Likelihood-Based Robust Classification with Bayesian Networks

Alessandro Antonucci Marco E. G. V. Cattaneo Giorgio Corani

Technical Report No. IDSIA-01-12
January 2012

IDSIA / USI-SUPSI
Istituto Dalle Molle di studi sull'intelligenza artificiale
Galleria 2, 6928 Manno, Switzerland

IDSIA was founded by the Fondazione Dalle Molle per la Qualità della Vita and is affiliated with both the Università della Svizzera italiana (USI) and the Scuola unversitaria professionale della Svizzera italiana (SUPSI).

M. Cattaneo is with the Department of Statistics, LMU Munich, München (Germany), `cattaneo@stat.uni-muenchen.de`

Likelihood-Based Robust Classification with Bayesian Networks

Alessandro Antonucci Marco E. G. V. Cattaneo Giorgio Corani

January 2012

Abstract

Bayesian networks are commonly used for classification: a structural learning algorithm determines the network graph, while standard approaches estimate the model parameters from data. Yet, with few data the corresponding assessments can be unreliable. To gain robustness in this phase, we consider a *likelihood-based* learning approach, which takes all the model quantifications whose likelihood exceeds a given threshold. A new classification algorithm based on this approach is presented. Notably, this is a *credal* classifier, i.e., more than a single class can be returned in output. This is the case when the Bayesian networks consistent with the threshold constraint assign different class labels to a test instance. This is the first classifier of this kind for general topologies. Experiments show how this approach provide the desired robustness.

1 Introduction

The development of classifiers, i.e., algorithms to assign *class* labels to instances described by a set of *features*, is a major problem of AI, with lots of important applications, ranging from pattern recognition to prediction to diagnosis. Probabilistic approaches to classification are particularly popular and effective. In particular, the naive Bayes (e.g., [8, Chap. 17]) assumes conditional independence for the features given the class. Despite the generally good performances of this classifier, these assumptions are often unrealistic and other models with less restrictive assumptions have been proposed. These can be expressed in the framework of *Bayesian networks* [11] by directed graphs.

Besides classifiers based on special topologies (e.g., tree-augmented [6]), *structural learning* algorithms (e.g., K2 [8, Chap. 18]) can learn the network structure from data. Regarding the learning of the parameters, this can be either based on Bayesian (e.g., a uniform Dirichlet prior) or frequentist (maximum-likelihood) approaches. The latter is unbiased and independent from the prior specification, but generally lead to inferior classification performances, especially on data sets where the contingency tables, which contain the counts of the joint occurrences of specific values of the features and the class, are characterised by several zeros [6]. To obtain more reliable estimates for the Bayesian network parameters, a *likelihood-based* approach [3, 10] can be considered. This is a generalization of the frequentist approach towards *imprecise probabilities* [12], i.e., robust models based on sets of probability distributions. Loosely speaking, the idea is to consider, instead of the single maximum-likelihood estimator, all the models whose likelihood is above a certain threshold level. When applied to classification with Bayesian networks, this approach produces a classifier based, instead of a single, on a collection of Bayesian networks (with the same topology) or, in other words, a *credal network* [5]. If different Bayesian networks associated to the classifier assign different classes on a same test instance, the classifier returns all these classes. This is an example of *credal classification*, comparable with those proposed in [4], being in fact an extension of what we proposed in [1] for the naive case. To the best of our knowledge, this is the first credal classifier for general topologies.[1] A notable feature of our classifier is that, in the likelihood evaluation, we also consider the test instance with missing value for the class. This is important to obtain more accurate classification performances when coping with zero counts. The paper is organised as follows. We review background material about classification with Bayesian networks (Sect. 2.1) and likelihood-based approaches (Sect. 2.2). Then, in Sect. 3.1, our approach is presented by means of a simple example. Discussion on how to cope with zero counts is in Sect. 3.2, while Sect. 3.3 reports the formula for the classifier. The classifier performances are empirically tested in Sect. 4. Conclusions and outlooks are finally in Sect. 5.

2 Background

2.1 Classification with Bayesian Networks

Consider a set of variables $\mathbf{X} := (X_0, X_1, \ldots, X_n)$, with X_i taking values in a finite set Ω_{X_i}, for each $i = 0, 1, \ldots, n$. Regard X_0 as the *class* and other variables as *features* of a classification task based on a data set of joint observations, i.e., $\mathcal{D} := \{(x_0^{(j)}, x_1^{(j)}, \ldots, x_n^{(j)})\}_{j=1}^N$. A *classifier* is an algorithm assigning a class label $x_0^* \in \Omega_{X_0}$ to a generic test instance $(\tilde{x}_1, \ldots, \tilde{x}_n)$ of the features. In particular, *probabilistic classifiers* learn from data a joint probability mass function $P(X_0, \ldots, X_n)$ and, with 0-1 losses, assign to the test instance the class label:

$$x_0^* := \arg \max_{x_0 \in \Omega_{X_0}} P(x_0 | \tilde{x}_1, \ldots, \tilde{x}_n). \tag{1}$$

The learning of a joint mass function from the data $\mathcal{D}$ can be approached within the framework of *Bayesian networks* [11]. A Bayesian network induces a compact specification of the joint based on independencies among its variables. These are depicted by directed acyclic graphs with nodes in one-to-one correspondence with the variables in $\mathbf{X}$. Markov condition gives semantics: every variable is conditionally independent of its non-descendants non-parents given its parents. *Structural learning* algorithms [8, Chap. 18] can learn the graph modeling independencies in this way. Let $\mathcal{G}$ be this graph. For each $i = 0, \ldots, n$, denote by Π_i the parents of X_i according to $\mathcal{G}$. The factorization induced by these independencies is:

$$P(x_0, x_1, \ldots, x_n) = \prod_{i=0}^{n} P(x_i | \pi_i), \tag{2}$$

where π_i is the value of Π_i consistent with $(x_0, x_1, \ldots, x_n)$. To do classification, i.e., to assign a class label as in (1) to the test instance, we check, for each $x_0', x_0'' \in \Omega_{X_0}$, whether or not:

$$\frac{P(x_0' | \tilde{x}_1, \ldots, \tilde{x}_n)}{P(x_0'' | \tilde{x}_1, \ldots, \tilde{x}_n)} = \frac{P(x_0', \tilde{x}_1, \ldots, \tilde{x}_n)}{P(x_0'', \tilde{x}_1, \ldots, \tilde{x}_n)} > 1. \tag{3}$$

This inequality can be rewritten as:

$$\frac{P(x_0' | \tilde{\pi}_0)}{P(x_0'' | \tilde{\pi}_0)} \cdot \prod_{i=1}^{n} \frac{P(\tilde{x}_i | \tilde{\pi}_i')}{P(\tilde{x}_i | \tilde{\pi}_i'')} = \frac{P(x_0' | \tilde{\pi}_0)}{P(x_0'' | \tilde{\pi}_0)} \prod_{i=1,\ldots,n : X_0 \in \Pi_i} \frac{P(\tilde{x}_i | x_0', \tilde{\pi}_i)}{P(\tilde{x}_i | x_0'', \tilde{\pi}_i)} > 1, \tag{4}$$

where $\tilde{\pi}_0$ is the value of the parents of X_0 consistent with $(\tilde{x}_1, \ldots, \tilde{x}_n)$; $\tilde{\pi}_i'$ and $\tilde{\pi}_i''$ are the values of Π_i consistent, respectively, with $(x_0', \tilde{x}_1, \ldots, \tilde{x}_n)$ and $(x_0'', \tilde{x}_1, \ldots, \tilde{x}_n)$ (for each $i = 1, \ldots, n$), and

[1] Other credal classifiers are based on the *imprecise Dirichlet model*, but there are no classification algorithms for general topologies [4].

the presence of X_0 among the parents of X_i is emphasized in the second product where (with a small abuse of notation) $\tilde{\pi}_i$ denote the state of $\Pi_i \setminus \{X_0\}$ consistent with $(\tilde{x}_1, \ldots, \tilde{x}_n)$. The first derivation in (4) follows from (2), the second comes from the fact that the terms in the products associated to variables X_i which are not children of X_0 (nor X_0 itself) are one. Hence, when doing classification with Bayesian networks, we can focus on the *Markov blanket* of X_0 (Fig. 1), i.e., (i) the class X_0; (ii) the parents of X_0; (iii) the children of X_0; and (iv) the parents of the children of X_0.

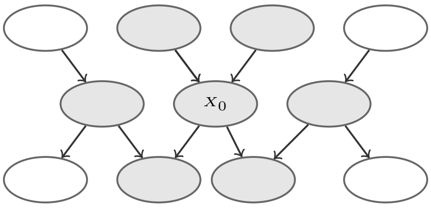

Figure 1: A Bayesian network. The nodes of the Markov blanket of X_0 are in grey.

Regarding the quantification of the conditional probabilities in (2) (or, after the above discussion, only of those in the Markov blanket of X_0) from data, standard techniques can be adopted. For the conditional $P(X_i|\pi_i)$, a Bayesian approach with Dirichlet prior with parameter st_i, for each $x_i \in \Omega_{X_i}$, would give:

$$P(x_i|\pi_i) := \frac{n(x_i, \pi_i) + st_i}{n(\pi_i) + s}, \tag{5}$$

for each $i = 1, \ldots, n$, $x_i \in \Omega_{X_i}$, $\pi_i \in \Omega_{\Pi_i}$, where $n(\cdot)$ is a *count function* returning the counts for the data in $\mathcal{D}$ satisfying the event specified in its argument. Similarly, a frequentist (maximum-likelihood) approach would use expression (5) with $s = 0$. These approaches are known to produce potentially unreliable estimates if only few data are available, this being particularly true if zero counts occur. An extension of the frequentist approach to partially overcome these problems is presented in the next section.

2.2 Likelihood-Based Learning of Imprecise-Probabilistic Models

Likelihood-based approaches [3, 10] are an extension of frequentist approaches intended to learn sets, instead of single, distributions, from data, this making the corresponding parameters estimates more robust and hence reliable. The basic idea is to start with a collection of candidate models, and then keep only those assigning to the available data a probability beyond a certain threshold. We introduce this with the following example.

Example 1. *Consider a Boolean X, for which N observations are available, and n of them report true. If $\theta \in [0,1]$ is the chance that X is true, likelihood of data is $L(\theta) := \theta^n \cdot (1-\theta)^{N-n}$ and its maximum $\theta^* = {}^{n}/_{N}$. For each $\alpha \in [0,1]$, consider the values of θ such that $L(\theta) \geq \alpha L(\theta^*)$. Fig. 2 depicts the behaviour of these probability (which can be seen as confidence [7]) intervals for increasing N.*

The above technique can be extended to the general case, and interpreted as a learning procedure [2, 9] in the following sense. Consider a *credal set* $\mathbf{P}$, i.e., a collection of probability distributions all over the same variable. Assume the elements of $\mathbf{P}$ are indexed by parameter $\theta \in \Theta$, i.e., $\mathbf{P} := \{P_\theta\}_{\theta \in \Theta}$. Given the data $\mathcal{D}$, consider the normalised likelihood:

$$L(\theta) := \frac{P_\theta(\mathcal{D})}{\sup_{\theta' \in \Theta} P_{\theta'}(\mathcal{D})}, \tag{6}$$

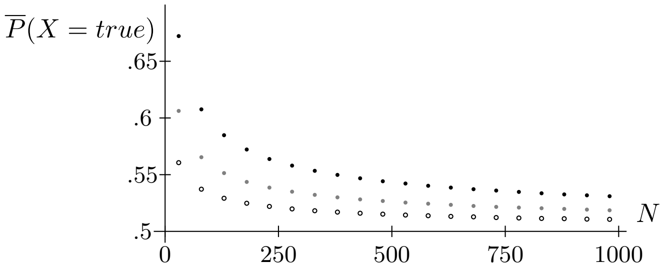

Figure 2: Probability intervals obtained by likelihood-based learning for different values of α for Ex. 1. The plot shows the upper bounds of the interval probability that the variable is true as a function of the sample size N, when $n/N = 1/2$ (lower bounds are symmetric). Black, gray and white points refer, respectively, to $\alpha = .8, .5, .15$.

likelihood-based learning consists in removing from $\mathbf{P}$ the distributions whose normalised likelihood is below a threshold. Thus, given $\alpha \in [0, 1]$, we consider the (smaller) credal set:

$$\mathbf{P}_\alpha := \{P_\theta\}_{\theta \in \Theta : L(\theta) \geq \alpha}. \tag{7}$$

Note that $\mathbf{P}_{\alpha=1}$ is a "precise" credal set including only the maximum-likelihood distribution, while $\mathbf{P}_{\alpha=0} = \mathbf{P}$. Likelihood-based learning is said to be *pure*, if the credal set $\mathbf{P}$ includes all the possible distributions that can be specified over the variable under consideration.

3 Robust Likelihood-Based Classifiers

3.1 A Demonstrative Example

Consider a classification task as in Sect. 2.1 with a single feature and both variables Boolean. Assuming that $X_0 \to X_1$ is the graph obtained from the data (note that this models no independence at all), (2) rewrites as:

$$P(x_0, x_1) := P(x_1|x_0) \cdot P(x_0), \tag{8}$$

for each x_0, x_1. As a probability mass function over a Boolean variable can be specified by a single parameter, all Bayesian networks over this graph are parametrized by $\theta = (\theta_1, \theta_2, \theta_3)$ with $\theta_1 := p(x_0)$, $\theta_2 := p(x_1|x_0)$, $\theta_3 := p(x_1|\neg x_0)$. Let P_θ denote the corresponding joint distribution as in (8). A pure likelihood-based approach consists in starting from $\Theta := [0, 1]^3 \subseteq \mathbb{R}^3$. The data set $\mathcal{D}$ to be used to refine this credal set can be equivalently characterized by four counts, i.e., $n_1 := n(x_0, x_1)$, $n_2 := n(x_0, \neg x_1)$, $n_3 := n(\neg x_0, x_1)$, $n_4 := (\neg x_0, \neg x_1)$. The corresponding likelihood, i.e.,

$$L(\theta) \propto (\theta_1 \cdot \theta_2)^{n_1} \cdot (\theta_1 \cdot (1 - \theta_2))^{n_2} \cdot ((1 - \theta_1) \cdot \theta_3)^{n_3} \cdot ((1 - \theta_1) \cdot (1 - \theta_3))^{n_4}, \tag{9}$$

attains its maximum when the parameters are estimated by the relative frequencies. For a more robust parameters estimation, given $\alpha \in [0, 1]$, all the quantifications satisfying (7) can be considered. A collection of Bayesian networks (all over the same graph), i.e., a *credal network* [5] is therefore considered as a more robust and reliable model of the process generating the data. This model can be used for classification. Yet, when evaluating the ratio as in (3), different P_θ can return different classes.

To decide whether or not a class is dominating another one, a possible, conservative, approach consists in assuming that a probability dominates another one if and only if this is true for all the distributions. In practice we extend (3) to sets of distributions as follows:

$$\inf_{P_\theta \in \mathbf{P}_\alpha} \frac{P_\theta(x_0'|\tilde{x}_1,\ldots,\tilde{x}_n)}{P_\theta(x_0''|\tilde{x}_1,\ldots,\tilde{x}_n)} > 1. \tag{10}$$

This is a well-known decision criterion for imprecise-probabilistic models called *maximality* [12]. Unlike (3), testing dominance with (10) for each pair of classes can lead to multiple undominated classes. This produces a *credal* classifier, which can assign more than a class to test instances. To check (10), the likelihood should be evaluated as a function of ratio (3). This can be done by sampling as in Fig. 3. Yet, different models with different likelihoods can have the same ratio, i.e., the function is not single-valued. Nevertheless, to check dominance it is sufficient to determine whether or not the models (points) with ordinate (likelihood) greater than α all have a ratio (abscissa) greater than one. To do that, it is possible to focus on the left-most point (α-cut) where the likelihood upper envelope is α.

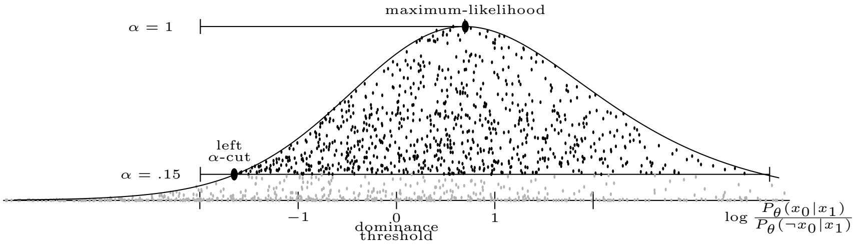

Figure 3: Likelihood-based classification of a test instance $X_1 = x_1$ for model in Sect. 3.1 with $[n_1, n_2, n_3, n_4] = [2, 2, 1, 3]$. Black line is the upper envelope (see Sect. 3.3). For this instance, (10) is not satisfied, i.e., x_0 does not dominate $\neg x_0$.

3.2 Coping with Zero Counts

In the above described procedure, likelihood was identified with the probability $P(\mathcal{D}|\theta)$ assigned by the Bayesian network associated to θ to the data. Yet, if there is an attribute $\tilde{x}_i$ in the test instance such that the relative counts are zero, the corresponding maximum-likelihood probability is zero, this preventing dominances in (10). Similar problems occur even within the Bayesian framework [11]. To prevent this issue, the test instance $(\tilde{x}_1,\ldots,\tilde{x}_n)$ can be regarded as an incomplete datum with class X_0 missing. This instance can be involved in the likelihood evaluation, i.e.,

$$L'(\theta) \propto P(\mathcal{D}|\theta) \cdot \sum_{x_0 \in \Omega_{X_0}} P_\theta(x_0, \tilde{x}_1, \ldots, \tilde{x}_n). \tag{11}$$

The maximum-likelihood estimate of θ can be calculated by means of the EM-algorithm, which completes the test instance with fractional counts for the different values of X_0. We denote by $\hat{n}(\cdot)$ the counting function obtained by augmenting the counts relative to $\mathcal{D}$ with these fractional counts. Note that, while $n() = N$, $\hat{n}() = N + 1$.

3.3 Analytic Formulae for the Upper Envelope of the Likelihood

Following the approach in Sect. 3.1 we would like to derive here an analytic expression for the upper envelope of the likelihood (11) as a function of the probability ratio (3). Each point of this upper envelope corresponds to a particular quantification P_θ of the Bayesian network. If $\theta(t)$ is a function of $t \in [a,b]$ such that there is a one-to-one correspondence between the quantifications $P_{\theta(t)}$ and the points of the upper envelope of the likelihood (when t varies in $[a,b]$), then

$$\left\{ \left(\frac{P_{\theta(t)}(x'_0, \tilde{x}_1, \ldots, \tilde{x}_n)}{P_{\theta(t)}(x''_0, \tilde{x}_1, \ldots, \tilde{x}_n)}, L'\left(\theta(t)\right) \right) : t \in [a,b] \right\} \tag{12}$$

is a parametric expression for the graph of the desired upper envelope.

A function $\theta(t)$ with the above property was obtained in [3] for the case when the test instance is not considered in the likelihood, i.e., for the case without summation in (11). This result is not directly applicable to the likelihood (11), but we can obtain an approximation of the desired upper envelope if we use the function $\hat{\theta}(t)$ resulting from the expected likelihood delivered by the EM-algorithm, i.e., the likelihood corresponding to the augmented counts $\hat{n}(\cdot)$. Our approximation is then given by (12) with $\hat{\theta}(t)$ instead of $\theta(t)$.

In order to simplify the analytic formulae, we assume that the children of X_0 in the Bayesian network are denoted by $X_1, \ldots, X_k$ (with $k \leq n$). We first define:

$$\begin{aligned} a &:= -\min\left\{\hat{n}(x'_0, \tilde{\pi}_0), \hat{n}(\tilde{x}_1, x'_0, \tilde{\pi}_1), \ldots, \hat{n}(\tilde{x}_k, x'_0, \tilde{\pi}_k)\right\}, & (13) \\ b &:= \min\left\{\hat{n}(x''_0, \tilde{\pi}_0), \hat{n}(\tilde{x}_1, x''_0, \tilde{\pi}_1), \ldots, \hat{n}(\tilde{x}_k, x''_0, \tilde{\pi}_k)\right\}. & (14) \end{aligned}$$

For each $t \in [a,b]$, let us consider the following functions:

$$\begin{aligned} x(t) &:= \frac{\hat{n}(x'_0, \tilde{\pi}_0) + t}{\hat{n}(x''_0, \tilde{\pi}_0) - t} \cdot \prod_{i=1}^{k} \frac{\frac{\hat{n}(\tilde{x}_i, x'_0, \tilde{\pi}_i) + t}{\hat{n}(x'_0, \tilde{\pi}_i) + t}}{\frac{\hat{n}(\tilde{x}_i, x''_0, \tilde{\pi}_i) - t}{\hat{n}(x''_0, \tilde{\pi}_i) - t}}, & (15) \\ y(t) &:= y_0(t) \cdot \left[\sum_{x_0 \in \Omega_{X_0}} k_{x_0}(t) \right], & (16) \end{aligned}$$

where:

$$\begin{aligned} y_0(t) := \; & [\hat{n}(x'_0, \tilde{\pi}_0) + t]^{n(x'_0, \tilde{\pi}_0)} \cdot [\hat{n}(x''_0, \tilde{\pi}_0) - t]^{n(x''_0, \tilde{\pi}_0)} \\ & \cdot \prod_{i=1}^{k} \left[\frac{[\hat{n}(\tilde{x}_i, x'_0, \tilde{\pi}_i) + t]^{n(\tilde{x}_i, x'_0, \tilde{\pi}_i)}}{[\hat{n}(x'_0, \tilde{\pi}_i) + t]^{n(x'_0, \tilde{\pi}_i)}} \cdot \frac{[\hat{n}(\tilde{x}_i, x''_0, \tilde{\pi}_i) - t]^{n(\tilde{x}_i, x''_0, \tilde{\pi}_i)}}{[\hat{n}(x''_0, \tilde{\pi}_i) - t]^{n(x''_0, \tilde{\pi}_i)}} \right], \end{aligned} \tag{17}$$

$$k_{x_0}(t) := \begin{cases} [\hat{n}(x_0, \tilde{\pi}_0) + t] & \cdot \prod_{i=1}^{k} \frac{\hat{n}(\tilde{x}_i, x_0, \tilde{\pi}_i) + t}{\hat{n}(x_0, \tilde{\pi}_i) + t} \quad \text{if} \quad x_0 = x'_0, \\ [\hat{n}(x_0, \tilde{\pi}_0) - t] & \cdot \prod_{i=1}^{k} \frac{\hat{n}(\tilde{x}_i, x_0, \tilde{\pi}_i) - t}{\hat{n}(x_0, \tilde{\pi}_i) - t} \quad \text{if} \quad x_0 = x''_0, \\ \hat{n}(x_0, \tilde{\pi}_0) & \cdot \prod_{i=1}^{k} \frac{\hat{n}(\tilde{x}_i, x_0, \tilde{\pi}_i)}{\hat{n}(x_0, \tilde{\pi}_i)} \quad \text{if} \quad x_0 \in \Omega_{X_0} \setminus \{x'_0, x''_0\}. \end{cases} \tag{18}$$

Theorem 1. *If $[x(a), x(b)] = [0, +\infty]$, our approximation of the upper envelope of the normalized likelihood (11) as a function of the probability ratio (3) is parametrized by $(x(t), {}^{y(t)}/_{y(0)})$ with $t \in [a,b]$.*

Theorem 2. *If $x(a) > 0$, the parametrization in Th. 1 holds in the region $[x(a), x(b)]$, while in the region $[0, x(a)]$, a parametrization is $(\tau \cdot x(a), {}^{y'(\tau)}/_{y(0)})$ with $\tau \in [0,1]$ and:*

$$y'(\tau) := \tau^{-a+n(x'_0)-\hat{n}(x'_0)} \cdot y_0(a) \cdot \left[\tau \, k_{x'_0}(a) + \sum_{x_0 \in \Omega_{X_0} \setminus \{x'_0\}} k_{x_0}(a) \right]. \tag{19}$$

The proofs of the two theorems are in the appendix. As a simple application of these results, it is straightforward to evaluate the upper envelope of the likelihood for the example in Fig. 3 when only the complete data are considered in the likelihood, i.e., only $y_0(t)$ is considered in (17). In this case, $t \in [-2, 1]$ and:

$$\left[\begin{array}{c} x(t) \\ y(t) \end{array} \right] = \left[\begin{array}{c} (2+t) \cdot (1-t)^{-1} \\ (2+t)^2 \cdot (1-t) \end{array} \right]. \tag{20}$$

Given the above parametrization of the likelihood upper envelope, classification can be performed by checking whether or not the left α-cut has abscissa greater than one. For the situation in Th. 1, this can be numerically done in few iteration by bracketing the (unique) zero of $g(t) := y(t) - \alpha y(0)$ in the region $t \in [a, 0]$, unless the corresponding bounds on $x(t)$ are greater (or smaller) than one (and similarly proceed for Th. 2).

4 Preliminary results

To describe the performance of a credal classifier, multiple indicators are needed. We adopt the following:

- *determinacy*: percentage of instances classified with a single class;
- *single accuracy* : accuracy over instances classified with a single class;
- *set-accuracy*: accuracy over instances classified with more classes;
- *indeterminate output size*: average number of classes returned when the classification is indeterminate.

Roughly speaking, a credal classifier identifies *easy* instances, over which it returns a single class, and *hard* instances, over which it returns more classes. A credal classifier is effective at recognizing hard instances if its precise counterpart undergoes a considerable drop of accuracy on them. As a counterpart of the likelihood-based classifier we consider a Bayesian network with the same graph, but whose parameters are learned precisely as in (5); this model is referred to as the *standard network* in the following. The considered data sets and their main characteristics are shown in Tab. 1. We run 5 runs of 5 folds cross-validation, for a total of 25 training/test experiments on each data set.

Dataset		Iris	Glass	Ecoli	Breast	Haberman	Diabetes	Ionosphere
Size	N	150	214	336	699	306	768	351
Features	k	4	7	6	9	2	6	33
Classes	$\lvert\Omega_{X_0}\rvert$	3	7	8	2	2	2	2

Table 1: Main characteristics of the data sets.

The determinacy of the likelihood-based classifier (with $\alpha = 0.15$) is generally around 90% or higher, as shown in the left plot of Fig. 4; in general, the determinacy increases on larger data sets. The likelihood-based classifier is effective at detecting hard instances; this can be appreciated by the right plot of Fig. 4, which compares the accuracy obtained by the standard network on the instances recognized as easy and hard by the likelihood-based classifier. The accuracy of the standard network clearly drops on the instances indeterminately classified by the likelihood-based model; the drop is statistically significant (Wilcoxon signed-rank test, p-value < 0.01).

The set-accuracy and the indeterminate output size are meaningful only on data sets with more than 2 classes. On such data sets, the likelihood-based classifier returns a number of classes which is on average 58% of the total classes, achieving on average a set-accuracy of 84%.

In future more extensive experiments should be carried out, comparing the likelihood-based model against credal classifiers already present in literature.

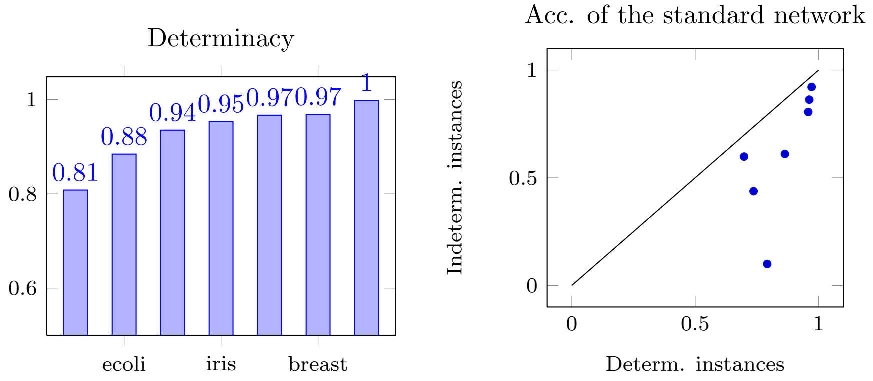

Figure 4: Experimental results: determinacy of the likelihood-based classifier (*left*) and comparison of the accuracy achieved by the standard network on the instances classified determinately and indeterminately by the likelihood-based classifier (*right*) .

5 Conclusions and Outlooks

A new, likelihood-based approach, to classification with Bayesian networks has been proposed. Instead of the single maximum-likelihood estimation of the network parameters, all the parametrizations assigning to the available data a likelihood beyond a given threshold are considered. All the classes which are optimal at least for a network parametrization consistent with this constraint are returned. This corresponds to a *credal classifier* which can eventually assign more than a single class label to the test instance. Preliminary experiments show that this approach is successful in discriminating hard- from easy-to-classify instances. In the latter case the single, correct, class label is returned, while for hard instances a set of classes, generally including the correct one, is returned. As a future work, we intend to compare this model with other credal classifiers and extend this approach to incomplete data sets.

Proof of the Theorems

Proof of Th. 1. Consider the function $\hat{\theta}(t)$ resulting from the expected likelihood delivered by the EM-algorithm, i.e., the likelihood corresponding to the augmented counts $\hat{n}(\cdot)$. Th. 1 of [3] implies that $P_{\hat{\theta}(t)}$ is the quantification corresponding to the relative frequencies described by $\hat{n}(\cdot)$ when we add t to the counts including x'_0 and we subtract t from the counts including x''_0, where $t \in [a, b]$. Hence, using the simplification (4), we obtain:

$$\frac{P_{\hat{\theta}(t)}(x'_0,\tilde{x}_1,\ldots,\tilde{x}_n)}{P_{\hat{\theta}(t)}(x''_0,\tilde{x}_1,\ldots,\tilde{x}_n)} = \frac{P_{\hat{\theta}(t)}(x'_0|\tilde{\pi}_0)}{P_{\hat{\theta}(t)}(x''_0|\tilde{\pi}_0)} \cdot \prod_{i=1}^{k} \frac{P_{\hat{\theta}(t)}(\tilde{x}_i|x'_0,\tilde{\pi}_i)}{P_{\hat{\theta}(t)}(\tilde{x}_i|x''_0,\tilde{\pi}_i)} = \frac{\frac{\hat{n}(x'_0,\tilde{\pi}_0)+t}{\hat{n}(\tilde{\pi}_0)}}{\frac{\hat{n}(x''_0,\tilde{\pi}_0)-t}{\hat{n}(\tilde{\pi}_0)}} \cdot \prod_{i=1}^{k} \frac{\frac{\hat{n}(\tilde{x}_i,x'_0,\tilde{\pi}_i)+t}{\hat{n}(x'_0,\tilde{\pi}_i)+t}}{\frac{\hat{n}(\tilde{x}_i,x''_0,\tilde{\pi}_i)-t}{\hat{n}(x''_0,\tilde{\pi}_i)-t}} = x(t). \tag{21}$$

The result $L'(\hat{\theta}(t)) \propto y(t)$ can be proved analogously. The likelihood $L'(\hat{\theta}(t))$ is maximal in $t = 0$, because $P_{\hat{\theta}(0)}$ is the quantification corresponding to the relative frequencies described by $\hat{n}(\cdot)$, i.e., $\hat{\theta}(0)$ is the maximum-likelihood estimate of θ. Therefore, our approximation of the upper envelope of the normalized likelihood is parametrized by:

$$\left(\frac{P_{\hat{\theta}(t)}(x_0', \tilde{x}_1, \ldots, \tilde{x}_n)}{P_{\hat{\theta}(t)}(x_0'', \tilde{x}_1, \ldots, \tilde{x}_n)}, \frac{L'(\hat{\theta}(t))}{L'(\hat{\theta}(0))} \right) = \left(x(t), \frac{y(t)}{y(0)} \right), \tag{22}$$

with $t \in [a, b]$. □

Proof of Th. 2. From the definitions of a and $x(t)$ it follows that when $x(a) > 0$, there must be a $j \in \{1, \ldots, k\}$ such that:

$$-a = \hat{n}(\tilde{x}_j, x_0', \tilde{\pi}_j) = \hat{n}(x_0', \tilde{\pi}_j) = n(x_0', \tilde{\pi}_j) + \hat{n}(x_0') - n(x_0'). \tag{23}$$

In general, Th. 1 of [3] implies that the function $\hat{\theta}(t)$ leads to our approximation (22) of the upper envelope of the normalized likelihood in the region $[x(a), x(b)]$, while in order to continue the same approximation in the region $[0, x(a)]$, we must modify $\hat{\theta}(a)$ as follows: For each $\tau \in [0, 1]$, let $\hat{\theta}(a, \tau)$ be equal to $\hat{\theta}(a)$ except for the conditional probability distribution of X_j given $\Pi_j = (x_0', \tilde{\pi}_j)$, which must satisfy $P_{\hat{\theta}(a,\tau)}(\tilde{x}_j | x_0', \tilde{\pi}_j) = \tau$ (while the probabilities of the other possible values of X_j can be arbitrarily chosen). With this definition we obtain:

$$\frac{P_{\hat{\theta}(a,\tau)}(x_0', \tilde{x}_1, \ldots, \tilde{x}_n)}{P_{\hat{\theta}(a,\tau)}(x_0'', \tilde{x}_1, \ldots, \tilde{x}_n)} = \tau \cdot x(a), \tag{24}$$

$$L'(\hat{\theta}(a, \tau)) \propto \tau^{n(x_0', \tilde{\pi}_j)} \cdot y_0(a) \cdot \left[\tau \, k_{x_0'}(a) + \sum_{x_0 \in \Omega_{X_0} \setminus \{x_0'\}} k_{x_0}(a) \right] = y'(\tau). \tag{25}$$

Therefore, our approximation of the upper envelope of the normalized likelihood in the region $[0, x(a)]$ is parametrized by:

$$\left(\frac{P_{\hat{\theta}(a,\tau)}(x_0', \tilde{x}_1, \ldots, \tilde{x}_n)}{P_{\hat{\theta}(a,\tau)}(x_0'', \tilde{x}_1, \ldots, \tilde{x}_n)}, \frac{L'(\hat{\theta}(a, \tau))}{L'(\hat{\theta}(0))} \right) = \left(\tau \cdot x(a), \frac{y'(\tau)}{y(0)} \right), \tag{26}$$

with $\tau \in [0, 1]$. □

References

[1] A. Antonucci, M. Cattaneo, and G. Corani. Likelihood-based naive credal classifier. In *ISIPTA '11*, pages 21–30. SIPTA, 2011.

[2] M. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich, 2007.

[3] M. Cattaneo. Likelihood-based inference for probabilistic graphical models: Some preliminary results. In *PGM 2010*, pages 57–64. HIIT Publications, 2010.

[4] G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Elsevier, 2010.

[5] F.G. Cozman. Credal networks. *Artif. Intell.*, 120:199–233, 2000.

[6] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Mach. Learn.*, 29:131–163, 1997.

[7] D.J. Hudson. Interval estimation from the likelihood function. *J. R. Stat. Soc., Ser. B*, 33:256–262, 1971.

[8] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

[9] S. Moral. Calculating uncertainty intervals from conditional convex sets of probabilities. In *UAI '92*, pages 199–206. Morgan Kaufmann, 1992.

[10] Y. Pawitan. *In All Likelihood*. Oxford University Press, 2001.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

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