Credal Nets with Probabilities Estimated with an Extreme Imprecise Dirichlet Model

A. Cano, M. Gómez-Olmedo, S. Moral
Dpto. Ciencias de la Computación
Universidad de Granada
18071 - Granada (Spain)
(acu,mgomez,smc)@decsai.ugr.es

Abstract

The propagation of probabilities in credal networks when probabilities are estimated with a global imprecise Dirichlet model is an important open problem. Only Zaffalon [21] has proposed an algorithm for the Naive classifier. The main difficulty is that, in general, computing upper and lower probability intervals implies the resolution of an optimization of a fraction of two polynomials. In the case of the Naive credal classifier, Zaffalon has shown that the function is a convex function of only one parameter, but there is not a similar result for general credal sets. In this paper, we propose the use of an imprecise global model, but we restrict the distributions to only the most extreme ones. The result is a model giving rise that in the case of estimating a conditional probability under independence relationships, it can produce smaller intervals than the global general model. Its main advantage is that the optimization problem is simpler, and available procedures can be directly applied, as the ones proposed in [7].

Keywords. Locally specified credal networks, global imprecise Dirichlet model, propagation algorithms, probability trees.

1 Introduction

Credal networks [12] are an extension of *Bayesian networks* where instead of having a joint precise global probability distribution we have a closed and convex set of possible distributions (a *credal set* [15]). This credal set produces a conditional credal set for each variable given its parents. There are two basic possibilities:

- The credal net is *separately specified* [12], i.e. the set of joint probability distributions is obtained by specifying a credal set of conditional probability distributions for each variable and each configuration of its parents, and then the joint credal set is the convex hull of the probability distributions obtained by multiplying the conditional probability distributions resulting by selecting one element from each conditional credal set (the joint credal set is the *strong extension* of the local conditional credal sets [12]).
- The credal net is *globally specified*, when only the joint credal set is given.

Most of the effort to design algorithms for computation in credal networks has been devoted to the case of separately specified credal nets. In general, this computation is equivalent to the resolution of a combinatorial optimization problem. One of the most promising approaches is based on the branch-and-bound technique [17, 7]. Also, there are several approximate algorithms, as the ones based on the simulated annealing technique [6] or the ones based on making the variables binaries in order to apply the efficient 2U algorithm [13, 3].

There is less work for globally specified credal networks. Preliminary models were proposed by Cozman [11, 12], but he followed a robust statistics methodology, considering credal sets that were neighborhoods of standard Bayesian networks. Recently, Antonucci and Zaffalon [2] have proposed a general method based on the use of auxiliary variables as in [5] to transform a globally specified credal network into a separately specified one. This allows the application of existing algorithms for separately specified networks to cases that initially were non-separately given.

However, the Antonucci and Zaffalon [2] transformation can not directly solve some important imprecise networks that can arise in practice. This is the case of credal nets in which conditional probabilities are estimated from a database of observations with an imprecise global Dirichlet model (IDM) [20]. The main problem is that in this situation we need auxiliary variables with infinite values (as the parameters can

have values in a continuum). If the IDM is locally applied to each conditional probability distribution (we consider a different IDM for each variable and each configuration of its parents), then there is no problem, as only the extreme parameters are relevant, and we can apply the transformation by Cano, Cano, and Moral [5]. This local application was initially proposed in Zaffalon [22]. But its main difficulty was that it has a tendency to produce too wide intervals that are too uninformative. For this reason Zaffalon [21] proposed[1] a global application of the IDM. This application has the problem that to compute lower and upper conditional probabilities, it is necessary the resolution of an optimization of a fraction of two polynomials in several parameters. In the case of the Naive credal classifier, Zaffalon [21] has shown that the function is a convex function of one parameter, and he proposes a numerical method for its optimization, but there is not a similar result for general networks.

In this paper, we propose the use of an imprecise global model, but we restrict the class IDM to the set of its extreme distributions. The result is a model giving rise to the same upper and lower probabilities, when estimating the uncertainty of a future simple event, but in the case of estimating a conditional probability under independence relationships, it can provide smaller intervals. Its main advantage is that the optimization problem is simpler, being possible to express the problem as a locally specified credal network for which standard algorithms for separately specified networks can be applied. In order to make the representation more efficient, we will represent conditional probability tables as probability trees as the ones used in [6, 7].

The paper is organized as follows: in Section 2 the basic concepts of credal sets and credal networks are given; in Section 3 we consider the IDM applied to estimating the probabilities in a credal network and introduce the extreme IDM; in Section 4 we show the transformation of a credal network with probabilities estimated with an IDM model (the general or the extreme one) into a locally specified credal network; in Section 5 the results of some preliminary experiments are shown; and finally Section 6 is devoted to the conclusions.

2 Credal Networks

Let $\mathbf{X}$ be a set of variables. Let us assume that each variable $X \in \mathbf{X}$ takes values on a finite set Ω_X (the frame of X). We shall use x to denote a generic value of X, $x \in \Omega_X$. If $\mathbf{Y} \subseteq \mathbf{X}$, then this variable will take values on he Cartesian product $\prod_{X \in \mathbf{Y}} \Omega_X$, denoted by $\Omega_{\mathbf{Y}}$. The elements of $\Omega_{\mathbf{Y}}$ are called *configurations* of $\mathbf{Y}$ and will be written as $\mathbf{y}$.

A *credal set* about $\mathbf{Y}$ is a closed and convex set of probability distributions on $\Omega_{\mathbf{Y}}$, denoted as $K_{\mathbf{Y}}$. If the number of extreme points is finite, then this convex set will be given by enumerating its extreme points: $K_{\mathbf{Y}} = \mathrm{CH}(\{P_1, \ldots, P_l\})$, where CH stands for the convex hull.

A *credal network* about variables $\mathbf{X}$ is a directed acyclic graph G, with a node for each $X \in \mathbf{X}$ and a credal set $K_{\mathbf{X}}$ such that every extreme distribution $P \in \mathrm{Ext}(K_{\mathbf{X}})$, factorizes according to the graph:

$$P(\mathbf{x}) = \prod_x P(x|\pi_X(\mathbf{x})) \tag{1}$$

where $\mathbf{\Pi}_X$ is the set of parents of X in G and $\pi_X(\mathbf{x})$ the configuration of these parents corresponding to $\mathbf{x}$.

A credal network is said to be *separately specified* [16] if the global credal set $K_{\mathbf{X}}$ can be obtained by giving a credal set, $K(X|\pi_X)$, for each variable X and each configuration of its parents π_X and then obtaining all the possible joint probabilities by expression (1).

A *locally specified credal set* [2] about $\mathbf{X}$ is composed of the following elements:

- The set of variables $\mathbf{X}$.
- An additional set of auxiliary variables which Antonucci and Zaffalon [2] call decision variables $\mathbf{D}$. Each variable $D \in \mathbf{D}$ takes values in a set Ω_D.
- A directed acyclic graph G with a node for each variable in $\mathbf{X} \cup \mathbf{D}$.
- A precise conditional probability distribution for each variable $X \in \mathbf{X}$ conditioned to its parents $\mathbf{\Pi}_X$ in G.
- A set $R_D(\pi_D) \subseteq \Omega_D$ for each decision variable D and each configuration of its parent variables π_D in G.

A locally specified credal net can define a credal net about $\mathbf{X} \cup \mathbf{D}$ and another about $\mathbf{X}$, by marginalization. These can be obtained by the following procedure:

- Consider for each $D \in \mathbf{D}$ the family of *decision functions* $f_D : \Omega_{\mathbf{\Pi}_D} \to \Omega_D$, such that $f_D(\pi_D) \in R_D(\pi_D), \forall \pi_D \in \Omega_{\pi_D}$.

[1] In the acknowledgments of the paper it is said that the model was suggested to him by Peter Walley

- Consider the set of *strategies*, where an strategy is given by a decision function for each decision variable.

- Each strategy defines a precise probability distribution: the one obtained by factorization according to G and given by the precise probability conditional distributions of each variable $X \in \mathbf{X}$ and the degenerate conditional probability distributions given by the decision functions of the decision variables of the given strategy: $P(d|\pi_D) = 1$, if $d = f_D(\pi_D)$ and 0, otherwise.

- The credal set about $\mathbf{X} \cup \mathbf{D}$ is the convex hull of the probabilities defined from the set of strategies. As the extreme points of this credal set factorize according to G, they define a credal network.

- The credal set about $\mathbf{X}$ is the one obtained by marginalization of the credal set about $\mathbf{X} \cup \mathbf{D}$ (equivalent to marginalizing each one of the probabilities). This set factorizes on the graph G' on $\mathbf{X}$, obtained by deleting nodes in $\mathbf{D}$ and connecting with arcs the parents of each decision node with all its children (if a decision node D has as parent another decision node D', then we also have to make a connection from the parents of D' to the children of D, and this also recursively applies to the parents of D').

The advantage of having a credal set about $\mathbf{X}$ locally expressed is that we can solve the computation of upper and lower conditional probabilities, or the dominance relationship [19], by means of an optimization problem in the set of strategies. If the sets Ω_D are finite, then both approximate [13, 3, 6] and exact [17, 7] algorithms able of solving medium size problems are currently available[2].

3 Credal Networks from the Imprecise Dirichlet Model

In this section we consider that we have a set of variables $\mathbf{X}$, a graph G and a database with N cases in which all the variables are observed (there are no missing data). For each configuration $\mathbf{y}$ of a subset of variables $\mathbf{Y} \subseteq \mathbf{X}$ we can measure the absolute frequency of it in the database $N_{\mathbf{y}}$. We want to estimate a credal set for graph G from the observations in the database.

[2]Most of these algorithms have been initially developed for separately specified nets, but with some small modifications they can be applied to locally specified ones. For example, for the model of this paper we did not need any modification of the algorithm in [7].

The *Imprecise Dirichlet Model* (IDM) [20] was introduced for estimating probability values from a set of observations and has been extensively used by its good theoretical properties and performance in experiments.

Assume the case of one variable X, if we want to estimate the probabilities $P(x)$ with the precise Dirichlet model, we have to assume a vector of positive parameters $(\alpha_x)_{x\in\Omega_X}$. The value $S = \sum_{x\in\Omega_X} \alpha_x$ is called the *equivalent sample size*. If we denote the probability $P(x)$ by θ_x, then the Dirichlet density is proportional to $\prod_x \theta_x^{\alpha_x - 1}$. In these conditions the estimation of the probability $P(x)$ for a future event is equal to $\frac{(N_x+\alpha_x)}{(N+S)}$ (the expected value of the posterior probability given the data).

The Imprecise Dirichlet Model (IDM) considers a set of prior distributions, those obtained by fixing a global sample size S and considering all the vectors of positive parameters $(\alpha_x)_{x\in\Omega_X}$ such that $S = \sum_{x\in\Omega_X} \alpha_x$. This gives rise to an interval estimation (corresponding to all the possible vectors compatible with a given S) of $P(x)$, which is given by:

$$\left[\frac{N_x}{N+S}, \frac{N_x+S}{N+S}\right] \tag{2}$$

Usually a parameter S in the interval $[1, 2]$ is considered, and recently some authors as Bernard [4] advocates for the use of $S = 2$.

When applying the IDM to obtain the credal set of a credal network, this can be done in two ways: local or global. In the local application we obtain a separable credal network. What we do is to apply an IDM to each variable X and each configuration of its parents π_X, considering only the part of the database compatible with configuration π_X, i.e. the cases that for variables $\mathbf{\Pi}_X$ have the same values than in configuration π_X. Then we obtain a local set for each variable and each configuration of its parents: the probabilities satisfying the intervals in equation (2) where the frequencies are measured in the restricted database (same values than the configuration of parent variables). The global credal set is obtained by strong extension (the convex hull of the set of all the probabilities equal to the multiplication of a conditional probability distribution for each variable given its parents, where this conditional probabilities are selected from the local conditional credal sets).

This was the method initially employed, but soon it was noticed that it can produce too wide posterior intervals [21], and a small imprecision in all the conditional probabilities can give rise to high degrees of imprecision in conditional probabilities that are a

function of all these conditional probabilities.

The other possible application is the global one [21]. According to it, the credal set is obtained by considering a global application of the IDM to all the variables $\mathbf{X}$. We consider the credal set given by the probabilities obtained from the IDM and that factorize according to G. When a global precise Dirichlet model is applied to $\mathbf{X}$ with parameters $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{X}}}$, then the estimated probabilities for any conditional probability of a variable X conditional to a configuration $\mathbf{Y}=\mathbf{y}$, coincides with the ones obtained with a Dirichlet model with a vector of parameters$(\alpha_{x,\mathbf{y}})_{x\in\Omega_X}$ which can be obtained from the original vector by adding in the non participating variables. If $\mathbf{Z}$ are the variables in $\mathbf{X}-\mathbf{Y}-\{X\}$, then we have that $\alpha_{x,\mathbf{y}}=\sum_{\mathbf{z}}\alpha_{x,\mathbf{y},\mathbf{z}}$.

When considering a global application of the IDM, the set of all conditional probability distributions for each single variable X given a configuration of its parents π_X is the same than in the local application. But in the global application restrictions in the parameters used in the different conditional probabilities. Imagine that we have two binary variables, X and Y, and that X is a parent of Y. If for the marginal of X we use a Dirichlet distribution with parameters $(\alpha_{x_1},\alpha_{x_2})$, then for the conditional probability of Y given $X=x_1$, we have to use a Dirichlet distribution with parameters $(\alpha_{y_1},\alpha_{y_2})$ with $\alpha_{y_1}+\alpha_{y_2}=\alpha_{x_1}$ and for the conditional probability of Y given $X=x_2$, the parameters has to verify $\alpha_{y_1}+\alpha_{y_2}=\alpha_{x_2}$. So the parameters use in one variable impose restrictions in the parameters used in the rest of variables. As a consequence, the joint credal set is not the one obtained by selecting an arbitrary conditional probability for each variable given its parents and multiplying them. We have to take into account the existing restrictions between the parameters which impose restrictions into the conditional probabilities for the different variables.

In the following section, we will show that it is possible to locally express the associated credal net, but there is an important problem: the decision variables are continuous and so we have to solve an optimization problem with continuous variables, which is not simple in general, and for which we do not know any paper reporting an implementation of a general algorithm to compute upper or lower conditional probabilities. Only Zaffalon [21] has reported an algorithm for the case of a Naive graph to compute the dominance relationship.

What we propose here is a modification of the IDM model that we will call the *extreme IDM*. In the extreme IDM, instead of considering all the prior Dirichlet with $S=\sum_{x\in\Omega_X}\alpha_x$ for a given S, only the extreme ones are considered: one for each $x_0\in\Omega_X$ given by parameters $(\alpha_x)_{x\in\Omega_X}$, where $\alpha_x=S$, if $x=x_0$ and 0.0, otherwise. This density will be called the *extreme density* concentrated in value x_0 with sample size S. These prior densities on the parameters are *improper* densities, i.e. their integral is not equal to 1.0, but infinite. Their use has been justified by the estimation they produce of the posterior probabilities after a sample. Some of them are the limit of proper density functions and have a simpler interpretation. Above density can be considered as the limit when ϵ approaches to 0 of the densities with parameters $(\alpha^{\epsilon}_x)_{x\in\Omega_X}$, where $\alpha^{\epsilon}_x=S$, if $x=x_0$ and ϵ, otherwise. The estimation of the future probabilities will be the limit of the estimation with the proper densities when epsilon tends to 0. When we consider the parameters $\alpha_x=0.0,\forall x\in\Omega_X$, the estimation we obtain for future probabilities coincide with the maximum likelihood estimation (relative frequencies), i.e. $P(x)$ is estimated by N_x/N.

The main fact about the new model is that instead of considering all the infinite densities determined by a simple size S, we only consider the extreme ones, in which all the sample size is concentrated in only one element[3]. This gives rise to one density for each one of the possible value of X.

When considering the extreme IDM for the estimation of future probabilities of a single variable X, what we obtain as estimation for $P(x)$ is the same interval than in formula (2). This is immediate, as the upper and lower limits of the intervals are obtained in the extreme densities. The densities in which the parameters are not concentrated in only one point, produce inner values of the intervals (2).

However, in a credal net we can have differences as we take into account the independence relationships represented by the graph. In general, we obtain intervals which are included into the intervals associated to the use of the global original IDM.

The global application of the extreme IDM with parameter S to a graph G and set of variables $\mathbf{X}$ is given by the credal set which is equal to the convex hull of all the probability distributions that factorizes according to the graph with conditional distributions obtained in the following way:

1. Consider a value $\mathbf{x}_0\in\Omega_{\mathbf{X}}$.
2. For each variable X and each conditional configuration of its parents π_X, estimate the probability distribution of X given this configuration in the following way:

[3]We consider that the use of improper densities is not essential for the extreme model.

(a) If the configuration π_X coincides with $\mathbf{x}_0$ in the set of parents of X then $P(x|\pi_X)$ is equal to $\frac{N_{x,\pi_X}+S}{N_{\pi_X}+S}$ if the value of X in configuration $\mathbf{x}_0$ is equal to x, and equal to $\frac{N_{x,\pi_X}}{N_{\pi_X}+S}$, otherwise; where N_{π_X} is the frequency of configuration π_X in the sample, and N_{x,π_X} the frequency of cases in which we have configuration π_X and $X = x$ in the sample.

(b) If the configuration π_X does not coincide with $\mathbf{x}_0$ in the set of parents of X then $P(x|\pi_X)$ is equal to $\frac{N_{x,\pi_X}}{N_{\pi_X}}$.

What we do is to consider all the extreme densities, one for each value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ given by parameters $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{X}}}$, where $\alpha_{\mathbf{x}_0} = S$ and 0.0, otherwise. With this vector of parameters, all the conditional probabilities are estimated. For a variable, X, and a configuration of its parents, π_X, if $\mathbf{x}_0$ coincides with this configuration in the set of parents of X, then the conditional probability about X is estimated with the extreme density concentrated in the value of X in configuration $\mathbf{x}_0$ with parameter S. If $\mathbf{x}_0$ does not coincide with this configuration in the set of parents of X, then we have to estimate the conditional probability with a vector of values which are all equal to 0.0, i.e. we apply maximum likelihood estimation. If applying the maximum likelihood estimation $N_{\pi_X} = 0$, then the estimation of the probability is not defined. We will consider the uniform distribution in this case[4].

Example 1 *We are going to show the differences between the global IDM model and the extreme IDM model in a very sample case.*

Assume three binary variables X, Y, Z and a single credal network in which X is a parent of Y and Z (as a Naive Bayes in which X is the root node). Consider a sample of size equal to 2 with observations:

X	Y	Z
x_1	y_1	z_1
x_2	y_1	z_1

Assume that we apply the extreme global IDM with global sample size $S = 2$ to estimate the conditional probabilities and we want to compute the upper probability of $X = x_1$ given that $Y = y_1, Z = z_1$. This probability, $P(x_1|y_1, z_1)$ is obtained by maximizing the result of Bayes rule, that taking into account the existing conditional independence relationships can be expressed as:

$$\frac{P(y_1|x_1).P(z_1|x_1).P(x_1)}{P(y_1|x_1).P(z_1|x_1).P(x_1) + P(y_1|x_2).P(z_1|x_2).P(x_2)}$$

The upper value with extreme prior densities is obtained when these probabilities are estimated with parameters $\alpha_{x_1,y_1,z_1} = 2$ and 0.0 otherwise, and the value of the upper probability is 0.75 (the value is obtained by estimating the probabilities with relative frequencies from a sample obtained from the original one by adding two cases in which $X = x_1, Y = y_1, Z = z_1$). This upper limit can be also obtained with another extreme parameter: $\alpha_{x_2,y_2,z_2} = 2$ and 0.0 otherwise.

If we consider the global IDM model, then more sets of parameters are allowed, and not only those concentrated in only one configuration of values. In particular, we can have $\alpha_{x_1,y_1,z_1} = 1, \alpha_{x_2,y_2,z_2} = 1$ and 0.0 otherwise. If we compute the conditional probability using this set of parameters (using relative frequencies to a sample in which two new cases are added: one in which $X = x_1, Y = y_1, Z = z_1$ and other in which $X = x_2, Y = y_2, Z = z_2$) we obtain a value of 0.8, which is the upper limit of the interval. So, in this case, when applying the global model, the upper limit is greater than when using the restricted model.

To give an idea of the differences between the two models, let us generalize above situation: imagine that we have a Naive Bayes model with X as root node and a number n of children variables ($n = 2$ in previous case). Assume that we also have a sample of size 2 similar to the above one (one in which $X = x_1$ and another in which $X = x_2$ and in both of them the first case of the remaining variables is observed), and that we want to compute the upper probability of $X = x_1$ conditioned to the first case of each variable. With variable, there is no difference between the models. With $n = 3$ the difference is very small, and with $n \geq 4$ both models produce again the same result.

4 Local Specification

In this section we will show that credal networks estimated with the IDM can be locally specified. First we will start with the complete model in which it will be necessary to use decision variables with infinite values.

Given a credal network with graph G learned with the IDM with global sample size S, we will consider the following credal network:

- For each variable X with parents $\mathbf{\Pi}_X$ in the graph, consider a decision variable D_X, which will be a parent of X. This variable

[4] Any conditional probability distribution will give rise to the same joint distribution, as these values are going to be multiplied by 0.0.

will have as set of values the set of vectors $(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}$, where $\alpha_{x,\pi_X} > 0$ and $\sum_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}} \alpha_{x,\pi_X} = S$.

- For each configuration π_X and vector $(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}$, the conditional probability of X is given by:

$$P(x|\pi_X,(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}) = \frac{N_{x,\pi_X}+\alpha_{x,\pi_X}}{N_{\pi_X}+S_{\pi_X}}$$

  where $S_{\pi_X} = \sum_{x\in\Omega_X} \alpha_{x,\pi_X}$.

- Consider an order of the variables which is compatible with the graph G. For each variable, X, in this order consider the set $\mathbf{T}_X = \mathbf{\Pi}_X \cup \{X\}$. Compute the intersections $\mathbf{R}_{X,Y} = \mathbf{T}_X \cap \mathbf{T}_Y$ with all the variables Y preceding X in the graph. Make as parents of D_X all the variables D_Y, for which $\mathbf{R}_{X,Y}$ is a non empty maximal set (there is not another $\mathbf{R}_{X,Y'}$ including it).

- f_{D_X} is defined as a function that associates to each configuration of its parents the set of possible values for D_X. This will be done, by determining the set of possible values for each one of its parents and then taking the intersection for all the parents. For a parent variable D_Y and a vector belonging to its domain $(\beta_{\mathbf{y}})_{\mathbf{y}\in\Omega_{\mathbf{T}_Y}}$, the set of possible values for D_X will be equal to the set of vectors $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{T}_X}}$ such that $\sum_{\mathbf{u}} \beta_{\mathbf{y}} = \sum_{\mathbf{v}} \alpha_{\mathbf{x}}$, where $\mathbf{U} = \mathbf{T}_Y - \mathbf{R}_{X,Y}$, $\mathbf{V} = \mathbf{T}_X - \mathbf{R}_{X,Y}$, i.e. the results of adding the vectors in the non common variables coincide.

With this procedure we only estimate conditional probabilities, considering that the joint probabilities can be obtained by multiplication. So all the probabilities factorize according to G.

In this local specification, for each variable X, the domain for the decision variable D_X is the set of possible parameters for the prior Dirichlet distributions if the joint probability has global parameter S. The conditional probability is determined for each parameter vector, by doing the corresponding estimation from the database and the given prior distribution. Finally, the role of functions f_{D_X} is to keep consistency among parameters taking into account the existing restrictions in the global application of the IDM. For that, we relate the vectors of parameters D_X and D_Y if the corresponding sets of variables $\mathbf{T}_X$ and $\mathbf{T}_X$ have non-empty intersection. Consistency is achieved if the marginalization of the vectors of parameters on the intersection of both sets of variables is the same, where the marginalization is computed by adding in the variables not in the intersection (in the same way than when computing a marginal probability). This is based on the properties of the Dirichlet densities (see [14, 4]).

The main problem of this description as a local network is that variables D_X take values in a continuous infinite set of parameters. This makes infeasible the application of existing algorithms for computing upper and lower conditional probabilities, which are designed for categorical variables. In the following, we will show that the use of the extreme IDM gives rise to a credal network that can be locally specified in a simple way by introducing categorical decision variables.

In the extreme IDM we have a prior density for each value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$, so in the posterior credal set after observing the database we will have a joint probability for each one of these values. We have to introduce decision auxiliary variables able of representing these values. This will be done by considering a decision variable R_X for each variable X with the same set of values than X: Ω_X. The set of values of variables $R_X, X \in \mathbf{X}$, will represent the configuration $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ in which the parameter S is concentrated.

Decision variables, R_X, do not have parents.

If we have variable X with parents $\mathbf{\Pi}_X$ in graph G, we add links from each variable R_Y where $Y = X$ or $Y \in \mathbf{\Pi}_X$ to X (we extend the parents of X by adding its decision variable and the decision variables of its parents). Let us call $\mathbf{R}_{\mathbf{\Pi}_X}$ the set of variables R_Y where $Y \in \mathbf{\Pi}_X$, and as usual (in lowercase), $\mathbf{r}_{\mathbf{\Pi}_X}$ will represent a configuration of this set of variables. The conditional probability of a variable X given $\mathbf{\Pi}_X = \pi_X, R_X = r_X$ and $\mathbf{R}_{\mathbf{\Pi}_X} = \mathbf{r}_{\mathbf{\Pi}_X}$ is computed as follows:

- If for one variable Y in Π_X, the value of Y in configuration π_X is not equal to the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, then

$$P(x|\pi_X,\mathbf{r}_{\mathbf{\Pi}_X},r_X) = \frac{N_{x,\pi_X}}{N_{\pi_X}} \tag{3}$$

  where the conditional distribution is the uniform if $N_{\pi_X} = 0$.

- If for any variable Y in Π_X, the value of Y in configuration π_X is the same than the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, and the value of X is the same than the value of R_X ($x = r_X$), then

$$P(x|\pi_X,\mathbf{r}_{\mathbf{\Pi}_X},r_X) = \frac{N_{x,\pi_X}+S}{N_{\pi_X}+S} \tag{4}$$

- If for any variable Y in Π_X, the value of Y in configuration π_X is the same than the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, and the value of X is not equal to the value of R_X ($x \neq r_X$), then

$$P(x|\pi_X, \mathbf{r}_{\mathbf{\Pi}_X}, r_X) = \frac{N_{x,\pi_X}}{N_{\pi_X} + S} \qquad (5)$$

It is immediate that this specification determines the same credal set over G as the one defined in Section 3, taking into account that the values of variables $R_X, X \in \mathbf{X}$, represent the value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ in which the extreme Dirichlet distribution is concentrated.

One important problem of this representation is that the number of variables in each conditional probability is duplicated, and as the size of conditional tables is exponential in the number of variables, then we can have tables of quadratic size with respect to the size of precise conditional probability tables in G. However, the size of the conditional probabilities can be smaller if we use an appropriate representation. In this paper we consider the use of the *probability tree* representation [9, 18, 7].

A *probability tree* $\mathcal{T}$ is a directed labelled tree, where each internal node represents a variable and each leaf represents a non-negative real number. Each internal node has one outgoing arc for each state of the variable associated with that node. The *size* of a tree $\mathcal{T}$, denoted by $size(\mathcal{T})$, is defined as its number of leaves.

A probability tree $\mathcal{T}$ on variables $\mathbf{Y}$ represents a potential (a joint or conditional probability distribution) in these variables $h : \Omega_{\mathbf{Y}} \to I\!R_0^+$ if for each $\mathbf{y} \in \Omega_{\mathbf{Y}}$ the value $h(\mathbf{y})$ is the number stored in the leaf node that is reached by starting from the root node and selecting the child corresponding to the value of Y in $\mathbf{y}$ for each internal node labelled Y.

A probability tree is usually a more compact representation of a potential than a table. This is illustrated in Figure 1, which displays a potential h and its representation using a probability tree. The tree contains the same information as the table, but using only five values instead of eight. Furthermore, trees enable even more compact representations to be obtained in exchange for loss of accuracy. This is achieved by pruning certain leaves and replacing them by the average value, as shown in the second tree in Figure 1.

All the necessary operations to compute with probability potentials in credal networks can be directly carried out in the probability tree representation, without transforming it into a table [9, 18, 7]. In the following we give the probability tree representation of the conditional probability distribution of a variable X in a local specification of an extreme IDM credal network. It is built with the following procedure $BuildTree(\mathcal{T}, \mathbf{Z}, \mathbf{y})$, where $\mathcal{T}$ is the tree we are building, $\mathbf{Z}$ is the set of variables from $\mathbf{\Pi}_X$ we have to consider and $\mathbf{y}$ the configuration of the variables already introduced in the tree and that corresponds to the path from the root to the present tree $\mathcal{T}$. Initially the procedure is called with $\mathcal{T}$ empty, $\mathbf{y}$ empty, and $\mathbf{Z} = \mathbf{\Pi}_X$. It performs the following steps:

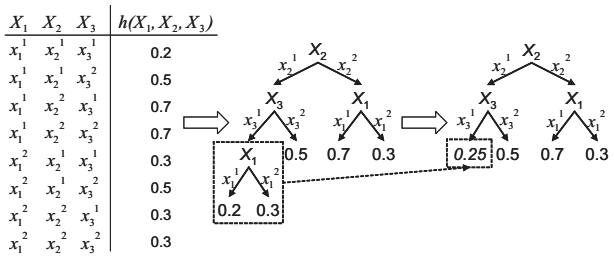

Figure 1: A probability potential h, its representation as a probability tree and its approximation after pruning various branches

- Take a variable $Z \in \mathbf{Z}$, branch $\mathcal{T}$ by Z, then branch also all its children by variable R_Z. Remove Z from $\mathbf{Z}$.

- For each one of the leaves $\mathcal{T}'$ of the resulting tree, consider the configuration $\mathbf{y}'$ equal to $\mathbf{y}$ plus the value of $Z = z$ corresponding to this leaf.

  - If for this leaf, the values of Z and R_Z are the same, then call recursively to $BuildTree(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$, continuing with the construction of the tree.
  - If for this leaf, the values of Z and R_Z are different, then call recursively to $BuildTree2(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$.

- If $\mathbf{Z} = \emptyset$, then the tree is finished by branching by X and all its children by R_X. For all the resulting leaves we store the conditional probability of $X = x$ given the configuration $\mathbf{y}$ (that now is a complete configuration of its parents). This probability is computed with expressions (4) or (5) depending on whether the leaf is obtained for the same value of R_X and X or for different values of these variables (in the corresponding expressions $\pi_{\mathbf{X}}$ is the current configuration $\mathbf{y}$).

$BuildTree2(\mathcal{T}, \mathbf{Z}, \mathbf{y})$ is a simpler procedure that obtains the conditional probability when the configuration of the parents is different of the configuration of the decision variables (by maximum likelihood):

- If $\mathbf{Z} = \emptyset$, then the tree is finished by branching by X and in its leaves we store the conditional probability of $X = x$ given the configuration $\mathbf{y}$. This probability is computed with expressions (3). As above, in the corresponding expression, $\pi_{\mathbf{X}}$ is the current configuration $\mathbf{y}$.
- If $\mathbf{Z} \neq \emptyset$, then take $Z \in \mathbf{Z}$, branch the tree by Z. Remove Z from $\mathbf{Z}$.
- For each one of the leaves $\mathcal{T}'$ of the resulting tree, consider the configuration $\mathbf{y}'$ equal to $\mathbf{y}$ plus the value of $Z = z$ corresponding to this leaf. Then, call recursively to $BuildTree2(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$.

As example, assume two binary variables X and Y for which we have the following table of frequencies:

	Y=0	Y=1
X=0	1	3
X=1	2	1

The resulting tree for the conditional probability of Y given X and $S = 2$, is given in Figure 2.

It can be shown that if n is the size of a table of X given $\mathbf{\Pi}_X$, then the number of leaves of this tree representation will be $n.\left(|\Omega_X| + \sum_{Y \in \mathbf{\Pi}_X}(|\Omega_Y| - 1)\right)$. In this example, we have represented a table of size $n = 4$ with a tree of 12 leaves. This is obtained from the following fact: the number of cases in which the value of the decision variables coincides with the conditioning configuration is n, and each one of then is branched by R_X of cardinal $|\Omega_X|$. Now, each conditioning variable Y defines $(|\Omega_Y|-1)$ branches in which the complete probability table of size n is estimated by maximum likelihood (no coincidence of the conditioning variables and the value of parents variables).

5 Experiments

The local estimation algorithm with the extreme IDM has been implemented in Elvira environment [10] producing the local specification at the same time. With this we have been able of applying the existing algorithms for credal networks as the ones described in [7] which have also been implemented in Elvira. We have done a very simple and preliminary experiment. We have selected a Naive Bayes graph with a class variable and 10 attributes (all binary variables). We have simulated samples with different sizes (from 10 to 1000). We have selected a Naive Bayes, as with no independencies the results are the same than with the complete IDM. So, we do the experiments with a graph in which many independence relationships among the variables are represented. In these conditions, we have estimated the locally specified credal network and computed the conditional probability for the class when all the attributes have been observed. We have considered 3 different situations: the observations are random, for each attribute we observe the most frequent value, and finally the case in which for each attribute we observe the least frequent value. We report the length of the computed posterior intervals. The intervals are computed with a simple exact deletion algorithm with probability trees (see details in [8]). The sample generation is repeated 50 times for each sample size and set of observations and in Table 1 we report the average and standard deviations of the lengths interval probabilities (Evi1 corresponds to random observations, Evi2 to observing the most frequent cases, and Evi3 to observing the least frequent cases).

We observe that the intervals decrease in size when the sample size is increased. Also when we observe the most frequent values the intervals are smaller than when the least frequent values are observed. Random observations give rise to intermediate intervals. In this stage, we can not say much more, except that the intervals are very wide with the smaller sample size (10) but that the imprecision is small with sample sizes of 1000. To our opinion, this imprecision is *reasonable*.

6 Conclusions

In this paper, we have proposed a new model to estimate probabilities for a credal network. This model is a restriction of the general IDM, where only the extreme densities are considered. Its main advantage of the new one is that the resulting credal network allows a simple local specification with categorical decision variables and then it is suitable for the application of existing algorithms for the computation of posterior intervals or dominance relationships.

We have shown the results of the imprecision in the intervals in some very preliminary experiments. But, really it would be necessary to carry out more tests to see the behaviour in real classification problems and to study the differences with the complete IDM. We believe that the differences between the two models are less important than the selection of parameter S and, at present, there is no general agreement about which is the most suitable value of S. We do not expect meaningful differences between them. We have also to take into account that it is possible that the fact that the new model is more restrictive could be compensated with a greater S (using $S = 2$ in all the situations).

Another point we would like to raise is that, though the IDM is a widely accepted model with very good

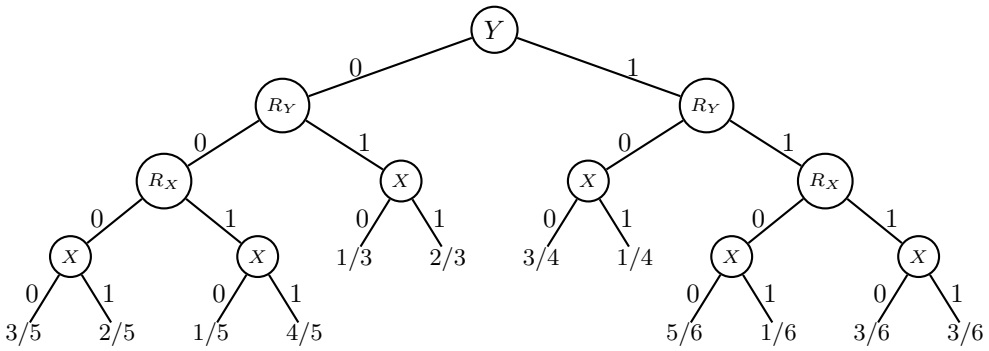

Figure 2: Tree representation of X given Y, R_Y, R_X

	Evi1		Evi2		Evi3	
Iter	aver.	dev.	aver.	dev.	ave.	dev.
10	0.948286	0.051853	0.608757	0.251884	0.999999	4.2475E-5
20	0.814695	0.172520	0.382576	0.247139	0.983606	0.115112
50	0.573315	0.144312	0.062716	0.067052	0.968920	0.121630
100	0.326327	0.126361	0.010081	0.007831	0.869229	0.238078
200	0.170638	0.053589	0.002283	0.001472	0.656434	0.209858
500	0.063706	0.017014	6.9051E-4	3.4956E-4	0.366218	0.123964
1000	0.032087	0.006731	3.0723E-4	1.1755E-4	0.181275	0.051277

Table 1: Average lengths standard deviations for the posterior conditional intervals ($S = 2$)

theoretical properties, it is not the only possible model for being used as prior information. In the problem we have studied in this paper, we see that the general model has computational problems. We also experimented difficulties with the global IDM when studying independence in [1] and we considered a different more restrictive IDM as it was impossible to make decisions about independence with the original IDM using a generalization of Bayesian scores (there was no dominance even with very large samples). So it is important to investigate alternative models for prior information, comparing their behaviour in solving different problems.

Acknowledgments

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

References

[1] J. Abellán and S. Moral. A new score for independence based on the imprecise Dirichlet model. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications (ISIPTA '05)*, pages 1–10. SIPTA, 2005.

[2] A. Antonucci and M. Zaffalon. Locally specified credal networks. In M. Studen and Vomlel, editors, *Proceedings of the third European Workshop on Probabilistic Graphical Models*, pages 25–34. Action M Agency, 2006.

[3] A. Antonucci, M. Zaffalon, J. S. Ide, and F. G. Cozman. Binarization algorithms for approximate updating in credal nets. In L. Penserini, P. Peppas, and A. Perini, editors, *Proceedings of the third European Starting AI Researcher Symposium*, pages 120–131. IOS Press, 2006.

[4] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39:123–150, 2005.

[5] A. Cano, J.E. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing. In *Proceedings of the Fith International Conference IPMU'94*, pages 4–8, Paris, 1994.

[6] A. Cano, J.M. Fernandez-Luna, and S. Moral. Computing probability intervals with simulated annealing and probability trees. *Journal of Applied Non-Classical Logics*, 12:151–171, 2002.

[7] A. Cano, M. Gómez, S. Moral, and J. Abellán. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44:261–280, 2007.

[8] A. Cano and S. Moral. Using probability trees to compute marginals with imprecise probabilities.

International Journal of Approximate Reasoning, 29:1–46, 2002.

[9] A. Cano, S. Moral, and A. Salmerón. Penniless propagation in join trees. *International Journal of Intelligent Systems*, 15:1027–1059, 2000.

[10] Elvira Consortium. Elvira: An environment for probabilistic graphical models. In J.A. Gámez and A. Salmerón, editors, *Proceedings of the 1st European Workshop on Probabilistic Graphical Models*, pages 222–230, 2002.

[11] F.G. Cozman. Robustness analysis of bayesian networks with global neighborhoods. Technical Report CMU-RI-TR96-42, Carnegie Mellon University, 1996.

[12] F.G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[13] E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106:77–107, 1998.

[14] D. Geiger and D. Heckerman. A characterization of the dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25.

[15] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[16] J.C.F. Rocha and F.G. Cozman. Inference with separately specified sets of probabilities in credal networks. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 430–437. Morgan & Kaufmann, 2002.

[17] J.C.F. Rocha and F.G. Cozman. Inference in credal networks with branch-and-bound algorithms. In *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications (ISIPTA03)*, pages 482–495, 2003.

[18] A. Salmerón, A. Cano, and S. Moral. Importance sampling in bayesian networks using probability trees. *Computational Statistics and Data Analysis*, 34:387–413, 2000.

[19] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[20] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[21] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T.L. Fine, and T. Seidenfeld, editors, *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, 2001.

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