# Tree-based credal networks for classification *

Marco Zaffalon (zaffalon@idsia.ch)
*IDSIA*
*Galleria 2, CH-6928 Manno (Lugano), Switzerland*

Enrico Fagiuoli (fagiuoli@disco.unimib.it)
*DiSCo, Università degli Studi di Milano-Bicocca*
*Via Bicocca degli Arcimboldi 8, I-20126 Milano, Italy*

**Abstract.** Bayesian networks are models for uncertain reasoning which are achieving a growing importance also for the data mining task of classification. Credal networks extend Bayesian nets to sets of distributions, or credal sets. This paper extends a state-of-the-art Bayesian net for classification, called tree-augmented naive Bayes classifier, to credal sets originated from probability intervals. This extension is a basis to address the fundamental problem of prior ignorance about the distribution that generates the data, which is a commonplace in data mining applications. This issue is often neglected, but addressing it properly is a key to ultimately draw reliable conclusions from the inferred models. In this paper we formalize the new model, develop an exact linear-time classification algorithm, and evaluate the credal net-based classifier on a number of real data sets. The empirical analysis shows that the new classifier is good and reliable, and raises a problem of excessive caution that is discussed in the paper. Overall, given the favorable trade-off between expressiveness and efficient computation, the newly proposed classifier appears to be a good candidate for the wide-scale application of reliable classifiers based on credal networks, to real and complex tasks.

**Keywords:** Credal classification, prior ignorance, imprecise Dirichlet model, credal sets, naive Bayes classifier, naive credal classifier, tree-augmented naive Bayes classifier, imprecise probabilities, Bayesian networks, data mining.

* This paper extends work published in the Proceedings of the 8th Information Processing and Management of Uncertainty in Knowledge-Based Systems Conference [15].

## 1. Introduction

Classification has a long tradition in statistics and machine learning [13]. The purpose of classifiers is to predict the unknown categorical *class* $C$ of objects described by a vector of *features* (or *attributes*), $A_1, \ldots, A_n$. Classifiers capture the relationship between $C$ and $(A_1, \ldots, A_n)$ by examining past joint realizations of the class and the features. Such abstract description of classification is easily mapped to a number of concrete applications. For example, medical diagnosis can be regarded as classification, the symptoms and the medical tests being the features and the class representing the possible diseases. Many other application domains exist, as image recognition, fraud detection, user profiling, text classification, etc.

Traditional classification is typically represented within Bayesian decision theory, which formally solves the prediction problem if the distribution that generates the data is known. Unfortunately, in practice there is usually little information about the distribution, apart from that carried by the data. This is even more radical when we come to data mining, which is regarded as the discipline that produces models from data alone, i.e. in the absence of any other form of knowledge. To produce reliable conclusions from the inferred models, the issue of learning from an initial state of ignorance must be properly addressed. This is also a fundamental methodological problem. Walley provides compelling criticism about the consolidated and most popular methods to cope with the prior ignorance problem (e.g., the Bayesian approach, see Section 2.2), showing that they are subject to largely arbitrary choices [35]. With respect to classification, it follows that traditional classifiers should not be regarded as reliable tools when there exist conditions of prior ignorance, especially when the learning sample is small[1]. Walley also proposes a new inferential method called the *imprecise Dirichlet model* (or IDM, see Section 2.2) that is able to reliably learn in conditions of prior ignorance. The application of the IDM to classification leads to reliable classifiers based on interval probability (or, more generally, *imprecise probability*), which have recently appeared in literature [39, 1, 30, 33, 40].

This paper proposes a new model for interval-probability classification based on the IDM, called *tree-augmented naive credal classifier* (TANC, Section 3). TANCs extend the precise-probability classifier called *tree-augmented naive* (TAN, see [18]). TANs are good classifiers, competitive with classifiers such as Quinlan's classification tree *C4.5* [32]. The extension involves replacing the model probabilities with in-

[1] Note that "small" and "large" depend on the sample space. Also a data set with, say, $10^6$ records is small if the sample space is sufficiently complex.

tervals of probability inferred by the IDM from data, and developing a classification procedure that computes the exact implications of having to work with intervals (see Section 3.2, Appendices A and B). The computational complexity of such a classification procedure is linear in the number of attributes, as shown in Section 3.3. This result enables TANCs to be used for large-size real domains.

We have tested the TANC model on eight real and artificial data sets, evaluating its prediction accuracy and robustness on new data (Section 4). The experiments show that TANCs share the good prediction capability with TANs. In comparison with the latter ones, TANCs are shown to be more reliable as the predictions are robust also when the sample only conveys limited information about a domain. The experiments also highlight that TANCs can suffer from an excessive caution of the inferences. Section 4.3 discusses this problem and proposes a general way to address it.

Two observations are worth considering. First, imprecise probability-based classifiers like TANCs generalize traditional classifiers in that they map objects to sets of classes (i.e., not always to single classes) as a natural consequence of the imprecise knowledge about probabilities. The partial indeterminate outputs are the key to obtain reliable answers; the less knowledge in the data, the larger the set of classes, and vice versa. We call *credal classification* (Section 2.4) this more general way to do classification, to emphasize that the generalization follows from adopting sets of probability distributions, also called *credal sets* after Levi [25]. (Probability intervals give rise to a special type of credal sets, see Section 2.1.) Accordingly, we refer to the new classifiers as *credal classifiers*.

Second, from a modelling point of view, TANCs can be regarded as a special case of *credal networks* [9], i.e. *Bayesian networks* [31] extended to manage credal sets (Section 2.3). Bayesian nets are probabilistic graphical models in which the nodes of the graph represent random variables and the arcs represent direct dependencies between variables. TANCs, as well as TANs, are classifiers that allow the dependencies between attributes conditional on the class to take the graphical form of a tree (see Figure 1 for an example), i.e. a graph in which each node has at most one parent.

This structure modelling is substantially more expressive than what existing credal network-based classifiers do [39, 40, 33]. Moreover, there is little hope that more general graphical structures than trees will allow efficient computation with credal nets. The computation with credal networks based on polytrees, i.e. the minimal graphical generalization of trees, is *NP-hard* [17].

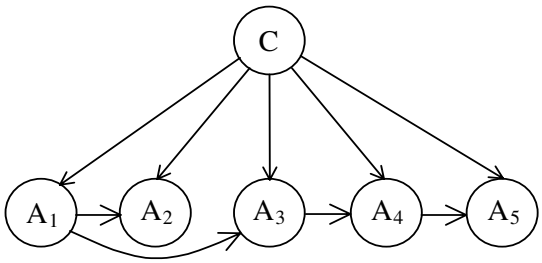


*Figure 1.* A simple TAN graphical structure.

## 2. Methods

### 2.1. Credal sets and probability intervals

In this paper a credal set is a closed convex set of probability mass functions [25]. Credal sets generalize probability theory, as well as probability intervals; possibility measures, belief functions, lower and upper probabilities and previsions are also special cases of credal sets. Credal sets are part of the wider theory of imprecise probability [34] (see http://www.sipta.org for up-to-date information), to which the reader interested in the foundations is referred.

In this paper we will restrict the attention to random variables which assume finitely many values (also called *discrete* or *categorical* variables). Denote by $\mathcal{X}$ the finite sample space for a discrete variable $X$. Let $x$ be the generic element of $\mathcal{X}$. Denote by $P(X)$ the mass function for $X$ and by $P(x)$ the probability of $x \in \mathcal{X}$. $\mathcal{P}_X$ denotes a generic credal set for $X$.

For any event $\mathcal{X}' \subseteq \mathcal{X}$, let $\underline{P}(\mathcal{X}')$ and $\overline{P}(\mathcal{X}')$ be the *lower and upper probability* of $\mathcal{X}'$, respectively. These are defined as follows:

$$\begin{aligned} \underline{P}(\mathcal{X}') &= \min_{P \in \mathcal{P}_X} P(\mathcal{X}') \\ \overline{P}(\mathcal{X}') &= \max_{P \in \mathcal{P}_X} P(\mathcal{X}'). \end{aligned}$$

Lower and upper expectations can be defined similarly. Note that a set of mass functions, its convex hull and its set of *vertices* (also called *extreme mass functions*, i.e. mass functions that cannot be expressed as convex combination of other mass functions in the set) produce the same lower and upper expectations and probabilities. This provides a justification to restrict the attention to convex sets.

Conditioning with credal sets is done by element-wise application of Bayes rule. The posterior credal set is the union of all posterior

mass functions. Denote by $\mathcal{P}_X^y$ the set of mass functions $P(X|Y=y)$, for generic variables $X$ and $Y$. As far as (conditional) independence is concerned, in this paper we adopt the concept of *strong independence*[2]. Two variables are strongly independent when every vertex in $\mathcal{P}_{(X,Y)}$ satisfies stochastic independence of $X$ and $Y$, i.e. for every extreme mass function $P \in \mathcal{P}_{(X,Y)}$, $P(x|y) = P(x)$ and $P(y|x) = P(x)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. See also [28] for a complete account of different strong independence concepts and [10] for a deep analysis of strong independence.

Let $\mathbb{I}_X = \{\mathbb{I}_x : \mathbb{I}_x = [l_x, u_x], 0 \leq l_x \leq u_x \leq 1, x \in \mathcal{X}\}$ be a set of probability intervals for $X$. The credal set originated by $\mathbb{I}_X$ is $\{P(X) : P(x) \in \mathbb{I}_x, x \in \mathcal{X}, \sum_{x\in\mathcal{X}} P(x) = 1\}$. $\mathbb{I}_X$ is said *reachable* or *coherent* if $u_{x'} + \sum_{x\in\mathcal{X}, x\neq x'} l_x \leq 1 \leq l_{x'} + \sum_{x\in\mathcal{X}, x\neq x'} u_x$, for all $x' \in \mathcal{X}$. $\mathbb{I}_X$ is coherent if and only if the related credal set is not empty and the intervals are tight, i.e. for each lower or upper bound in $\mathbb{I}_X$ there is a mass function in the credal set at which the bound is attained [5, 34]. Coherent intervals produce, by means of the related credal sets, upper and lower probabilities that are a special case of *Choquet capacities* of order two [5]. For Choquet capacities of order two it holds that, given two mutually exclusive events $\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}$, there is always [36] a mass function $P$ in the related credal set for which

$$P(\mathcal{X}') = \underline{P}(\mathcal{X}'), P(\mathcal{X}'') = \overline{P}(\mathcal{X}''). \tag{1}$$

We will use this property in Section 3.2.

### 2.2. The imprecise Dirichlet model

Consider a random sample, i.e. a set of values $x \in \mathcal{X}$ of the discrete random variable $X$, drawn independently with chances $\theta_x$, $x \in \mathcal{X}$. The chances $\theta_x$, i.e. the parameters of an underlying multinomial distribution, are rarely known in practice, and here we assume that all the information about them is represented by the sample. To approximate the actual chances from the data one can use methods of statistical inference.

The Bayesian approach [4] to the statistical inference regards the parameters of an unknown distribution as random variables. The uncertainty about the parameters prior to the availability of the sample is modelled by a density function, called *prior*. This is done also when there is little or no prior knowledge about the chances by using special so-called *noninformative* priors, e.g. the *uniform* prior. After observing

[2] Here we follow the terminology introduced by Cozman. Note that other authors use different terms [8].

the data, the chosen prior is updated by Bayes theorem to a new density called *posterior*. A parameter of the original unknown distribution can then be approximated by taking its expectation with respect to the posterior. In the case of multinomial samples, the Dirichlet density is the traditional choice for the prior. The Dirichlet $(s, \boldsymbol{t})$ density for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the vector of the chances and $\boldsymbol{t}$ is the vector of the $t_x$ hyperparamenters ($x \in \mathcal{X}$), is

$$\boldsymbol{\pi}(\boldsymbol{\theta}) \propto \prod_{x \in \mathcal{X}} \theta_x^{st_x - 1}, \tag{2}$$

where $s > 0$, $0 < t_x < 1$ for each $x \in \mathcal{X}$, $\sum_{x \in \mathcal{X}} t_x = 1$, and the proportionality constant is determined by the fact that the integral of $\boldsymbol{\pi}(\boldsymbol{\theta})$ over the simplex of possible values of $\boldsymbol{\theta}$ is 1. The constant $s$ is chosen arbitrarily and determines the weight of the prior towards the number of units in the sample. The larger $s$, the larger the number of units needed to smooth the effect of the prior information.

Literature presents strong criticism about the above modelling in the frequent case of prior ignorance. In particular, Walley shows that the use of a single prior density to model ignorance is unjustified and can ultimately lead to fragile models and unreliable conclusions. Walley proposes that prior ignorance be modelled by *a set* of prior densities. In the case of multimomial samples, the densities are again Dirichlet: in particular, all the densities obtained by fixing $s$ in (2) and letting the $t$-hyperparameters take all the possible values in their domain of definition. The resulting model is called *imprecise Dirichlet model* (or IDM [35]). By updating the set of priors to a set of posteriors conditional on the data, the IDM allows posterior inference to be realized similarly to the Bayesian case, with an important difference: the IDM leads to lower and upper expectations for a chance. These are achieved by (i) computing the expected value of the chance with respect to the posterior obtained for generic value of $\boldsymbol{t}$ and (ii) by minimizing and maximizing this quantity over $\boldsymbol{t}$'s domain of definition.

In summary, in the absence of prior knowledge the IDM still produces a reliable posterior estimate for a chance in the form of an interval whose extremes are the above lower and upper expectations. This interval estimate for $x$ is given by

$$\left[\frac{\#(x)}{N+s}, \frac{\#(x)+s}{N+s}\right] \tag{3}$$

where $\#(x)$ counts the number of units in the sample in which $X = x$, $N$ is the total number of units, and $s$ is like in (2). The IDM lower and upper probabilities can also be interpreted as lower and upper bounds on the relative frequency of $x$ if we imagine that there are $s$ hidden

observations as well as the $N$ revealed ones. With the IDM, $s$ is also interpreted as a degree of caution of the inferences and is usually chosen in the interval $[1, 2]$ (see [35] for discussion about the value of $s$).

Expression (3) permits to infer probability intervals from multinomial samples and hence to infer credal sets, given the relationship between them highlighted in Section 2.1. Note that the sets of probability intervals obtained using (3) are reachable. It is also worth noting that the above interval is independent of the definition of the sample space. This is a desirable property that avoids the problems due to the arbitrariness in the definition of $\mathcal{X}$.

In this paper all the model probabilities inferred from data will be of the type (3).

### 2.3. Credal networks

A Bayesian network is a pair composed of a directed acyclic graph and a collection of conditional mass functions. A node in the graph is identified with a random variable $X_i$ (we use the same symbol to denote them and we also use "node" and "variable" interchangeably). Each node $X_i$ holds a collection of conditional mass functions $P(X_i \,|pa\,(X_i))$, one for each possible joint state $pa\,(X_i)$ of its direct predecessor nodes (or *parents*) $Pa(X_i)$.

Bayesian nets satisfy the *Markov condition*: every variable is independent of its nondescendant non-parents given its parents. From the Markov condition, it follows [31] that the joint mass function $P(\boldsymbol{X}) = P(X_1, \ldots, X_t)$ over all the $t$ variables of the net is given by

$$P(x_1, \ldots, x_t) = \prod_{i=1}^{t} P(x_i \,|pa\,(X_i)\,) \quad \forall (x_1, \ldots, x_t) \in \times_{i=1}^{t} \mathcal{X}_i, \qquad (4)$$

where $pa\,(X_i)$ is the assignment to the parents of $X_i$ consistent with $(x_1, \ldots, x_t)$.

Similarly to the definition above, a credal network is a pair composed of a directed acyclic graph and a collection of conditional credal sets. The graph is intended to code strong dependencies, according to Section 2.1: every variable is strongly independent of its nondescendant non-parents given its parents. A generic variable, or node of the graph, $X_i$ holds the collection of credal sets $\mathcal{P}_{X_i}^{pa(X_i)}$, one for each possible joint state $pa\,(X_i)$ of its parents $Pa(X_i)$.

We assume that the credal sets of the net are *separately specified* [17, 34]: this implies that selecting a mass function from a credal set does not influence the possible choices in others. This assumption is natural within a sensitivity-analysis interpretation of credal nets. In

the present context, the assumption helps in keeping the computational complexity of algorithms low.

Now we can define the *strong extension* [10], i.e. the set $\mathcal{P}$ of joint mass functions associated with a credal net:

$$\mathcal{P} = CH\left\{P(\boldsymbol{X}) \text{ as from (4)} : P(X_i\,|pa\,(X_i)) \in \mathcal{P}_{X_i}^{pa(X_i)}, i = 1, \ldots, t\right\} \tag{5}$$

where $CH$ denotes the convex hull operation. In other words, $\mathcal{P}$ is the convex hull of the set of all the joint mass functions as from (4), that are obtained by selecting conditional mass functions from the credal sets of the net in all the possible ways.

Strong extensions can have a huge number of extreme mass functions. Indeed, the computation of lower and upper probabilities with strong extensions is NP-hard [17][3] also when the graph is a polytree (with binary random variables the computation is still polynomial [14]). Polytrees are the minimal graphical generalization of trees with the characteristic that forgetting the direction of arcs, the resulting graph has no undirected cycles. So this paper shows that the representation level of trees is the more expressive one that permits easy computation. Notably, the expressiveness of trees was shown to be a good approximation to that of polytrees [11].

### 2.4. Credal classification

Bayesian networks can be used to do classification. Consider a network with the class $C$ and the attributes $A_1, \ldots, A_n$ as nodes. According to Bayesian decision theory, the predicted class for a new instance of the attributes $(a_1, \ldots, a_n) = \boldsymbol{a}$ should be $c^* = \arg\max_{c \in \mathcal{C}} P(c|\boldsymbol{a})$. The computation of $P(c|\boldsymbol{a})$ is well-known with Bayesian nets and is called "updating" or "inference" (the latter should not be confused with statistical inference).

Credal networks can be used for similar purposes, but with a major difference. For each $c \in \mathcal{C}$, we have an interval for $P(c|\boldsymbol{a})$ rather than a number: $\left[\underline{P}(c|\boldsymbol{a}), \overline{P}(c|\boldsymbol{a})\right]$, where $\underline{P}(c|\boldsymbol{a}) = \min_{P \in \mathcal{P}} P(c|\boldsymbol{a})$, $\overline{P}(c|\boldsymbol{a}) = \max_{P \in \mathcal{P}} P(c|\boldsymbol{a})$, and $\mathcal{P}$ is the strong extension of the net. This prevent us in general from having a total order on the classes which is needed

[3] However, it should be observed that Ferreira da Rocha and Cozman's result is proved for the subset of polytrees in which the local credal sets are convex hulls of degenerate mass functions that assign all the mass to one elementary event. As such, it does not tell anything on the complexity to work with the more realistic case of polytrees whose credal sets are made by mass functions that assign positive probability to any event.

to have an optimal class $c^*$. We are only left with a partial order, which depends on the chosen dominance criterion. We use *credal dominance* [40, Def. 4.1].

We say that class $c'$ *credal-dominates* class $c''$ if and only if $P(c'|\boldsymbol{a}) > P(c''|\boldsymbol{a})$ for all mass functions $P \in \mathcal{P}$ so that $P(\boldsymbol{a}) > 0$.

Credal dominance is a special case of *strict preference* [34, Sect. 3.7.7] justified by Walley on the basis of behavioral arguments. It was previously proposed by Seidenfeld in the commentary of a paper by Kyburg [23, p. 260, P-III′]. It is easy to shown [40, Def. 4.1] that $c'$ credal-dominates $c''$ if and only if

$$\min_{P \in \mathcal{P}: P(\boldsymbol{a})>0} \left[P(c', \boldsymbol{a}) - P(c'', \boldsymbol{a})\right] > 0. \tag{6}$$

The classification procedure follows easily. It takes each pair of classes, tests credal dominance by (6) and discards the dominated class, if any. The output of the classifier is the set of classes that are not dominated. Thus credal networks give rise to credal classifiers.

Credal classification was introduced in [38] and discussed more widely in [40]. Let us stress that a credal classifier is not only a new classifier, it implements a new idea of classification based on the possibility to (partially or totally) suspend the judgment.

This characteristic of credal classification should not be confused with the *rejection option* of traditional classifiers [13]. By the rejection option, a traditional classifier produces the most probable class for an instance only if its posterior probability is greater than a given threshold, otherwise the instance under examination is rejected and no class is produced. By the rejection option, traditional classifiers suspend the judgement on a set of instances for convenience, in order to reduce the misclassification errors on the remaining ones. In contrast, credal classifiers suspend the judgment because some classes cannot be compared with each another under the weak assumptions the model make. There is a principled and practical difference between the two approaches which can be clarified by focusing on the case when a classifier is inferred from a very small learning set. In this case, a credal classifier will be very likely to suspend the judgment on a new instance, as the learning set provides it with little knowledge on the domain. On the contrary, a traditional classifier using the rejection option will suspend the judgment in a way that is roughly independent of the size of the learning set. In fact the possibility to make a classification depends in this case on the best-class posterior probability, which carries no indication on its variability due to the small sample size. Finally, observe also that the rejection option could be generalized to credal classifiers, by allowing them to produce an output only if a

certain posterior lower probability (e.g., of the output set of classes) is greater than a threshold.

More broadly speaking, credal classification can be motivated and better explained by focusing on the special case of sequential learning tasks, in which typically the classifier starts in condition of prior ignorance. Every new instance is first classified and only then stored in the knowledge base together with the actual class, which is unknown at the classification stage. The classifier's knowledge grows incrementally, so that its predictions become more reliable as more units are collected. A credal classifier naturally shows this behavior. Initially it will produce all the classes (i.e., complete indeterminacy); with more units, the average output set size will decrease approaching one in the limit. If one compares this behavior with that of traditional classifiers that always produce a single class, even when very few examples have been processed, these will appear to be overconfident.

## 3. Credal TANs

Figure 1 shows a simple example of a TAN graphical structure. In general, the TAN structure is a directed graph with nodes $C, A_1, \ldots, A_n$, such that there is an arc from $C$ to any other node, and that the subgraph connecting the attributes is a tree. In a directed tree each node has one parent at most.

The special graphical structure makes TANs a special case of Bayesian networks. According to Section 2.3, by extending them to credal sets, we create the special case of credal networks here called TANCs. Let us consider the learning and classification with TANCs.

### 3.1. Learning

As the focus of this paper is on designing an efficient classification algorithm for TANCs, we only discuss briefly the important issue of learning TANCs from data. We report our simplifying choices below and some possible improvements at the end of this section.

With regard to incomplete data, we assume that data are *missing at random* [26] so that missing data can be ignored in doing the empirical estimates. In practice, we discard the units for which at least the value of one variable, involved in the estimate under consideration, is missing. This approach leads to unbiased estimates under MAR.

With respect to learning the tree structure from data, we base this step on the well-known procedure from Chow and Liu [7]. This is one of the most common choices to infer trees from data, as it was shown

to come up with the best approximating tree (on the basis of Kullback-Leiber's divergence [22]).

Finally, we use (3) for the inference of the credal sets of the model. The generic conditional probability $P(x|pa\,(X))$ is obtained applying (3) to a sub-sample for which $Pa(X) = pa(X)$. Each application of (3) is treated independently of the others, so forcing the resulting credal sets to be separately specified, as assumed in Section 2.3. The separate treatment of credal sets is important in this work to obtain an efficient classification algorithm. The price to pay is some excess of caution in the classifications. This is a general property of imprecise probability models, i.e. the possibility to trade caution for computational complexity, which is exactly the approach followed here.

We can devise some extensions for future research to improve the current learning procedure. The treatment of ignorable missing data could be refined by more sophisticated approaches that better exploit the information in the sample, e.g. see [41, Sect. 5.2]. However, these generally do not provide closed-form expressions for the estimates, which should be approximated by the EM algorithm [6]. This further complication does not seem to be strictly necessary unless the data are largely incomplete. There is room for sophistication also in the case of structure learning. For example, one could think of inferring a partial tree (or *forest*) by suppressing the arcs that are not supported by the information in the data (this issue if discussed further on in Section 4.3). The rationale here is to avoid excessively complicated models that can degrade classification accuracy. This is possible by using traditional methods such as *confidence* or *credibility intervals* [20, 41]. Alternatively, it would be possible to investigate the inference of robust trees to bypass the fragility of the structures obtained from small data sets. Recent achievements from imprecise probabilities would be helpful in this case [42].

### 3.2. Classification

We focus on testing whether or not class $c'$ credal-dominates $c''$, according to (6), as the classification procedure is already reported in Section 2.4.

Let us consider a possibly incomplete instance in which only the attributes $\{A_m, \ldots, A_n\} = E$, $1 \leq m \leq n$, are in a known state, say

$\boldsymbol{e} = (a_m, \ldots, a_n)$. Problem (6) can be rewritten as follows.[4]

$$\begin{aligned} \min_{P \in \mathcal{P}} \left[ P(c', \boldsymbol{e}) - P(c'', \boldsymbol{e}) \right] \; &> \; 0 \Leftrightarrow \\ \min_{P \in \mathcal{P}_C} P(c') \underline{P}(\boldsymbol{e}|c') - P(c'') \overline{P}(\boldsymbol{e}|c'') \; &> \; 0 \Leftrightarrow \qquad (7) \\ \underline{P}(c') \underline{P}(\boldsymbol{e}|c') - \overline{P}(c'') \overline{P}(\boldsymbol{e}|c'') \; &> \; 0. \qquad (8) \end{aligned}$$

The first passage is possible given that the credal sets of the net are separately specified. The quantities $P(\boldsymbol{e}|c')$ and $P(\boldsymbol{e}|c'')$ can be optimized separately since they depend on disjoint collections of credal sets. This should be clear since they depend on different classes. The last passage is a straightforward application of (1).

Now there are two cases, depending on $m$. If $m = 1$ the instance is complete and for each $c \in \mathcal{C}$,

$$\underline{P}(\boldsymbol{e}|c) \; = \; \prod_{i=1}^{n} \underline{P}(a_i|c, pa(A_i)) \qquad (9)$$

$$\overline{P}(\boldsymbol{e}|c) \; = \; \prod_{i=1}^{n} \overline{P}(a_i|c, pa(A_i)) \qquad (10)$$

again because of the separate specification of the credal sets ($pa(A_i)$ is, of course, consistent with $\boldsymbol{e}$ for each $i = 1, \ldots, n$). Substituting the products related to $c'$ and $c''$ in (8), we have a closed-form expression for testing credal dominance.

The case of incomplete instance, $m > 1$, is more complicated, the nodes that do not belong to $E$ must be marginalized out and this involves a propagation of intervals in the tree. In the following we give a procedure to compute the lower and upper probabilities of $(\boldsymbol{e}|c)$, $c$ being a generic class. The procedure takes a TANC and the class $c$ as input, and is based on two operations called *absorption* and *removal* of nodes, defined below.

Absorption and removal of nodes apply to a node $X$ whose children, $Y_1, \ldots, Y_k$ ($k \geq 1$), are all leaves. Figure 3 shows the generic portion of tree related to $X$, its children and its parent $U$ (the dependency on the class is not displayed in the following figures). When $X \in E$ (with value $x$), node absorption clusters node $X$ and all its children into a new node, like in Figure 4. The relevant intervals of the new node are

[4] In the following we drop the condition $P(\boldsymbol{a}) > 0$ that appears in (6) as each mass function in the strong extension assigns positive probability to any event when the probability intervals are inferred by the IDM.

1. Remove the arcs from the node $C$ to the attributes;
2. recursively remove the leaves that do not belong to $E$;
3. if the resulting tree is a single node, output the lower and upper probabilities held by the node and stop;
4. select an inner (non-leaf) node, say $X$, such that all its children are leaves;
5. If $X \in E$ absorb $X$, else remove $X$;
6. go to step 3.

*Figure 2.* Algorithm for the computation of the lower and upper probability of $(\boldsymbol{e}|c)$ with an incomplete instance $\boldsymbol{e}$.

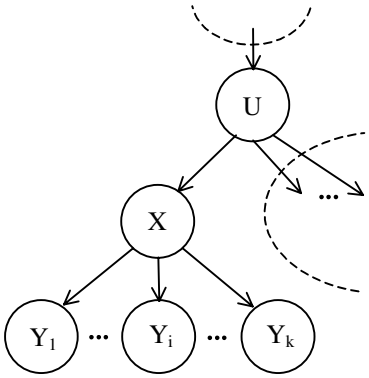


*Figure 3.* A portion of the tree.

computed by

$$\underline{P}(x, y_1, \ldots, y_k|u, c) \;=\; \underline{P}(x|u, c) \prod_{i=1}^{k} \underline{P}(y_i|x, c) \tag{11}$$

$$\overline{P}(x, y_1, \ldots, y_k|u, c) \;=\; \overline{P}(x|u, c) \prod_{i=1}^{k} \overline{P}(y_i|x, c) \tag{12}$$

for each $u \in \mathcal{U}$.

If $X \notin E$, node removal removes node $X$, and clusters all its children into a new node, like in Figure 5. The relevant intervals of the new node

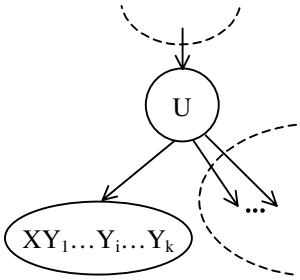


*Figure 4.* The graph after the absorption of $X$.

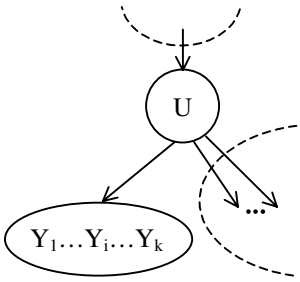


*Figure 5.* The graph after the removal of $X$.

are computed by

$$\underline{P}(y_1,\ldots,y_k|u,c) = \min_{P\in\mathcal{P}_X^{u,c}} \sum_{x\in\mathcal{X}} \left[ P(x|u,c) \prod_{i=1}^{k} \underline{P}(y_i|x,c) \right] \quad (13)$$

$$\overline{P}(y_1,\ldots,y_k|u,c) = \max_{P\in\mathcal{P}_X^{u,c}} \sum_{x\in\mathcal{X}} \left[ P(x|u,c) \prod_{i=1}^{k} \overline{P}(y_i|x,c) \right] \quad (14)$$

for each $u \in \mathcal{U}$. Each of these optimization problems is easily solved in time linear with the size of $\mathcal{X}$, as shown in Appendix B.

Appendix A proves the correctness and termination of the entire procedure.

### 3.3. Classification complexity

The complexity of credal classification with TANCs depends on whether or not the instance is complete. We first analyze the more complicated case of incomplete instance. We derive the complexity in this case by using a bottom-up approach.

Let us start by evaluating the complexity of the operation of node removal. Let $M = \arg\max_{i=1,\ldots,n} |\mathcal{A}_i|$. $M$ is an upper bound on the

number of values of a generic node (such as $X$) in the net. Consider expression (13), for a class $c$, a node $X$, and $u \in \mathcal{U}$ (the case of expression (14) is analogous). Doing all the products of the conditional probabilities in (13) takes $O\left(od\left(X\right)M\right)$, denoting by $od\left(\cdot\right)$ the out-degree of a node, i.e. the number of arcs that leave the node. The optimization problem in (13) is solved in time $O(M)$, according to Appendix B. Overall, we have $O(od\left(X\right)M + M)$ to set up and solve the problem for a single $u \in \mathcal{U}$. Repeating it for all $u \in \mathcal{U}$, the removal of node $X$ takes $O\left(\left(od\left(X\right)+1\right)M^2\right)$.

The complexity of node removal is an upper bound on the complexity of node absorption as it is clear comparing (11) with (13). By summing the last expression over all the nodes in a tree, we have a bound on the complexity of computing the lower and upper probabilities of $(\boldsymbol{e}|\, c)$ (Section 3) for a specific class $c$, i.e. $O\left(nM^2\right)$. This follows because the removal of $X$ does not change the out-degree of $U$; and because the sum of the out-degrees is the sum of the number of arcs in the tree, which is bounded by the number of nodes $n$.

Extending the last expression to all $c \in \mathcal{C}$ and taking into account the $|\mathcal{C}|^2$ tests of credal dominance (8), we obtain the overall complexity to credal-classify an incomplete instance, i.e.

$$O\left(n\left|\mathcal{C}\right|M^2 + \left|\mathcal{C}\right|^2\right). \tag{15}$$

With complete instances, the complexity is

$$O\left(n\left|\mathcal{C}\right| + \left|\mathcal{C}\right|^2\right) \tag{16}$$

as it follows from (8), (9), and with similar arguments to those used to derive (15).

Observe that in common applications, both $M$ and $|\mathcal{C}|$ are relatively low.

## 4. Experimental analysis

We considered 8 data sets, listed in Table I. These are real and artificial data sets from different domains. For example, "Breast Wisconsin" reports patients data for diagnosing breast cancer; "Pima-indians diabetes" is another medical data set; "Vehicle" contains data for car recognition; etc.

Table I. Data sets used in the experiments, together with their number of classes, of features, of instances and of missing data. All the data sets are available from the UCI repository of machine learning data sets [29].

| Name | # cl. | # feat. | # inst. | # m.d. |
|---|---|---|---|---|
| Breast Wisconsin | 2 | 9 | 699 | 16 |
| Chess (kr_kp) | 2 | 36 | 3196 | 0 |
| Crx | 2 | 15 | 690 | 37 |
| German | 2 | 20 | 1000 | 0 |
| Iris | 3 | 4 | 150 | 0 |
| Molec. biology (splice) | 3 | 60 | 3170 | 0 |
| Pima-indians diabetes | 2 | 8 | 768 | 0 |
| Vehicle | 4 | 18 | 946 | 0 |

### 4.1. Experimental methodology

The data sets containing non-discrete features were discretized by the common entropy-based discretization [16]. Then we ran 5 repetitions of the empirical evaluation scheme called *5-fold cross validation* [21]. According to $r$-fold cross-validation, a data set $\mathcal{D}$ is split into $r$ mutually exclusive subsets (the folds) $\mathcal{D}_1, \ldots, \mathcal{D}_r$ of approximately equal size. The classifier is inferred and tested $r$ times; each time $t$ it is inferred from $\mathcal{D} \setminus \mathcal{D}_t$ and tested on $\mathcal{D}_t$, where the actual classes are hidden to the classifier. The statistics for the quantities of interest, like the percentage of successful predictions on the test set $\mathcal{D}_t$ (i.e. the *prediction accuracy*), are collected over the $r$ folds. In order to collect better estimates, $r$-fold cross validation is usually repeated a number of times and the results are averaged over the repetitions.

The experiments involved the comparison of the TANC model with its precise probability counterpart, the TAN. As far as TANCs are concerned, we set the parameter $s$ to 1 (Section 2.2), one of the choices suggested in [35]. TANs were inferred by using the noninformative uniform prior distribution again with $s = 1$. Recall also that the choice of the weight of the Bayesian prior is arbitrary, as it happens usually with Bayesian models. The IDM inherits this characteristics, though Walley gives reasonable motivations to choose $s$ in the interval $[1, 2]$. The effect of $s$ on the TANC classifications follows easily under the interpretation of $s$ as caution parameter: the larger $s$, the larger in general the output

sets of classes for a given instance. Notably, the sets related to larger $s$'s will always include those following from smaller ones.

In the following, when the TANC produces more than one class for an instance, we say that the classification is (partially or totally) indeterminate and the classifier is imprecise, or suspends the judgment, with respect to the instance. In the opposite case we speak of determinate classifications and of precision with regard to classifiers.

Finally, the notation $a \pm b$ used in the following section denotes, each time, a value $a$ and its standard deviation $b$ ($b$ is computed according to [21]; the possible upper bounds greater than 100 should be regarded as 100%). All the comparisons reported as *significative* were tested for statistical significance using the two-samples one-tail t-test at level 0.05.

### 4.2. Results

It is convenient to examine the data sets with two classes separately from the others. Table II shows the results of the experiments for the former ones.

Table II. Results of the experimental evaluation of TANCs on the two-classes data sets. The cells contain percentages.

| Name | M% | $C_1$% | T% | N% | Ts% | S% |
|---|---|---|---|---|---|---|
| Breast Wisconsin | 65.5 | 100.0±0.0 | 94.4±0.9 | 97.4±1.1 | 86.8±4.4 | 42.4±1.9 |
| Chess (kr_kp) | 52.0 | 93.3±1.0 | 92.7±0.5 | 87.8±1.3 | 60.6±14.0 | 1.9±0.2 |
| Crx | 55.5 | 90.5±3.5 | 84.1±3.1 | 85.4±3.3 | 70.1±6.9 | 31.2±3.9 |
| German | 70.0 | 80.0±3.6 | 73.6±1.4 | 74.6±2.2 | 61.8±5.8 | 35.3±1.5 |
| Pima-indians diabetes | 65.1 | 80.2±3.6 | 76.1±1.6 | 74.4±2.6 | 58.3±9.2 | 18.8±1.4 |

The columns are described below. Let us recall that the prediction accuracy corresponds to the relative number of correct predictions.

- M% is the relative frequency of the mode (i.e., the majority class) of the data sets.
- $C_1$% is the accuracy of the TANC on the subset of instances where it is precise.
- T% is the accuracy of the TAN on the entire test set.
- N% is the accuracy of the *Naive Bayes classifier* [12] on the entire test set. This information will be used for discussion in the following section. It can be also used to verify the improvement of TAN

over the naive Bayes that can occur as the latter assumes mutual independence of the attributes conditional on the class.

- Ts% is the accuracy of the TAN on the subset of instances for which the TANC is imprecise.

- S% is the percentage of instances for which the TANC is imprecise.

First, the comparison of $C_1$% with M% shows that the TANC model does non-trivial predictions; these are better than those achieved by the trivial classifier that each time outputs the mode. The situation is similar with the TAN. Comparing $C_1$% with T% shows that when the TANC is precise, its predictions are significantly better then the TAN ones (significance does not hold with the Chess data set). In the area of TANC indeterminacy (S%), the TAN accuracy drops considerably, as it is clear comparing Ts% with T%, so that the TAN is quite unreliable. Furthermore, Ts% is always significantly worse than $C_1$%.

The results for the data sets with more than two classes are in Table III.

Table III. Results of the experimental evaluation of TANCs on the data sets with more than two classes. The cells contain percentages, except Ps.

| Name | M% | $C_1$% | Cs% | T% | N% | Ts% | Rs% | S% | Ps |
|---|---|---|---|---|---|---|---|---|---|
| Iris | 33.3 | 97.0±3.3 | 100.0±0.0 | 92.0±2.2 | 94.7±2.9 | 52.9±27.1 | 30.4±24.9 | 11.3±2.6 | 2.6/3 |
| Molec. biology (splice) | 51.9 | 97.7±0.6 | 97.8±1.7 | 94.7±0.4 | 95.4±0.9 | 70.4±5.4 | 47.1±5.8 | 11.1±0.6 | 2.1/3 |
| Vehicle | 25.4 | 85.1±4.0 | 85.8±3.7 | 73.6±1.5 | 62.3±3.1 | 63.2±5.1 | 36.8±5.1 | 52.4±1.7 | 2.1/4 |

There are three more columns with respect to Table II.

- Cs% is a set-accuracy, i.e. it represents the probability that the actual class belongs to the set of classes proposed by the TANC. This measure is computed in the subset of instances where the TANC is imprecise.

- Rs% is the accuracy of a uniformly random predictor that randomly chooses one of the classes proposed by the TANC. This measure is computed in the subset of instances for which the TANC is imprecise and the actual class belongs to the set the TANC proposes.

- Ps is the average relative number of classes produced by the TANC in the area of imprecision.

Again, the TANC does non-trivial predictions ($C_1$% vs. M%) and these are significantly better than T% and Ts%. The TAN is again quite unreliable (T% vs. Ts%, and Ts% vs. Rs%) in the area of indeterminacy (S%). The TANC adopts a cautious approach for the instances of S%: it produces more than one class (Ps); and these contain the actual one with high probability (Cs%). Note that the set of classes produced by the TANC is meaningful, because it is strictly included in the set of all the classes on average (Ps is less than 1).

### 4.3. Discussion

An important experimental evidence is that the TANC shows a very good prediction performance on the instances where it is precise. Furthermore, the experiments highlight that the TANC can isolate an area (S%) of hard instances to classify, given the scarce knowledge relevant to them in the learning set. On these instances, the TAN loses performance considerably (T% vs. Ts%), in a way that it sometimes approaches the performance of a random predictor (i.e., 50% accuracy in the first group of data sets, and Rs% in the second). In contrast, the TANC realizes that the knowledge is scarce and, rather than losing reliability, gives a non-informative output (with the first group of data sets) or a partially informative one (with the second) that contains the actual class with high probability.

TANCs appear to be safer classifiers than TANs; in the area of imprecision TANs are quite unreliable. However, TANCs have another type of drawback: they appear to be overly cautious. The fraction of instances where the TANC is imprecise can be large, e.g., in the "Breast Wisconsin" or "Vehicle" data sets, in a way that does not always seem to be reasonable. We think that this characteristic is related to a problem that arises when the model is too complex with respect to the available data. In this case the increase in variance of the model parameters (e.g., the empirical probabilities) leads to bad estimates.

Let us focus on the "Breast Wisconsin" data set. The naive Bayes classifier which is much simpler than the TAN achieves very good prediction accuracy with it, as it follows from column N% in Table II. The TAN does a worse performance (T% in Table II) because it appears to unreasonably increase the request of parameters to estimate: TANs are forced to infer a dependency tree for the attributes, by definition, also when such a complex structure is not justified, as it appears to be the case of "Breast Wisconsin".

Large variance of the parameters produces performance loss with traditional classifiers. With TANCs or, more generally, with credal classifiers, such variability is likely to produce excessive caution. In fact,

a largely imprecise probability interval can be the result of the large variability of an empirical probability. Large probability intervals in the TANC model raise the chance to produce (partially) indeterminate classifications. Similar analysis on the remaining data sets seems to confirm this conjecture.

The issue of model complexity is related in the classification literature to the issue of *overfitting*. The experimental analysis carried out here raises a critical point: avoiding overfitting seems to be important for credal classifiers as well as for traditional ones. With TANCs avoiding overfitting means, for example, suppressing the arcs of the tree that connect weakly dependent attributes, i.e. moving to models based on forests. This is readily possible with TANCs (see the observation at the end of Appendix A), and we plan to investigate this point in future research.

## 5. Conclusions

There are different issues that must be tackled in order to build reliable classifiers. In this paper we have focused on one of them that is very often neglected: the careful modelling of prior ignorance. We have adopted the imprecise Dirichlet model as a well-founded approach to the common data mining problem of learning from data when these are the only source of knowledge about a domain.

The IDM naturally leads to probability intervals. The present work has shown how to propagate these intervals in a graphical structure to carry out reliable classification. The resulting TANC model is a credal classifier that extends the pre-existing tree-augmented naive classifier to imprecise probabilities. We have shown that the classification with TANCs takes a time linear in the number of attributes. This is an important result for applications that follows from a careful choice of the representation level that allows a favorable trade-off between efficient computation and model expressiveness.

The experiments have shown that TANCs are good and reliable classifiers. On the negative side, they exhibit an excess of caution in some cases, which can be related to a problem of overfitting. This is likely to be a characteristic of credal classification, not only of TANCs. Future research should investigate the effect on credal classifiers of the traditional tools to avoid overfitting, such as feature selection and model pruning [37].

With respect to future research, we believe that there are two other major lines of investigation. The first involves extending the structure learning to interval or imprecise probability. Some work has already

been done on the subject [3, 20, 42], and this seems to be another substantial step to the direction of reliability. The second line involves addressing the problems posed by non-ignorable missing data. Literature presents a principled approach to the problem that would be worth extending to TANCs [39, 33].

**Acknowledgements**

The authors are grateful to Paola Chiesa and Valeria Nobile for support with software development. Thanks to Monaldo Mastrolilli for help with knapsack problems. This research was partially supported by the Swiss NSF grant 2100-067961.02.

## Appendix

### A. Correctness proof

We show the correctness of the operations of node absorption and removal defined in Section 3.2 and then the correctness of the entire procedure in Figure 2.

Let us recall that the operations apply to a node $X$ whose children, $Y_1, \ldots, Y_k$ ($k \geq 1$), are all leaves. Refer to Figure 3, which shows the portion of tree related to $X$, its children and its parent $U$. Note that when the operations are applied in step 5 of the procedure, $Y_1, \ldots, Y_k$ are in $E$.

The operations work under the following conditions: credal sets of different nodes are separately specified as well as credal sets in the same node when this is non-leaf. When the node is a leaf a weaker condition suffices. Recall that we are only interested in the instance of the leaves $(y_1, \ldots, y_k)$ that is consistent with $\boldsymbol{e}$ and hence the probabilistic information that we need about $Y_i$ ($\forall i \in \{1, \ldots, k\}$) reduces itself to the intervals $[\underline{P}(y_i\,|x, c), \overline{P}(y_i\,|x, c)]$, $x \in \mathcal{X}$, for a given $c \in \mathcal{C}$. Given $i \in \{1, \ldots, k\}$, we require that there exists a mass function $P \in \mathcal{P}$ so that $P(y_i\,|x, c) = \underline{P}(y_i\,|x, c)$ for all $x \in \mathcal{X}$; similarly for the right extremes (the two mass functions need not be the same). We call this condition *consistency of extremes* (or *consistency*, for short). Consistency allows us to simultaneously consider either all the left extremes $\underline{P}(y_i\,|x, c)$, $x \in \mathcal{X}$, or all the right extremes; but it does not allow us to simultaneously consider, for example, $\underline{P}(y_i\,|x', c)$ and $\overline{P}(y_i\,|x'', c)$, $x', x'' \in \mathcal{X}$, $x' \neq x''$, as it would be possible if separate specification of credal sets was assumed.

With reference to node absorption, we prove the following proposition.

PROPOSITION 1. *For each $u \in \mathcal{U}$, the lower and upper probabilities of $(x, y_1, \ldots, y_k|u, c)$ are given by (11) and (12), respectively, and they are consistent.*

*Proof.* Consider the left extremes (the remaining case is analogous).

Let $u \in \mathcal{U}$. $\underline{P}(x, y_1, \ldots, y_k|u, c) = \underline{P}(x|u, c) \prod_{i=1}^{k} \underline{P}(y_i|x, c)$ follows with analogous arguments to that leading to (7). Given $u', u'' \in \mathcal{U}$, $u' \neq u''$, $\underline{P}(x|u', c) \prod_{i=1}^{k} \underline{P}(y_i|x, c)$ is consistent with $\underline{P}(x|u'', c) \prod_{i=1}^{k} \underline{P}(y_i|x, c)$ because in both cases $P(y_i|x, c)$ is set at the same value ($\forall i \in \{1, \ldots, k\}$) and because $P(x|u', c)$ and $P(x|u'', c)$ are separately specified.

Now we focus on node removal.

PROPOSITION 2. *For each $u \in \mathcal{U}$, the lower and upper probabilities of $(y_1, \ldots, y_k|u, c)$ are given by (13) and (14), respectively, and they are consistent.*

*Proof.* Consider $P(y_1, \ldots, y_k|u, c)$, for a given $u \in \mathcal{U}$.

By marginalization and for the independence of $Y_1, \ldots, Y_k$ and $U$, given $X$ and $C$, we have $P(y_1, \ldots, y_k\,|u, c) = \sum_{x \in \mathcal{X}} [P(x\,|u, c) P(y_1, \ldots, y_k\,|x, c)]$. This is also equal to the expression $\sum_{x \in \mathcal{X}} [P(x\,|u, c) \prod_{i=1}^{k} P(y_i\,|x, c)]$ because $Y_1, \ldots, Y_k$ are mutually independent given $X$ and $C$. We take the minimum of the last expression (the maximum is analogous). Recall that the values $\underline{P}(y_i\,|x, c)$, $x \in \mathcal{X}$, are consistent ($\forall i \in \{1, \ldots, k\}$) and that credal sets of different nodes are separately specified. Given that all the numbers involved are non-negative, we have $\underline{P}(y_1, \ldots, y_k\,|u, c) = \min_{P \in \mathcal{P}_X^{u,c}} \sum_{x \in \mathcal{X}} [P(x\,|u, c) \prod_{i=1}^{k} \underline{P}(y_i\,|x, c)]$.

Let us now show that given $u', u'' \in \mathcal{U}$, $u' \neq u''$, $\underline{P}(y_1, \ldots, y_k\,|u', c)$ is consistent with $\underline{P}(y_1, \ldots, y_k\,|u'', c)$. First, observe that the credal sets $\mathcal{P}_X^{u',c}$ and $\mathcal{P}_X^{u'',c}$ are separately specified. Second, $P(y_i\,|x, c)$ is fixed at the same value in both cases ($\forall i \in \{1, \ldots, k\}, \forall x \in \mathcal{X}$).

(Observe that both operation can be applied also when $X$ is the root of the tree.)

Now consider the procedure in Figure 2.

THEOREM 1. *Given a TANC, a class $c \in \mathcal{C}$ and an instance $\boldsymbol{e} = (a_m, \ldots, a_n)$, the procedure in Figure 2 terminates and computes $\underline{P}(\boldsymbol{e}|c)$ and $\overline{P}(\boldsymbol{e}|c)$.*

*Proof.*

Consider correctness first. Step 1 produces a credal net that represents the strong extension conditional on $c$, namely $\mathcal{P}^c$. Step 2 corresponds to marginalize out the leaves that do not belong to $E$, for each

mass function in $\mathcal{P}^c$. The operation in steps 1 and 2 are well-known with Bayesian networks. They hold with a credal network, too, because in the present formulation the credal net is equivalent to a set of Bayesian nets (as it follows from the expression for the strong extension (5)), and for each of them the operations in the above steps hold.

Next we us focus on the steps 4 and 5. First, trivially, note that if the tree has more than one node, there is always an inner node $X$ the children of which are all leaves, as required by 4. Second, both operations in point 2 produce a new tree for which (i) the relevant information concerning the nodes in $E$ is preserved; (ii) only a new leaf node is created the intervals of which are consistent. Therefore, the procedure can iteratively be applied to the new tree. The last operation creates a single node that represents $E$, so providing us with $[\underline{P}(\boldsymbol{e}|c), \overline{P}(\boldsymbol{e}|c)]$ in step 3.

Termination is trivially proved considering step 3 and observing that the size of the tree strictly decreases with each execution of step 5.

As a marginal note, it is worth observing that the algorithm can be applied to a *forest* (i.e., a set of disjoint trees) in a straightforward way. The above procedure turns a tree into a single node. By applying the procedure to each tree, the forest in turned into a set of nodes. The nodes are unconnected and for this reason they are strongly independent. Furthermore, their related intervals are separately specified by construction. By analogous arguments used to derive (7), $\underline{P}(\boldsymbol{e}|c)$ is then the product of the lower probabilities of the nodes; similarly for $\overline{P}(\boldsymbol{e}|c)$. This observation allows TANCs to be defined more generally by allowing a forest to model the dependencies between attributes.

## B. Lower and upper expectations as knapsack problems

We formalize the computation of lower and upper expectations with probability intervals by the following optimization problem,

$$\begin{aligned} & opt \sum_{x \in \mathcal{X}} \alpha_x p_x \\ & \sum_{x \in \mathcal{X}} p_x = 1 \\ & l_x \leq p_x \leq u_x \quad x \in \mathcal{X} \end{aligned}$$

where $opt \in \{\min, \max\}$, and for each $x \in \mathcal{X}$, $p_x$ is a decision variable, $\alpha_x \geq 0$ (but there is $x \in \mathcal{X}$ so that $\alpha_x > 0$), and $0 \leq l_x \leq u_x \leq$

1. The problem can be solved in time $O(|\mathcal{X}|)$ by means of simple transformations, as shown below[5].

For the moment consider the minimization problem. The variables of the problem are $p_{x'}$, $x' \in \mathcal{X}' = \{x \in \mathcal{X} \mid u_x > l_x\}$, as the others are constant and are removed from the formulation, yielding

$$\begin{aligned} \sum_{x \in \mathcal{X} \setminus \mathcal{X}'} \alpha_x l_x + \quad & \min \sum_{x' \in \mathcal{X}'} \alpha_{x'} p_{x'} \\ & \sum_{x' \in \mathcal{X}'} p_{x'} = 1 - \sum_{x \in \mathcal{X} \setminus \mathcal{X}'} l_x \\ & l_{x'} \leq p_{x'} \leq u_{x'}, \quad x' \in \mathcal{X}'. \end{aligned}$$

The problem is re-written by using the transformation $q_{x'} = \frac{p_{x'} - l_{x'}}{\delta_{x'}}$, where $\delta_{x'} = u_{x'} - l_{x'}$, for each $x' \in \mathcal{X}'$:

$$\begin{aligned} \textstyle\sum_{x \in \mathcal{X}} \alpha_x l_x + \quad & \min \textstyle\sum_{x' \in \mathcal{X}'} (\alpha_{x'} \delta_{x'})\, q_{x'} && (17) \\ & \textstyle\sum_{x' \in \mathcal{X}'} \delta_{x'} q_{x'} \geq 1 - \sum_{x \in \mathcal{X}} l_x && (18) \\ & 0 \leq q_{x'} \leq 1, \quad x' \in \mathcal{X}'. && (19) \end{aligned}$$

Note that the constants $\delta_{x'}$ are positive and all the constants $(\alpha_{x'} \delta_{x'})$ and $(1 - \sum_{x \in \mathcal{X}} l_x)$ are non-negative (the case $\sum_{x \in \mathcal{X}} l_x = 1$ admits the trivial solution $q_{x'} = 0$ for each $x' \in \mathcal{X}'$; and hence we can treat it apart, and assume that $\sum_{x \in \mathcal{X}} l_x$ is strictly less than 1 in the formulation). Note also that the constraint $\sum_{x' \in \mathcal{X}'} \delta_{x'} q_{x'} = 1 - \sum_{x \in \mathcal{X}} l_x$ has been replaced by $\sum_{x' \in \mathcal{X}'} \delta_{x'} q_{x'} \geq 1 - \sum_{x \in \mathcal{X}} l_x$ , since at the optimum the latter one must be satisfied with the equality sign if the original problem is feasible, otherwise the optimum might be strictly improved.

As final step, consider the new decision variables $r_{x'} = 1 - q_{x'}$, $x' \in \mathcal{X}'$, yielding a new formulation,

$$\begin{aligned} \sum_{x \in \mathcal{X}} \alpha_x l_x + \textstyle\sum \alpha_{x'} \delta_{x'} + \quad & \min \textstyle\sum_{x' \in \mathcal{X}'} (-\alpha_{x'} \delta_{x'})\, r_{x'} \\ & \sum_{x' \in \mathcal{X}'} (-\delta_{x'} r_{x'}) \geq 1 - \sum_{x \in \mathcal{X}} l_x - \sum_{x' \in \mathcal{X}'} \delta_{x'} \\ & 0 \leq r_{x'} \leq 1, \quad x' \in \mathcal{X}' \end{aligned}$$

[5] There are simpler algorithms to solve this problem, but they usually require the $\alpha_x$ coefficients to be sorted, which introduces a logarithmic term in the complexity that is avoided by the procedure described here.

which is also,

$$\sum_{x\in\mathcal{X}} \alpha_x l_x + \sum\nolimits_{x'\in\mathcal{X}'} \alpha_{x'}\delta_{x'} - \quad \max \sum\nolimits_{x'\in\mathcal{X}'} (\alpha_{x'}\delta_{x'})\, r_{x'}$$

$$\sum_{x'\in\mathcal{X}'} \delta_{x'} r_{x'} \leq -1 + \sum_{x\in\mathcal{X}} l_x + \sum_{x'\in\mathcal{X}'} \delta_{x'}$$

$$0 \leq r_{x'} \leq 1, \quad \forall x' \in \mathcal{X}'.$$

The latter maximization is a special case of the *continuous knapsack problem* (notice that it is always $\sum_{x'\in\mathcal{X}'} \delta_{x'} \geq 1 - \sum_{x\in\mathcal{X}} l_x$). Both Lawler [24, Sect. 4] and Balas and Zemel [2] (see also [27, Sect. 2.2.1–2.2.2]) give a procedure that solves continuous knapsack problems in time $O(|\mathcal{X}'|)$. (The connection between probability intervals and knapsack problems was also noticed in [19, Sect. 3].)

The maximization case is completely analogous up to formulation in (17)–(19), where the problem is already in knapsack form, thus not needing any further transformation.

## References


1. Abellán, J. and S. Moral: 2001, 'Building classification trees using the total uncertainty criterion'. In: G. de Cooman, T. Fine, and T. Seidenfeld (eds.): *ISIPTA'01*. The Netherlands, pp. 1–8.
2. Balas, E. and E. Zemel: 1980, 'An algorithm for large zero-one knapsack problems'. *Operations Research* **28**, 1130–1154.
3. Bernard, J.-M.: 2002, 'Implicative analysis for multivariate binary data using an imprecise Dirichlet model'. *Journal of Statistical Planning and Inference* **105**(1), 83–103.
4. Bernardo, J. M. and A. F. M. Smith: 1996, *Bayesian Theory*. New York: Wiley.
5. Campos, L., J. Huete, and S. Moral: 1994, 'Probability intervals: a tool for uncertain reasoning'. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2**(2), 167–196.
6. Chen, T. T. and S. E. Fienberg: 1974, 'Two-Dimensional Contingency Tables with both Completely and Partially Cross-Classified Data'. *Biometrics* **32**, 133–144.
7. Chow, C. K. and C. N. Liu: 1968, 'Approximating discrete probability distributions with dependence trees'. *IEEE Transactions on Information Theory* **IT-14**(3), 462–467.
8. Couso, I., S. Moral, and P. Walley: 2000, 'A survey of concepts of independence for imprecise probability'. *Risk, Decision and Policy* **5**, 165–181.
9. Cozman, F. G.: 2000a, 'Credal networks'. *Artificial Intelligence* **120**, 199–233.
10. Cozman, F. G.: 2000b, 'Separation Properties of Sets of Probabilities'. In: C. Boutilier and M. Goldszmidt (eds.): *UAI-2000*. San Francisco, pp. 107–115.
11. Dasgupta, S.: 1999, 'Learning polytrees'. In: *UAI-99*. San Francisco, pp. 134–141.
12. Duda, R. O. and P. E. Hart: 1973, *Pattern classification and scene analysis*. New York: Wiley.

13. Duda, R. O., P. E. Hart, and D. G. Stork: 2001, *Pattern classification*. Wiley. 2nd edition.
14. Fagiuoli, E. and M. Zaffalon: 1998, '2U: an exact interval propagation algorithm for polytrees with binary variables'. *Artificial Intelligence* **106**(1), 77–107.
15. Fagiuoli, E. and M. Zaffalon: 2000, 'Tree-augmented naive credal classifiers'. In: *IPMU 2000: Proceedings of the 8th Information Processing and Management of Uncertainty in Knowledge-Based Systems Conference*. Spain, pp. 1320–1327.
16. Fayyad, U. M. and K. B. Irani: 1993, 'Multi-interval discretization of continuous-valued attributes for classification learning'. In: *Proceedings of the 13th international joint conference on artificial intelligence*. San Francisco, CA, pp. 1022–1027.
17. Ferreira da Rocha, J. C. and F. G. Cozman: 2002, 'Inference with separately specified sets of probabilities in credal networks'. In: A. Darwiche and N. Friedman (eds.): *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*. pp. 430–437.
18. Friedman, N., D. Geiger, and M. Goldszmidt: 1997, 'Bayesian networks classifiers'. *Machine Learning* **29**(2/3), 131–163.
19. Ha, V., A. Doan, V. Vu, and P. Haddawy: 1998, 'Geometric foundations for interval-based probabilities'. *Annals of Mathematics and Artificial Intelligence* **24**(1–4), 1–21.
20. Kleiter, G. D.: 1999, 'The posterior probability of Bayes nets with strong dependences'. *Soft Computing* **3**, 162–173.
21. Kohavi, R.: 1995, 'A study of cross-validation and bootstrap for accuracy estimation and model selection'. In: *IJCAI-95*. San Mateo, pp. 1137–1143.
22. Kullback, S. and R. A. Leiber: 1951, 'On information and sufficiency'. *Ann. Math. Statistics* **22**, 79–86.
23. Kyburg, H. E. J.: 1983, 'Rational Belief'. *The behavioral and brain sciences* **6**, 231–273.
24. Lawler, E.: 1979, 'Fast approximation algorithms for knapsack problems'. *Mathematics of Operations Research* **4**(4), 339–356.
25. Levi, I.: 1980, *The Enterprise of Knowledge*. London: MIT Press.
26. Little, R. J. A. and D. B. Rubin: 1987, *Statistical Analysis with Missing Data*. New York: Wiley.
27. Martello, S. and P. Toth: 1990, *Knapsack Problems: Algorithms and Computer Implementations*. Chichester: Wiley.
28. Moral, S. and A. Cano: 2002, 'Strong conditional independence for credal sets'. *Annals of Mathematics and Artificial Intelligence* **35**(1–4), 295–321.
29. Murphy, P. M. and D. W. Aha: 1995, 'UCI Repository of Machine Learning Databases'. http://www.sgi.com/Technology/mlc/db/.
30. Nivlet, P., F. Fournier, and J.-J. Royer: 2001, 'Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition'. In: G. de Cooman, T. Fine, and T. Seidenfeld (eds.): *ISIPTA'01*. The Netherlands, pp. 284–292.
31. Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.
32. Quinlan, J. R.: 1993, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
33. Ramoni, M. and P. Sebastiani: 2001, 'Robust Bayes classifiers'. *Artificial Intelligence* **125**(1–2), 209–226.
34. Walley, P.: 1991, *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall.

35. Walley, P.: 1996, 'Inferences from multinomial data: learning about a bag of marbles'. *J. R. Statist. Soc. B* **58**(1), 3–57.
36. Walley, P. and T. L. Fine: 1982, 'Towards a frequentist theory of upper and lower probability'. *Ann. Statist.* **10**, 741–761.
37. Witten, I. H. and E. Frank: 1999, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.
38. Zaffalon, M.: 1999, 'A credal approach to naive classification'. In: G. de Cooman, F. Cozman, S. Moral, and W. P. (eds.): *ISIPTA'99*. Univ. of Gent, Belgium, pp. 405–414.
39. Zaffalon, M.: 2001, 'Statistical inference of the naive credal classifier'. In: G. de Cooman, T. Fine, and T. Seidenfeld (eds.): *ISIPTA'01*. The Netherlands, pp. 384–393.
40. Zaffalon, M.: 2002, 'The naive credal classifier'. *Journal of Statistical Planning and Inference* **105**(1), 5–21.
41. Zaffalon, M. and M. Hutter: 2002, 'Robust feature selection by mutual information distributions'. In: A. Darwiche and N. Friedman (eds.): *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*. pp. 577–584.
42. Zaffalon, M. and M. Hutter: 2003, 'Robust inference of trees'. Technical Report IDSIA-11-03, IDSIA.