P(X_2 = 1|\ s) \leq 0.458, 0.542 \leq P(X_1 = -|\ s) \leq 0.75\}$