2nd International Symposium on Imprecise Probabilities and Their Applications, Ithaca, New York, 2001

Statistical inference of the naive credal classifier

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

Abstract

In the wish list of the characteristics of a classifier, there are a reliable approach to small data sets and a clear and robust treatment of incomplete samples. This paper copes with such difficult problems by adopting the paradigm of credal classification. By exploiting Walley's imprecise Dirichlet model, it defines how to infer the naive credal classifier from a possibly incomplete multinomial sample. The derived procedure is exact and linear in the number of attributes. The obtained classifier is robust to small data sets and to all the possible missingness mechanisms. The results of some experimental analyses that compare the naive credal classifier with naive Bayesian models support the presented approach.

1 Introduction

Classification (also known as pattern recognition, identification, or selection) is a multivariate technique concerned with allocating new objects to previously defined groups on the basis of observations on several characteristics of the objects [4]. Formally, a *classifier* is a function that maps an instance of a set of variables, called attributes or features, to a state of a categorical class variable. The range of application of classifiers is wide and comprises, among others, pattern recognition, prediction and diagnosis.

The *naive credal classifier* (NCC) [21, 23] is the extension of the naive Bayes classifier (NBC) [3] to sets of probability distributions (or *credal sets* [13]). The NCC is an example of a *credal classifier*: this is a function that maps an instance of a set of features to a set of states of a categorical class variable. A credal classifier enables imprecision to be taken into account, as generated by unobserved or rare events, small sample sizes and missing data. As a consequence, for a given pattern of the attributes, imprecision in the input may prevent a single output class from being obtained; then the result of a credal classifier is a set of classes, all of which are candidates to be the correct category. In other words, a credal classifier recognizes that the available knowledge may not suffice to isolate a single class and thus gives rise to a set of alternatives.

In this paper I cope with the statistical inference of the NCC from a possibly incomplete multinomial sample. I initially exploit Walley's *imprecise Dirichlet model* (IDM) [20] to derive the expressions defining the NCC as inferred from a complete sample, in Sect. 2. This is firstly made by imposing the assumption of probabilistic independence of the attributes conditional on the class that is central both to the NBC and the NCC. Secondly, the choice of the class of prior densities characterizing the IDM completes the definition of the NCC. This derivation is an original development that improves upon an early proposal to infer the NCC from a complete sample [21]. This proposal defined the NCC by means of a set of joint distributions that properly encompasses the set proposed in this paper: the past set was, in other words, less precise. Greater precision is obtained here by implicitly relaxing the assumption made in the past work of dealing with *logically independent* credal sets (also known as *separately specified* credal sets [19]).

The focus of the paper is then moved to incomplete samples. By regarding an incomplete sample as a collection of complete samples [22], I extend the above expressions to the case of missing data in Sect. 2.5.1. The classifier inferred in this way is robust to all the possible replacements of missing data with admissible values. To the best of my knowledge, only another classifier has an analogous characteristic [18], though it cannot be considered a credal classifier as it is, since it does not output sets of classes (also, it neglects the imprecision due to the prior uncertainty on the multinomial model). The problem of missing data is widely recognized as one of the most critical and important topics in the field of classification [1]; the approach proposed here seems a significant step towards the

proper treatment of this problem.

It is also important to emphasize that the presented approach to inference does not involve approximations. Nevertheless, both the computational complexity of inferring the classifier and that of doing a classification are linear in the number of attributes, as shown in Sect. 4.

The proposed method of inference is discussed by means of some experimental analyses on real and artificial data sets. These permit to introduce the problem of the experimental evaluation of a credal classifier that is not a straightforward extension of the case of standard classifiers. Furthermore, the experiments, reported in Sect. 5, point to some counter-intuitive situations that are interpreted in Sect. 6.

2 Inferring the NCC by the IDM

This paper infers the NCC by exploiting Walley's imprecise Dirichlet model [20]. The IDM models prior ignorance about the chances of the multinomial distribution by a set of Dirichlet distributions and makes posterior inferences by combining this set with the observed likelihood function. In the following subsections I consider a complete random sample.

2.1 Notation and basic assumption

Let us denote the classification variable by C, taking values in the finite set $\mathcal{C}$, where the possible classes are denoted by lower-case letters. We measure k features $(A_1, \ldots, A_k)$ taking generic values $(a_1, \ldots, a_k) = \mathbf{a}$ from the sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$, which are assumed to be finite.

Let the unknown chances of the multinomial distribution be denoted by $\theta_{c,\mathbf{a}}$ $((c, \mathbf{a}) \in \mathcal{C} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k)$. Denote by $\theta_{a_i|c}$ the chance that $A_i = a_i$ conditional on c; similarly, let $\theta_{\mathbf{a}|c}$ be the chance that $(A_1, \ldots, A_k) = (a_1, \ldots, a_k)$ conditional on c.

Let N be the total number of observed units, each with known values of the attributes and the class. The units are assumed to be generated independently from the multinomial process. Let $n(c)$ and $n(a_i\,|c)$ be the observed frequencies of class c and of $(a_i\,|c)$ in the N observations, respectively. We have the structural constraints: $0 \leq n(a_i\,|c) \leq n(c)$ for all c and $(a_i\,|c)$; $\sum_c n(c) = N$; and $\sum_{a_i \in \mathcal{A}_i} n(a_i\,|c) = n(c)$ for all c and i. $t(c)$ and $t(a_i\,|c)$ are used to denote the corresponding values of the t-hyperparameters in the IDM.

Let $\mathbf{n}$ denote the vector of all the above frequencies, $\mathbf{t}$ the corresponding vector of all the t-hyperparameters and $\boldsymbol{\theta}$ be the vector whose elements are the chances $\theta_{c,\mathbf{a}}$ $((c, \mathbf{a}) \in \mathcal{C} \times \mathcal{A}_1 \cdots \times \mathcal{A}_k)$.

Both the naive Bayes classifier and the naive credal classifier are based on the assumption of probabilistic independence of the attributes conditional on the class (this assumption is widely discussed in literature [2, 6]):

$$\theta_{\mathbf{a}|c} = \prod_{i=1}^{k} \theta_{a_i|c} \quad \forall\, (c, \mathbf{a}) \in \mathcal{C} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k. \quad (1)$$

Inferring the NCC means to determine a set of joint distributions from data, each satisfying the mutual independence of the attributes conditional on the class.

2.2 Likelihood function

Under assumption (1), the chance $\theta_{c,\mathbf{a}}$ that a unit will have values $(c, \mathbf{a})$ can be expressed as a product of theta-parameters:

$$\theta_{c,\mathbf{a}} = \theta_c \prod_{i=1}^{k} \theta_{a_i|c}\,. \quad (2)$$

Hence, after observing data $\mathbf{n}$, the likelihood function for the problem can be expressed as a product of powers of the theta-parameters:

$$L(\boldsymbol{\theta}|\mathbf{n}) \propto \prod_{c \in \mathcal{C}} \left[\theta_c^{n(c)} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{n(a_i|c)} \right].$$

2.3 Definition of the IDM

The prior densities in the IDM class are proportional to an expression that is similar to the likelihood function, except that frequencies $n(\cdot)$ are replaced everywhere by $st(\cdot) - 1$, where $s > 0$ is a fixed constant and $t(\cdot)$ is the hyperparameter that corresponds to frequency $n(\cdot)$, i.e.,

$$f\,(\boldsymbol{\theta}\,|\mathbf{t}, s\,) \propto \prod_{c \in \mathcal{C}} \left[\theta_c^{st(c)-1} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{st(a_i|c)-1} \right].$$

This is a product of Dirichlet prior densities. If we take the prior class to consist of densities of this type, we can get different classes by imposing different constraints on the hyperparameters $\mathbf{t}$. I will use the following constraints:

$$\sum_c t(c) = 1 \quad (3)$$

$$\sum_{a_i \in \mathcal{A}_i} t(a_i\,|c) = t(c) \quad \forall (i, c) \quad (4)$$

$$t(a_i\,|c) > 0 \qquad \forall (i, a_i, c). \quad (5)$$

These constraints correspond exactly to the structural constraints that are satisfied by the observed

frequencies $\mathbf{n}$ and this is natural because $st(\cdot)$ plays essentially the same role in the prior densities as $n(\cdot)$ does in the likelihood function. Define the *imprecise Dirichlet model* to be the set of all prior densities of the above (product-Dirichlet) form, where s is a fixed positive number and $\mathbf{t}$ satisfies the preceding constraints. Note that the acronym IDM is used in this paper to refer to the model defined here and not to Walley's original definition [20].

By multiplying the prior density and the likelihood function, we obtain a posterior density of the same form as the prior, with $st(\cdot)$ replaced by $st(\cdot) + n(\cdot)$. Thus the posterior density for the theta-parameters is a product of independent Dirichlet densities.

2.4 Basic formulae

Let us focus on the computation of the expectation of $\theta_{c,\mathbf{a}}$ with respect to the posterior density for the theta parameters. This is denoted by $E[\theta_{c,\mathbf{a}}|\mathbf{n},\mathbf{t}]$. From (2) and the posterior independence of all the theta-parameters that appear there, each of which has a posterior beta distribution, it is easily shown that:

$$E[\theta_{c,\mathbf{a}}|\mathbf{n},\mathbf{t}] = P(c,\mathbf{a}|\mathbf{n},\mathbf{t}) = P(c|\mathbf{n},\mathbf{t}) \prod_{i=1}^{k} P(a_i|c,\mathbf{n},\mathbf{t}), \tag{6}$$

where $P(c|\mathbf{n},\mathbf{t}) = E[\theta_c|\mathbf{n},\mathbf{t}] = [n(c) + st(c)]/[N + s]$ and $P(a_i|c,\mathbf{n},\mathbf{t}) = E[\theta_{a_i|c}|\mathbf{n},\mathbf{t}] = [n(a_i\,|c) + st(a_i\,|c)]/[n(c) + st(c)]$. Here $P(\cdot|\mathbf{n},\mathbf{t})$ denotes epistemic posterior probability and $E(\cdot|\mathbf{n},\mathbf{t})$ denotes epistemic posterior expectation, with respect to the product-Dirichlet prior distribution with hyperparameters $\mathbf{t}$, after observing frequencies $\mathbf{n}$. These expressions are exact—no approximations are involved.

2.5 Checking for preference (credal dominance)

First note that if the attribute vector $\mathbf{a}$ is fixed, as it is for the inferences below, then the only constraints on $\mathbf{t}$ are: $0 < t(a_i\,|c) < t(c)$ for all values of (i,c), and $\sum_c t(c) = 1$. Now consider the problem of predicting the class of a new unit whose attribute values $\mathbf{a}$ are known. Let $E[U(c)|\mathbf{a},\mathbf{n},\mathbf{t}]$ denote the expected utility from choosing class c, given $\mathbf{a}$, the previous data $\mathbf{n}$ and a vector $\mathbf{t}$ of hyperparameters. I consider 0-1 valued utility functions, i.e., we receive utility 1 if we choose the correct class c and 0 if we do not, so that $E[U(c)|\mathbf{a},\mathbf{n},\mathbf{t}] = P(c|\mathbf{a},\mathbf{n},\mathbf{t})$.

We say that class c' *is preferred* to class c'' (or that c' *credal-dominates* c'') if and only if $E[U(c')|\mathbf{a},\mathbf{n},\mathbf{t}] > E[U(c'')|\mathbf{a},\mathbf{n},\mathbf{t}]$ for all values of $\mathbf{t}$ in the IDM, which holds if and only if $P(c'|\mathbf{a},\mathbf{n},\mathbf{t}) > P(c''|\mathbf{a},\mathbf{n},\mathbf{t})$ for all values of $\mathbf{t}$ in the IDM, which holds if and only if $P(c',\mathbf{a}|\mathbf{n},\mathbf{t}) > P(c'',\mathbf{a}|\mathbf{n},\mathbf{t})$ for all values of $\mathbf{t}$ in the IDM. The last step is based on the fact that $P(c,\mathbf{a}|\mathbf{n},\mathbf{t}) > 0$ and hence $P(\mathbf{a}|\mathbf{n},\mathbf{t}) > 0$, for all $\mathbf{t}$ in the IDM. To check whether c' is preferred to c'', it then suffices to solve the following optimization problem:

$$\inf \frac{P(c',\mathbf{a}|\mathbf{n},\mathbf{t})}{P(c'',\mathbf{a}|\mathbf{n},\mathbf{t})} \tag{7}$$

$$\sum_c t(c) = 1 \tag{8}$$

$$0 < t(a_i\,|c) < t(c) \quad \forall (i,c) \tag{9}$$

and to compare the result with 1 [21]. Recall that $P(c,\mathbf{a}|\mathbf{n},\mathbf{t})$ is given by (6). The criterion of credal dominance reported here is a special case of strict preference as defined by Walley ([19], Sect. 3.7.7).

2.5.1 Solution of the optimization problem

Problem (7) with constraints (8) and (9) can be rewritten as

$$\inf \left\{ \left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{k-1} \prod_i \frac{n(a_i\,|c') + st(a_i\,|c')}{n(a_i\,|c'') + st(a_i\,|c'')} \right\}$$
$$\sum_c t(c) = 1$$
$$0 < t(a_i\,|c) < t(c) \quad \forall (i,c),$$

after using (6) in the objective function (i.e., the function to optimize) and after some algebraic manipulation. It is possible to immediately observe that the infimum of the problem is obtained when each $t(a_i\,|c') \to 0$ and each $t(a_i\,|c'') \to t(c'')$, so these values are used into the objective function. Also, the constraint $\sum_c t(c) = 1$ can be replaced by $t(c') + t(c'') = 1$, since this is the only possibility at the infimum. Suppose it is not, i.e. $t(c') + t(c'') < 1$. Then we might hold $t(c'')$ fixed and increase $t(c')$ up to $1 - t(c'')$, so decreasing the infimum. By these considerations, the new form of the optimization problem becomes:

$$\inf \left\{ \left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{k-1} \prod_i \frac{n(a_i\,|c')}{n(a_i\,|c'') + st(c'')} \right\} \tag{10}$$

$$t(c') + t(c'') = 1 \tag{11}$$

$$t(c'), t(c'') > 0. \tag{12}$$

The cases when $n(a_i\,|c') = 0$ for some i can be treated separately: in such conditions, for any value of $t(c')$ and $t(c'')$ the function attains the minimum value zero (it is minimum because the function is non negative). We can thus assume for the subsequent derivation that $n(a_i\,|c') > 0$ for each i. Furthermore, let us also consider $k \geq 1$ in the rest of the paper, since the case $k = 0$ can be solved trivially.

The problem is rewritten by using the following notations, for short: $\alpha_i = n(a_i\,|c')$, $\beta_i = n(a_i\,|c'')$, $\alpha = n(c')$, $\beta = n(c'')$ and $x = st(c'')$. It becomes:

$$\inf h(x) = \inf \left\{ \left[\frac{\beta + x}{\alpha + s - x} \right]^{k-1} \prod_i \frac{\alpha_i}{\beta_i + x} \right\} \quad (13)$$

$$0 < x < s. \quad (14)$$

The objective function is always positive over the domain, so we can compute the logarithmic derivative of $h(x)$:

$$\frac{d \ln h(x)}{dx} = \frac{k-1}{\beta + x} + \frac{k-1}{\alpha + s - x} - \sum_i \frac{1}{\beta_i + x}.$$

Another differentiation leads to:

$$\frac{d^2 \ln f(x)}{dx^2} = -\frac{k-1}{(\beta + x)^2} + \frac{k-1}{(\alpha + s - x)^2} + \sum_i \frac{1}{(\beta_i + x)^2}.$$

This is always positive. In fact $\frac{1}{(\beta_i + x)^2} \geq \frac{1}{(\beta + x)^2}$, because $\beta_i \leq \beta$ and then $\sum_i \frac{1}{(\beta_i + x)^2} \geq \sum_i \frac{1}{(\beta + x)^2} = \frac{k}{(\beta + x)^2} > \frac{k-1}{(\beta + x)^2}$. It follows that $\ln h(\cdot)$ is convex and has a single infimum. By applying the exponential function we have that also $h(\cdot)$ is convex and admits a single point of infimum over the open interval $(0, s)$.

Let us now consider the behavior of $h(x)$ when $x \to 0$. $h(0)$ is not defined if there exists i such that $n(a_i\,|c'') = \beta_i = 0$. However, being $h(x)$ convex in the open interval implies that it tends to $+\infty$ when $x \to 0$ and so the minimum must be in $(0, s]$. We also know that $-\infty$ is the limit of the logarithmic derivative when $x \to 0$. This is used below, where, on the basis of the above considerations, I describe a simple algorithm for computing the infimum.

1. If there exists i such that $n(a_i\,|c') = 0$, let $\inf h(x) = 0$. Stop.
2. If there exists i such that $n(a_i\,|c'') = 0$, let $(\ln h(0))' = -\infty$, else compute $(\ln h(0))'$.
3. Compute $(\ln h(s))'$.
4. If $(\ln h(0))' \geq 0$, let $\inf h(x) = h(0)$. Stop.
5. If $(\ln h(s))' \leq 0$, let $\inf h(x) = h(s)$. Stop.
6. If $(\ln h(0))' < 0$ and $(\ln h(s))' > 0$, approximate the minimum numerically. Stop.

The first point is just the cited separate treatment of the cases where $n(a_i\,|c') = 0$ for some i. The rest of the procedure is a simple test which takes into account the values of the logarithmic derivative at the extremes of the interval. If the function was bounded, the points 4, 5 and 6 would obviously identify the infimum. The point 2 allows this test to be extended to the case of $h(0)$ being infinite, by introducing the value $-\infty$ for the derivative and by treating it homogeneously with the others.

As far as the numerical way to search for the minimum is concerned, one of the best choices seems Newton-Raphson's method because the first and second derivatives are available. The method is very fast, in very few iterations it can compute an approximation of the minimum at a very reasonable precision. The problem with the basic Newton-Raphson algorithm can be obtaining the convergence, but this is guaranteed when the algorithm is combined with bracketing ([17], p. 366). Note that the limitation of machine precision may prevent the test of credal dominance to be carried out; in fact, if the minimum of $\ln h(\cdot)$ is within machine precision from zero, it will not be possible to determine its actual sign. It seems reasonable to adopt a conservative approach defining that c'' is not credally dominated in this case (this follows naturally by treating the zero of the machine as the actual zero). Moreover, in all the other cases, by choosing a sufficient number of iterations, the overall test of credal dominance will be solved exactly: when the minimum of $\ln h(\cdot)$ is more distant than machine precision from zero, the bracket maintained by the algorithm will narrow until zero will be excluded, thus determining the sign of the minimum exactly even if the minimum will only be known approximately.

3 Extension to incomplete samples

I now turn to the problem of inferring the classifier from an incomplete multinomial sample. The underlying assumption behind the following development is that the incomplete sample is the result of two processes: a multinomial sampling and a subsequent unknown mechanism that turns some values into missing data [22]. I also assume that the class is never missing, as it is reasonable since classification is a supervised learning approach by definition.

In order to make the classifier be robust to all the possible missingness mechanisms, all the complete samples that are consistent with the incomplete one are taken into account. For each of them, the arguments in the preceding sections apply and the test of credal dominance can be expressed by problem (10) with constraints (11) and (12). Then it is taken the minimum of the set of infima generated when the missing data are replaced by known values in all the possible ways. This is obtained by a slight change in the test of credal dominance:

$$\inf \left\{ \left[\frac{n(c'') + st(c'')}{n(c') + st(c')} \right]^{k-1} \prod_i \frac{\underline{n}(a_i \,|c')}{\overline{n}(a_i \,|c'') + st(c'')} \right\}$$
$$t(c') + t(c'') = 1$$
$$t(c'), t(c'') > 0.$$

Here $\underline{n}(a_i \,|c')$ and $\overline{n}(a_i \,|c'')$ denote the minimum value of $n(a_i \,|c')$ and the maximum value of $n(a_i \,|c'')$, respectively, when all the possible replacements of missing data are considered. The form of the problem is the same of the original problem (10) with constraints (11) and (12), so that the solution procedure given in Sect. 2.5.1 applies.

The naive credal classifier can then cope with missing data without any increase in computational complexity. This holds when the NCC must be inferred from an incomplete sample, but it also holds when the NCC must classify an incomplete instance of the attributes: it is enough to do the test of credal dominance by considering $\mathbf{a}$ as the vector of the non-missing attributes (i.e. in this case k represents the number of observed attributes). This follows from an analogous characteristic of the NBC and by regarding the NCC as a set of NBCs.

4 Computational complexity

Inferring the NCC from an incomplete sample is a task linear in the number of units. In fact, it is matter of collecting the following set of two-way counts: $\underline{n}(a_i \,|c)$ and $\overline{n}(a_i \,|c)$ for all $i = 1, \ldots, k$, all $c \in \mathcal{C}$ and all $a_i \in \mathcal{A}_i$.

The classification of an instance requires to compare all the pairs of classes by the procedure to solve the test of credal dominance given in Sect. 2.5.1. Such procedure is based on the functions $(\ln h(\cdot))'$ and $(\ln h(\cdot))''$, whose values, for a given argument, can be computed in time $O(k)$ (see [4], p. 633, for the definition of $O(\cdot)$). This is also the complexity of the procedure if the Newton-Raphson's numerical approximation can be shown to work in constant time.

Let us analyze this point. Consider $\ln h(\cdot)$, where $h(\cdot)$ is given by (13). We can start the method in the middle of the interval, $x = s/2$. By using $\beta = N$, $\alpha = 0$ and $\alpha_i = N$, $\beta_i = 0$ for all i, we obtain the upper bound $\ln h(s/2) < (2k-1)\ln(2N/s+1)$. Following Walley's recommendation [20] to choose s in the interval $[1, 2]$, the upper bound is maximized at $s = 1$. The speed of the method of Newton-Raphson guarantees that even in the case of classification problems of very large size, $(2k-1)\ln(2N+1)$ can be reduced to a very good approximation of the minimum in few iterations. Consider, for instance, $k = 10^3$ and $N = 10^9$, which lead to $(2k-1)\ln(2N+1) \simeq 42811$. If we used a simple binary search that each time halves the interval containing the optimum, we would at most need $\log_2(42811 \cdot 10^9) \simeq 46$ iterations to have an error as low as 10^{-9}. This is an upper bound on the number of iteration of Newton-Raphson, which is in facts much faster: near the optimum the number of significant digits approximately *doubles* with each step ([17], p. 365). For all practical purposes we can therefore consider the number of iterations of Newton-Raphson's method a constant.

By considering all the tests of credal dominance between each pair of classes, it follows that the time required for the classification of a complete instance, which in the worst case, is $O(k\,|\mathcal{C}|^2)$.

5 Experiments

This section analyzes the behavior of the NCC on some real and artificial data sets. Experimental analyses are useful to emphasize that there is the need to define new methods to make experiments when credal classifiers are concerned. One such method is proposed in the following. Experiments can further highlight behaviors that are peculiar of credal classifiers and that can be unclear at first sight, so that they should be interpreted. I discuss some of these issues in Sect. 6.

5.1 Data sets

Four data sets are considered from the repository of the University of California at Irvine [15]: Breast, Corral, German and LetterAB.

- Breast is a breast cancer database obtained from the University of Wisconsin Hospitals, Madison [14]. It is composed of 699 records of patients made up by 9 continuous attributes related to the responses of the clinical analyses and a binary class: "benign cancer" (65%), "malignant cancer" (34.5%). 16 values are missing.

- Corral is an artificial data set [9] with 6 boolean attributes: A0, A1, B0, B1, IR and CR. The target concept is the boolean value: (A0 and A1) or (B0 and B1). IR is an irrelevant feature and CR is an attribute highly associated with the class, but with 25% error rate. There are 128 records where the class "false" appears in 56% of cases.

- German was donated by Professor Dr. Hans Hofmann from the Institut für Statistik und Ökonometrie, Universität Hamburg. The data set is related to the prediction of the type of German customer ("good" or "bad") as far as the

release of credit cards is concerned. The prediction is based on the customer's profile, defined by 13 categorical features and 7 continuous features. There are 1000 records, no missing values, and the classes appear with percentages of 70% ("good") and 30% ("bad").

- LetterAB is the restriction of the Letter database to the first two classes. Letter is a set of data for letter image recognition [5]. The objective is to identify each of a number of black-and-white rectangular pixel displays as letters in the English alphabet. LetterAB is composed of 16 continuous attributes and 1555 complete records. The two classes are almost equally represented.

I used the discretization utility of MLC++ [11], with default parameters, to convert the databases to sets of categorical data, since dealing with categorical attributes is an assumption of this paper. Note that discretizing the entire data sets once for all the subsequent experiments generally gives rise to a slight optimistic bias in the evaluation of the prediction accuracy. This is not problematic here since the focus of the analysis is not on evaluating the prediction accuracy on the original databases.

5.2 Bayesian models

It is important to compare the IDM inferences with Bayesian inferences, which means to compare the present version of the NCC with the NBC inferred according to some Bayesian prior. The following Bayesian models seem worthy of consideration.

- Haldane [7]: this uses a single product-Dirichlet density with $s = 0$ (which is the limit of the IDM as $s \to 0$). It gives the very simple formulae: $P(c|\mathbf{n}, \mathbf{t}) = n(c)/N$, $P(a_i|c, \mathbf{n}, \mathbf{t}) = n(a_i|c)/n(c)$. Note that the Haldane prior can give rise to undefined classification, due to null probabilities of the observed instance of the attributes.

- Perks [16]: we have $P(c|\mathbf{n}, \mathbf{t}) = [n(c) + 1/|\mathcal{C}|]/[N + 1]$ and $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1/|\mathcal{A}_i|]/[n(c)+1]$. By analogy with the IDM proposed in Sect. 2.3, I also consider a slightly different model in which $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1/(|\mathcal{C}|\,|\mathcal{A}_i|)]/[n(c) + 1/|\mathcal{C}|]$. In this way, the t-parameters that define the prior respect the IDM constraint (4); and the prior is also given less weight compared to the former definition.

- Uniform [12]: we obtain $P(c|\mathbf{n}, \mathbf{t}) = [n(c) + 1]/[N + |\mathcal{C}|]$ and $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1]/[n(c) + |\mathcal{A}_i|]$, or possibly $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1/|\mathcal{A}_i|]/[n(c) + 1]$ (again by analogy with the IDM).

- Jeffreys [8]: $P(c|\mathbf{n}, \mathbf{t}) = [n(c) + 1/2]/[N + |\mathcal{C}|/2]$, $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1/2]/[n(c) + |\mathcal{A}_i|/2]$, or possibly $P(a_i|c, \mathbf{n}, \mathbf{t}) = [n(a_i|c) + 1/(2\,|\mathcal{A}_i|)]/[n(c) + 1/2]$ (again by analogy with the IDM).

5.3 Results

The experiments aim to compare the NCC with seven versions of the NBC, based on the seven Bayesian priors described in Sect. 5.2. The comparison is principally based on the prediction accuracy, which is defined as the relative number of correct guesses.

In the following, I present the results of the evaluation of the classifiers on previously unseen sets of units. I used the scheme called r-fold cross-validation [10]. According to cross-validation, a data set $\mathcal{D}$ is split in 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 $\mathcal{D}_t$. The statistics for the quantities of interest, like the percentage of successful predictions (i.e. the accuracy), are collected over the r folds. Since the variance of the statistics can be large, especially for small samples, r-folds cross validation is repeated, by always making random splits of the database, until the standard deviation of the mean measures is lower than an arbitrary threshold which makes them be reasonably stable. I set the threshold to 0.33%; furthermore I used $r = 5$.

The results of the experiments are reported in the Tables 1–4. Each row in a table refers to a different prior distribution for the NBC. Notice that the priors modified by analogy with the IDM are marked with a prime (e.g., Perks′). The tables have the following columns.

- "Trials" is the number of repetitions of the 5-folds cross-validation or, also, the number of times that the entire data set was used to compute the statistics. This number depends on the size of the data set, on the measured percentages and on the chosen threshold for the standard deviation (i.e. 0.33%).

- "C_1%" is the accuracy of the NCC on the subset of instances where there is a single dominant class according to the NCC. That is, this column reports the results of the NCC when a reasonable confidence allows it to isolate a single class.

- "N%" is the accuracy of the NBC on the subset of instances whose probability is positive. This

value should be compared with C_1%.

- "Ns%" is the accuracy of the NBC on the subset of instances of positive probability, for which the NCC outputs more than one class. This is the most important measure. It is reasonable to expect that when the NCC suspends the judgment (i.e., it outputs two classes), this means that there is not enough knowledge in the data to do a reliable classification. So it is also reasonable to expect that the NBC accuracy in such cases is worse than that shown in the column N%.
- "S%" is the percentage of instances for which the NCC outputs more than two classes.
- "U%" is the percentage of instances with zero probability for the NBC. This value can be non zero only when the Haldane prior is used.

	Trials	C_1%	N%	Ns%	S%	U%
Haldane	7266	97.43	97.19	43.00	0.43	0.00
Perks	7354	97.43	97.19	42.55	0.43	0.00
Perks′	7372	97.43	97.19	43.01	0.43	0.00
Uniform	7234	97.43	97.18	40.09	0.43	0.00
Uniform′	7004	97.43	97.19	42.66	0.43	0.00
Jeffreys	7278	97.43	97.18	40.94	0.43	0.00
Jeffreys′	7372	97.43	97.19	43.01	0.43	0.00

Table 1: Experimental results on the Breast data set.

	Trials	C_1%	N%	Ns%	S%	U%
Haldane	4175	88.19	86.44	46.50	4.19	0.00
Perks	3955	88.19	86.54	48.98	4.20	0.00
Perks′	4188	88.19	86.52	48.28	4.19	0.00
Uniform	4188	88.19	86.66	51.76	4.19	0.00
Uniform′	4194	88.19	86.58	49.90	4.19	0.00
Jeffreys	3955	88.19	86.54	48.98	4.20	0.00
Jeffreys′	4188	88.19	86.52	48.28	4.19	0.00

Table 2: Experimental results on the Corral data set.

	Trials	C_1%	N%	Ns%	S%	U%
Haldane	497	76.10	75.17	55.41	4.48	0.00
Perks	498	76.10	75.14	54.74	4.48	0.00
Perks′	497	76.10	75.16	55.09	4.48	0.00
Uniform	500	76.10	75.09	53.63	4.48	0.00
Uniform′	498	76.10	75.14	54.65	4.48	0.00
Jeffreys	499	76.10	75.13	54.40	4.48	0.00
Jeffreys′	497	76.10	75.16	55.09	4.48	0.00

Table 3: Experimental results on the German data set.

	Trials	C_1%	N%	Ns%	S%	U%
Haldane	701	97.05	96.83	79.67	1.32	0.94
Perks	610	97.05	96.26	61.62	2.24	0.00
Perks′	604	97.06	96.28	62.59	2.25	0.00
Uniform	633	97.05	95.93	46.82	2.24	0.00
Uniform′	610	97.05	96.26	61.62	2.24	0.00
Jeffreys	645	97.05	96.05	52.33	2.24	0.00
Jeffreys′	604	97.06	96.28	62.59	2.25	0.00

Table 4: Experimental results on the LetterAB data set.

The discussion that follows is based on the column Ns% and on the obvious observation that when there are two classes, predicting at 50% is equivalent to randomly guessing. For all the priors, in the cases of the Corral and the German data sets, the prediction of the NBC where the NCC suspends the judgment is almost equivalent to a coin tossing (˜50%). As far as the Breast database, the column Ns% shows a negative departure from random guessing in all the cases. This seems to indicate that the inner bias of every precise-probability classifier that was considered had a bad effect on the prediction. In the case of the LetterAB data set, some priors exhibit a prediction slightly different from randomly guessing: e.g. Haldane's prior (in this case the accuracy is about 80%, but note that it is computed only on the 1.32% of instances with positive probability, where the NCC suspends the judgment, not on the 2.24% as in the other cases).

It appears that the NCC is able to isolate an area of ignorance for the databases Corral, German and Breast, where each precise-probability classifier cannot do reliable predictions. The results for LetterAB are more difficult to interpret: the most important evidence arising is that a precise-probability classifier can realize significant predictions in the set of instances where the NCC suspends the judgment. The discussion about this point is demanded to Sect. 6, which provides easy examples for such phenomenon and their justification.

There are two further points worth highlighting. First, for all the data sets, the accuracy in column C_1% is greater than that in column N%: the NCC always isolated a set of *hard* instances to classify. Consequently, the prediction on the rest of units (column C_1%) is always an improvement compared to the prediction on the entire data set (Ns%). Thus the NCC realized a robust prediction when it deemed that there was sufficient knowledge to isolate a single class.

Second, the mentioned area of ignorance can be quite large and therefore it cannot be neglected. For instance, it is made of about 45 units out of 1000 in the

German database. We should expect such a value to be larger for classifiers with weaker assumptions compared to (1), because, as it is well-known [2, 6], the variability of the model probabilities would be larger and so would be the prediction accuracy. This fact strongly supports the need of a proper treatment of imprecision, as the one the NCC realizes.

Finally, let us emphasize how the comparison of the credal classifier with its precise-probability counterparts was useful to the analysis. This seems to be important as far as the experimental methodology is concerned.

6 Interpreting NCC-vs-NBC behaviors

The results on the LetterAB data set show that the NBC can sometimes have a good accuracy on the subset of instances where the NCC suspends the judgment. At first glace this seems to contradict the ability of the NCC to isolate an area of ignorance. The following subsections describe this phenomenon by some examples and show that the behavior of the NBC, though seemingly successful, is instead unreasonable.

6.1 A class does not appear in the sample

We analyze the behavior of the NCC on a data set where a class never appears. Let us consider a binary class taking values in the set $\{c', c''\}$ and 20 binary attributes, with values in $\{a_i, a_i'\}$ for each i. We use the notation introduced for problem (13). In particular, we consider the values: $s = 1$, $k = 20$, $\alpha = n(c') = 10^6$, $\beta = n(c'') = 0$, $\alpha_i = n(a_i\,|c') = \alpha/2$ and $\beta_i = n(a_i\,|c'') = 0$ for each i. By (13), we have $h(1) = 10^6/2^{20} \simeq 0.95 < 1$, so that c'' is not credal-dominated by c' when the instance $\mathbf{a} = (a_1, a_2, \ldots, a_k)$ is observed; and the output of the NCC is thus $\{c', c''\}$.

This seems strange because $n(c') = 10^6$ is much larger than $n(c'') = 0$. Compare this with the behavior of the NBC inferred by using Haldane's prior (similar arguments can be used for the other priors), whose output is c'. Which classifier is right?

We can use the following argument to decide. We should prefer c' to c'' iff we were confident that $\theta_{c',\mathbf{a}} > \theta_{c'',\mathbf{a}}$ and intuitively this is true iff $n(\mathbf{a}\,|c')$ is substantially larger than $n(\mathbf{a}\,|c'')$. Under assumption (1), we are allowed to write $n(\mathbf{a}\,|c') \simeq n(c') \prod_{i=1}^{k} \theta_{a_i|c'} \simeq n(c') \prod_{i=1}^{k} \frac{n(a_i|c')}{n(c')} \simeq 1$ and similarly $n(\mathbf{a}\,|c'') \simeq 0$. We have that $n(\mathbf{a}\,|c')$ is *not* substantially larger than $n(\mathbf{a}\,|c'')$, so we should not discard the case $C = c''$. The NCC appears to be right and the NBC to be wrong. (Note that this is also an example of how the NCC can deal with rare events, which are another critical problem in the field of classification.)

Counter-intuitive facts like this should be taken into account especially when experimental evaluations were concerned, because they generally provide us only with a partial view. For example, if we used cross-validation on a data set like the one above, the NBC would have a very good prediction accuracy in the subset of instances where the NCC suspends the judgment, even up to 100%. This would not contradict the ability of the NCC to isolate an area of ignorance; and also, the NBC should *not* be the classifier to choose. Its behavior on the finite sample does not provide us with any reasonable confidence of being extendable to a larger sample.

6.2 If the independence assumption fails

We consider another data set where a class does not appear. In this case the behavior of the classifiers is similar to the one reported in the preceding section, but the explanation differs.

Consider Tab. 5. It represents a sample where each unit is an instance of three binary attributes and a binary class. Notice that c'' is never observed.

A_1	A_2	A_3	C
a_1	a_2	a_3	c'
a_1	a_2	a_3	c'
a_1	a_2	a_3	c'
a_1	a_2	a_3	c'
a_1'	a_2'	a_3'	c'
a_1'	a_2'	a_3'	c'
a_1'	a_2'	a_3'	c'
a_1'	a_2'	a_3'	c'

Table 5: A set of data that does not support the assumption of independence.

Consider $s = 1$. Given observation $\mathbf{a} = (a_1, a_2, a_3)$, it is easy to verify that the output of the NCC is $\{c', c''\}$, whereas the NBC inferred by using Haldane's prior outputs c'. Note that in Tab. 5 we have $n(\mathbf{a}\,|c') = 4$ and $n(\mathbf{a}\,|c'') = 0$, so, in contrast with the discussion in Sect. 6.1, in this case we should discard the case $C = c''$. Thus the NBC, proposing the right solution, seems to be right and the NCC to be wrong.

The situation is reversed if we interpret the question more carefully. As in Sect. 6.1, we can write $n(\mathbf{a}\,|c') \simeq n(c') \prod_{i=1}^{k} \frac{n(a_i|c')}{n(c')} = 1$ and similarly $n(\mathbf{a}\,|c'') \simeq 0$. These values do not allow us to discard the class c''. Hence the NCC is methodologically right. It remains to understand what is the actual source of

the problem. The problem lies in the poor evaluation of $n(\mathbf{a}\,|c')$, caused by a serious violation of assumption (1) in the data. Both classifiers do not realize that $n(\mathbf{a}\,|c')$ is sufficiently larger than $n(\mathbf{a}\,|c'')$, as it appears from the data set. Their reaction to this fact is then different, because of their different attitudes towards risk. The optimistic bias of the NBC makes it choose the class c', but this is not justified by the evidence that is available to the classifier.

The discussion highlights a useful question as far as experiments are concerned: when the NBC has a good accuracy on the instances on which the NCC suspends the judgment, this may suggest that assumption (1) seriously disagrees with the evidence coming from data and that more structured classifiers may be needed.

6.3 Another failure of the independence assumption

We can obtain a behavior of the NCC as that described in the preceding section also when all the classes appear in the sample, as in Tab. 6. In the sample there are four binary attributes and a binary class. We consider $s = 2$ for the NCC and Haldane's prior for the NBC.

A_1	A_2	A_3	A_4	C
a_1	a'_2	a_3	a'_4	c'
a_1	a'_2	a_3	a'_4	c'
a_1	a'_2	a_3	a'_4	c'
a'_1	a'_2	a'_3	a'_4	c'
a'_1	a'_2	a'_3	a'_4	c'
a_1	a_2	a_3	a'_4	c''
a_1	a_2	a_3	a'_4	c''
a_1	a_2	a_3	a'_4	c''
a_1	a_2	a_3	a'_4	c''
a_1	a_2	a_3	a'_4	c''

Table 6: Another failure of assumption (1).

Given observation (a_1, a'_2, a_3, a'_4), we have that $P(c'', a_1, a'_2, a_3, a'_4) = 0$ for the NBC, because $P(a'_2|c'') = 0$, and hence the NBC prefers c' to c''. In contrast, c'' is not credally dominated for the NCC because $h(7/4) = 225/253.2 < 1$ (see Sect. 2.5.1).

It is possible to follow the same arguments given in Sect. 6.2 to interpret this behavior. The example enforces the evidence related to the problems that can arise for a wrong assumption of independence. This might be the case of the LetterAB data set (see Sect. 5.3), where the configurations of pixels for a letter are likely to seriously violate (1).

6.4 A note on missing data

The above discussion on the violation of the independence assumption puts the treatment of missing data proposed in Sect. 3 under a new light. Consider the following incomplete sample for two independent binary variables (A_1, A_2): $[(a_1, *), (*, a_2), (a_1, *), (*, a_2), (a_1, *), (*, a_2)]$. By considering all the possible replacements of the missing data with known values, we also take into account complete samples that clearly violate the assumption of independence, as in this case: $[(a_1, a'_2), (a'_1, a_2), (a_1, a'_2), (a'_1, a_2), (a_1, a'_2), (a'_1, a_2)]$. However, the independence assumption (1) is commonly used as a convenient approximation for the purposes of classification. In this spirit, it is reasonable to consider the latter sample, too.

7 Conclusions

The classification literature seems to neglect the topic of properly treating imprecision. This is unreasonable as also for simple classifiers like those presented here, imprecision renders the classification indeterminate in a number of cases; and because treating imprecision soundly and efficiently is possible, as shown here.

The bottom-line advantage of using credal classifiers is the robustness of the classification: we know that, given the chosen model, unreasonable predictions are automatically discarded. The need of doing post-classification analysis is greatly reduced. This particularly suits the current needs of the field—due to the several possible applications and great availability of databases—such as obtaining quick and robust responses.

But robustness is a core characteristic that extends far beyond this point: e.g., to missing data. These constitute a pervasive problem in the practice of classification and are recognized as a critical theoretical topic. By the proposed approach, it is possible to build a classifier that is robust to every missingness mechanism. Again, this does not require expensive computations and provides us with a tool based on a clear approach that we can easily trust.

More broadly speaking, relaxing the assumption of precision seems to be a way to cope with the most critical problems. Apart from the discussed cases of small and incomplete samples, also massive databases might profit from credal classification. These are huge samples for which it is impractical to scan the entire database, so that the inference phase must be based on a subset of the observations. Choosing when to stop reading the data and obtaining robust classifications

are then two important problems. By using a credal classifier these might be addressed in a natural way: ideally the inference phase might be stopped when the classifications were precise—i.e. when there was enough evidence to isolate a single class; whenever it had to be stopped earlier, the obtained classification would be robust by definition.

There is evidence suggesting that credal classifiers can have a deep impact on classification. More research efforts are needed to develop new credal classifiers and to improve upon the existing tools.

Acknowledgements

This work owes a great debt to Peter Walley. He proposed the present definition of the IDM and also gave substantial contributions to Sect. 6. Very unfortunately, he could not collaborate until the end of this project. He chose to resign as a co-author of the paper for this reason.

References

[1] P. Domingos. Machine learning. In W. Klosgen and J. Zytkow, editors, *Handbook of data mining and knowledge discovery*. Oxford University Press, New York. To appear.

[2] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, 1997.

[3] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001. 2nd edition.

[5] P. W. Frey and Slate D. J. Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, 6(2):161–182, 1991.

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

[7] J. B. S. Haldane. On a method of estimating frequencies. *Biometrika*, 33:222–225, 1945.

[8] H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1983. 3rd edition.

[9] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In W. W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference on Machine Learning*, pages 121–129, New York, 1994. Morgan Kaufmann.

[10] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI-95*, pages 1137–1143, San Mateo, 1995. Morgan Kaufmann.

[11] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.

[12] de P. S. Laplace. *Théorie Analytique des Probabilités*. Courcier, Paris, 1812.

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

[14] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 23(5):1–18, 1990.

[15] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1995. http://www.sgi.com/Technology/mlc/db/.

[16] W. Perks. Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar.*, 73:285–312, 1947.

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1993. 2nd edition.

[18] M. Ramoni and P. Sebastiani. Robust Bayes classifiers. *Artificial Intelligence*, 125(1–2):209–226, 2001.

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

[20] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[21] M. Zaffalon. A credal approach to naive classification. In G. de Cooman, F. Cozman, S. Moral, and Walley P., editors, *ISIPTA'99*, pages 405–414, Univ. of Gent, Belgium, 1999. The Imprecise Probabilities Project.

[22] M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 2001. To appear.

[23] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 2001. To appear.