Credal Model Averaging: an Extension of Bayesian Model Averaging to Imprecise Probabilities

Giorgio Corani and Marco Zaffalon

IDSIA
(Istituto Dalle Molle di Studi sull'Intelligenza Artificiale)
Manno, Switzerland
{giorgio,zaffalon}@idsia.ch

Abstract. We deal with the arbitrariness in the choice of the prior over the models in *Bayesian model averaging* (BMA), by modelling prior knowledge by a set of priors (i.e., a prior *credal set*). We consider Dash and Cooper's BMA applied to *naive Bayesian networks*, replacing the single prior over the naive models by a credal set; this models a condition close to prior ignorance about the models, which leads to *credal model averaging* (CMA). CMA returns an *indeterminate* classification, i.e., multiple classes, on the instances for which the learning set is not informative enough to smooth the effect of the choice of the prior. We give an algorithm to compute exact credal model averaging for naive networks. Extensive experiments show that indeterminate classifications preserve the reliability of CMA on the instances which are classified in a prior-dependent way by BMA.

Keywords: Credal model averaging, Bayesian model averaging, imprecise probabilities, naive Bayes, classification, naive Bayesian networks.

1 Introduction

In the last ten years, data mining and statistical research has been paying increasing attention to the question of model uncertainty. Loosely speaking, *model uncertainty* refers to a situation where more than one model is consistent with the available data. Many researchers have argued, both theoretically and empirically, that taking such an uncertainty into account leads to improved inference (see [1] for a recent overview). In this context, *Bayesian model averaging* (BMA) [2] has proven to be an effective way to deal with model uncertainty.

BMA is based on a very simple observation: that the posterior probability for an event of interest given the data, say $P(X = x|\mathbf{d})$, can be re-written (in the case of finitely many models) as

$$P(X = x|\mathbf{d}) = \sum_{j=1}^{l} P(X = x|M_j, \mathbf{d})P(M_j|\mathbf{d}), \tag{1}$$

thus making explicit its dependency on the possible model M_j; in particular, $P(M_j|\mathbf{d}) = P(\mathbf{d}|M_j)P(M_j)/P(\mathbf{d})$ formalizes how much one should trust each model after having observed the data if the prior beliefs were $P(M_j)$.[1]

Despite BMA being a sound approach to deal with model uncertainty, it gives rise to challenges, such as the fact that the exhaustive sum in Eq. (1) can become intractable because of the number of models, thus requiring to adopt approximate solutions which are in general computationally expensive.

Another important issue concerns the arbitrariness inherent in the choice of the prior over the models; in fact, the results produced by BMA can be sensitive to such a choice. Traditionally, a very common choice is to adopt a uniform prior over the models; this, however, can be criticized from different standpoints (see for instance the discussions in the rejoinder of [2]). Alternatively, in [3] a prior is adopted which favors simple models over complex ones. Although all these choices are reasonable in some situation, it is more difficult to justify their use in general. The problem is that the specification of any single prior implies some arbitrariness, which entails the risk of drawing prior-dependent conclusions that may be fragile. In fact, the way the prior over the models should be specified is a serious open problem of BMA.

In this paper we focus in particular on pattern *classification*, where BMA is often related to *feature selection*. In fact, given a set of N feature variables, one can design 2^N different subsets of feature variables; *feature selection* is indeed concerned with selecting the supposedly best subset of the feature variables, which corresponds to a supposedly best classifier. An appealing alternative to the selection of a single classifier is to use BMA to average over all the 2^N classifiers.

This avenue has been taken by Dash and Cooper, who focused in particular on *Bayesian networks* [4]. In case of *naive* networks, in particular, their approach allows one to compute BMA *exactly* and *efficiently*, as their algorithm does not introduce any approximation and has complexity $\mathcal{O}(N)$. Moreover, Dash and Cooper show that exact BMA over the 2^N naive networks can be implemented by a single summary naive network. Yet, they do not discuss the problem of the sensitivity to the prior, nor does so a subsequent approach that implements another form of averaging for naive nets [3].

Our standpoint is that solving the problem of the prior in BMA may require to drop the idea of specifying a unique, precise prior, and to model instead prior knowledge by a set of priors; such a set is referred to as the *credal set*. In Section 2.3 we extend Dash and Cooper's BMA to *imprecise probability* [5], substituting the single prior by a credal set. We call the resulting approach *credal model averaging* (CMA). While traditional non-informative priors model a condition of indifference between the different models, the prior that we define for CMA models a condition close to *prior ignorance* by expressing very weak beliefs a priori about the relative credibility of the 2^N naive nets. Then we use Dash and Cooper's algorithm to efficiently turn each precise prior in the credal set

[1] The more traditional approach to inference that considers only one model $M_{\bar{j}}$ as possible is indeed recovered when $P(M_{\bar{j}}) = 1$.

into a posterior. The set of posteriors obtained in this way is referred to as the *posterior credal set*.

Having multiple posteriors instead of one leads to a generalized form of classification that we have called *credal classification* in some previous work [6,7]. In particular, CMA returns a *determinate* classification, i.e, a single class, only if the probability of such a class is larger than that of any other *for all* the precise posteriors in the posterior credal set. Otherwise, if different classes are found to be the most probable, depending on the specific posterior considered from the posterior credal set, CMA returns an *indeterminate* classification, i.e., multiple classes. We call 'hard to classify' the instances in the test set that give rise to indeterminate classifications, meaning that the learning set is not informative enough about them (in order to smooth the effect of *all* the priors in the credal set in favor of a single class). We expect Dash and Cooper's BMA to behave unreliably on the instances recognized as hard by CMA, as their classification is prior-dependent indeed.

In Section 3 we investigate this point empirically using 31 data sets from the UCI repository. We split the test instances according to whether they are deemed to be hard or not by CMA. Then we evaluate the predictive performance of BMA separately on the two types of instances. What we observe is indeed a striking drop in the predictive accuracy of BMA moving from the instances that are not hard to the others. The drop is observed on every data set we consider, with no exception. Moreover, we show indeterminate classification to be valuable, as they are informative (they return on average only a minority of the classes, not all of them) and reliable (they do contain the actual class with very high frequency). Summing up, extensive experiments show that CMA is a more robust approach than BMA.

Moreover, CMA implements an idea of model averaging that overcomes the arbitrariness in the choice of the prior in a novel way, which could be used more generally than what we do here. In fact, CMA leads very naturally to classification robustness. This is achieved, in particular, by relying on the paradigm of credal classification, which has already proven to be suitable for data mining purposes: in a recent work [7], we have extended naive Bayes to imprecise probabilities, in order to deal robustly with the specification of the prior density over the parameters of the model and with the treatment of missing data, achieving a remarkable reliability improvement compared to naive Bayes. Hence, in our view, allowing classifiers to give weaker answers than the determinate ones we are used to in classification may enhance the overall classification reliability.

2 Credal model averaging

In the following section we show how we extend the BMA framework of Dash and Cooper [4] to manage a set of priors over the models. Our setting is in fact characterized by the same assumptions of Dash and Cooper and by a similar notation.

2.1 Setup

We consider a supervised classification problem; there is a vector of N feature variables $\mathbf{F} := (F_1, F_2, \ldots, F_N)$ and a set of N_c classes $\mathcal{C} := \{c_1, c_2, \ldots, c_{N_c}\}$. The i-th instance of the data set $\mathbf{d}$ is the pair $(\mathbf{f}_i, c_i)$, where $\mathbf{f}_i := (f_{1i}, f_{2i}, \ldots, f_{Ni})$ is the instance of the feature variables in the instance under consideration. The data set contains n instances, generated by an *independently and identically distributed* mechanism.

We consider a Bayesian network with $N+1$ nodes, i.e., a single class node and N feature nodes; we assume the network to be *naive*, i.e., a feature node is either linked to the class or it is isolated. We denote by $\mathbf{X}$ the collection of nodes of the network; they are indexed by i so that $X_0 := C$, while, for $i \neq 0$, $X_i := F_i$. Moreover, the class node has no parent. A certain layout of the network, in which certain feature nodes are linked to the network and the remaining ones are isolated, is referred to as a *graph.* Given the N feature variables, we can hence design 2^N different graphs.

All feature variables are assumed to be *categorical*; i.e., each node X_i represents a categorical random variable with r_i possible states. In practice, this requires to discretize the numerical features before inducing the classifier.

We denote by θ_{ijk} the physical probability (or *chance*), about which we are uncertain, of X_i to be in state k when the parent node is in state j. The vector $\boldsymbol{\theta}_{ij}$ (made of r_i elements) contains hence the chances of the states of node i conditional on the j-th state of the parent; finally, $\boldsymbol{\theta}$ collects all the vectors $\boldsymbol{\theta}_{ij}$, i.e., it contains all the parameters of the network.

We take a Dirichlet density $Dir(\alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijr_i})$ as prior over each vector $\boldsymbol{\theta}_{ij}$, with $\alpha(\cdot) > 0$. We adopt the following setting: for the i-th feature node, we set $\alpha_{ijk} = 1/(N_c \cdot r_i)$. For the class node, we set[2] $\alpha = 1/N_c$.

As usual with Bayesian networks, we assume moreover *parameter independence* and moreover we assume the data set to be *complete*, i.e., without missing data.

2.2 Overview of Dash and Cooper's BMA

In this section we briefly recall Dash and Cooper's approach to BMA. Let us denote by $\mathcal{G}$ the set of the 2^N graphs which can be designed given the N feature variables, and by g a generic graph in $\mathcal{G}$. BMA computes a weighted average of the probabilities produced by all the graphs as follows:

$$P(\mathbf{X}=\mathbf{x}|\mathbf{d}) = \sum_{g \in \mathcal{G}} P(\mathbf{X}=\mathbf{x}|g,\mathbf{d}) P(g|\mathbf{d}) \propto \sum_{g \in \mathcal{G}} P(\mathbf{X}=\mathbf{x}|g,\mathbf{d}) P(\mathbf{d}|g) P(g), \quad (2)$$

where $P(\mathbf{X} = \mathbf{x}|g, \mathbf{d})$ is the posterior probability of the instance to classify assuming that the underlying graph is g (in which some feature variables are

[2] To be more precise, the parameter referring to the class node should be denoted as α_{00k}.

linked to the class and some others are isolated), $P(\mathbf{d}|g)$ represents the (so-called *marginal*) *likelihood* of graph g and $P(g)$ represents the *prior probability* of graph g. The last relation in Eq. (2) is due to Bayes' rule.

Let us give the explicit form for the first term in the sum:

$$P(\mathbf{X}=\mathbf{x}|g,\mathbf{d}) = \prod_{i=0}^{N} \hat{\theta}_{iJK} := \prod_{i=0}^{N} \frac{\alpha_{iJK}+n_{iJK}}{\alpha_{iJ}+n_{iJ}}. \tag{3}$$

Here the coefficients n_{ijk} are counts collected from the data set that report how many times feature variable i is in state k when its parent is in state j; coefficients α_{ijk} refer instead to the Dirichlet densities introduced in Section 2.1. Moreover, $n_{ij} := \sum_k n_{ijk}$ and $\alpha_{ij} := \sum_k \alpha_{ijk}$. The uppercase letters J and K denote the specific states of nodes and parents which have been read off from vector $\mathbf{x}$.

In practice, the coefficients n_{ijk} are computed differently depending on whether they refer to the class, to a feature node linked to the class or to an isolated feature node. In particular:

- for the class node, n_{0jk} has the meaning of class frequency, i.e., it indicates how many times class k occurs in the training set; at the denominator, n_{0j} corresponds to the data set size n. Note that the value of $\hat{\theta}_{0jk}$ is the same for all graphs;
- for feature nodes linked to the class, n_{ijk} represents a conditional frequency, i.e., it indicates how many times in the training set feature variable i assumes value k while the class has value j; at the denominator, n_{ij} represents the total occurrences of class j in the training set. Given a feature X_i $(i \neq 0)$, $\hat{\theta}_{ijk}$ has the same value for all graphs in which X_i is linked to the class node; let us denote this quantity by $\hat{\theta}^{C}_{ijk}$;
- for isolated feature nodes, n_{ijk} represents an unconditional frequency, i.e., it indicates how many times feature variable i assumes value k in the training set; at the denominator, n_{ij} corresponds to the data set size n. Given a feature X_i $(i \neq 0)$, $\hat{\theta}_{ijk}$ has the same value for all graphs in which node X_i is isolated; let us denote this quantity by $\hat{\theta}^{\emptyset}_{ijk}$.

The coefficients $\alpha(\cdot)$ can be interpreted in the same way of coefficients $n(\cdot)$, provided that they are regarded as referring to the so-called *hypothetical sample* rather than to the actual data set.

Let us now consider the marginal likelihood. We have:

$$P(\mathbf{d}|g) = \prod_{i=0}^{N} M_i := \prod_{i=0}^{N} \left(\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})} \right), \tag{4}$$

where the coefficients $\alpha(\cdot)$ and $n(\cdot)$ have the meaning already discussed. Hence, M_0 is a fixed value for all the graphs, while M_i $(i \neq 0)$ is a fixed value M_i^C for all the graphs in which X_i is linked to the class, and another fixed value $M_i^{\emptyset}$ for all the graphs in which X_i is isolated. Let i^C and as $i^{\emptyset}$ denote the set of indexes to feature variables which in graph g are respectively linked to the class node and isolated. We can eventually express Eqs. (3) and (4) in a more compact way as

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}|g,\mathbf{d}) &= \hat{\theta}_0 \prod_{i \in i^C} \hat{\theta}^C_{iJK} \prod_{i \in i^\emptyset} \hat{\theta}^\emptyset_{iJK}, \qquad (5) \\ P(\mathbf{d}|g) &= M_0 \prod_{i \in i^C} M^C_i \prod_{i \in i^\emptyset} M^\emptyset_i . \qquad (6) \\ & \qquad (7) \end{aligned}$$

Concerning the prior over the graphs, corresponding to the term $p(g)$ in Eq. (2), Dash and Cooper require it to be a *modular* prior, which means it should also factorize into a product of $N+1$ terms, each one corresponding to a node. Then they do not detail the prior any further, much probably because they use a flat prior that cancels out of the calculations. Since this will not be our case, we give here a few more details about the prior. Call p_i the probability that node i is connected to a parent.We design a modular prior by simply requiring that

$$P(g) = \prod_{i \in i^C} p_i \prod_{i \in i^\emptyset} (1-p_i), \tag{8}$$

and in addition that $p_0 = 0$, because we know that the class variable has always no parents. Note that to recover the flat prior over the graphs, it would be sufficient to set $p_i := 0.5$ for all $i = 1, \ldots, N$.

We are finally in the condition to write an explicit formula for $P(\mathbf{X}=\mathbf{x}|\mathbf{d})$. Let us introduce the following quantities (which are all positive):

$$\begin{aligned} \rho_{0K} &:= \theta_{0JK} M_0, \\ \rho^C_{iJK} &:= \theta^C_{iJK} M^C_i, \\ \rho^\emptyset_{iK} &:= \theta^\emptyset_{iJK} M^\emptyset_i, \end{aligned} \tag{9}$$

where we have dropped index J in the definition of ρ_{0K} and $\rho^\emptyset_{iK}$; in fact, these quantities refer to the class node and to the isolated feature nodes, which have no parents. It turns out then that

$$P(\mathbf{X}=\mathbf{x}|\mathbf{d}) \propto \rho_{0K} \cdot \prod_{i=1}^{N} \left[(1-p_i)\rho^\emptyset_{iK} + p_i \rho^C_{iJK} \right]. \tag{10}$$

This way of expressing $P(\mathbf{X}=\mathbf{x}|\mathbf{d})$ is an achievement from Dash and Cooper that is particularly important for computations: in fact, it means that once the $\rho(\cdot)$ coefficients have been computed, Eq. (10) is computed in $\mathcal{O}(N)$ time, without the need for implementing the 2^N models and without introducing any approximation. Dash and Cooper also show that computing BMA according to Eq. (2) is equivalent to implementing a single summary network characterized by a new vector $\hat{\boldsymbol{\theta}}^*$; assuming to adopt a uniform prior over the graphs, it holds that for $i=0$, $\hat{\theta}^*_{0JK} \propto \rho_{0K}$ and, for $i \neq 0$, $\hat{\theta}^*_{iJK} \propto (\rho^C_{iJK} + \rho^\emptyset_{iK})$.

2.3 Extension of BMA to imprecise probabilities

We extend BMA to imprecise probabilities by considering a set $\mathcal{P}$ of priors over the graphs, instead of a single prior; $\mathcal{P}$ is referred to as *prior credal set*. Before detailing the construction of the prior credal set, let us consider the motivations behind such a choice and some of its consequences.

A major motivation behind using a credal set rather than a single prior is related to modeling prior ignorance. The point is that by a single prior it is possible to model indifference; in order to model ignorance, one should use a credal set.[3] Therefore credal sets allow us to express more satisfactorily the fact that initially we do not know about the relative credibility of the models; this naturally makes the resulting classifier more robust than BMA. Indeed, especially when the learning set is small, the class returned by BMA may well vary depending on the specification of the prior over the graphs; in this case, the classification is defined as *prior-dependent* and its reliability is questionable. On the other hand, since CMA considers a set of priors as possible, it is aware by construction that some classifications may change with the choice of the prior in the credal set, and this enables it to keep reliability. The way this is done in practice is related to the definition of the optimality criterion for the classes in the imprecise setting.

Let us recall that a Bayesian classifier returns as *optimal prediction* the class with the highest probability (in the case of 0-1 loss function), identified on the basis of a uniquely computed posterior, derived from a unique prior. In the imprecise probability setting, one specifies a set of priors that is turned into a set of posteriors by element-wise application of Bayes' rule. According to Section 3.9.2 of [5], the optimality criterion in this case is to return all the *non-dominated* classes. The definition of dominance is as follows: class c_1 dominates c_2 if for all

```
IDENTIFICATION OF NON-DOMINATED CLASSES

1. set NonDominatedClasses := $\mathcal{C}$;
2. for class $c' \in \mathcal{C}$
     – for class $c'' \in \mathcal{C}$, $c'' \neq c'$
         • if $c''$ is dominated by $c'$ (to be assessed via the below procedure), drop $c''$
           from NonDominatedClasses;
         • exit;
     – exit
3. return NonDominatedClasses.
```

Fig. 1. Identification of non-dominated classes via pairwise comparisons.

[3] Yet, complete prior ignorance is not compatible with learning, see Section 7.3.7 of [5]. This issue is re-considered later in this section when we define the credal set.

the computed posteriors, the probability of c_1 is greater than that of c_2; hence, c_2 is non-dominated if no class dominates c_2.

A key point is that there can be several non-dominated classes; in this case, the classifier returns an indeterminate (or set-valued) classification. Classifiers that issue set-valued classifications are called *credal classifiers* in [6]. Summing up, a credal classifier will become indeterminate on the instances whose classification would be prior-dependent when a single prior is used; on these instances, it will return all the non-dominated classes as a way to maintain reliability. It is important to realize that non-dominated classes are *incomparable*; this means that there is no information in the model that allows us to rank them. In other words, credal classifiers are models that allow us to drop the dominated classes, as sub-optimal, and to express our indecision about the optimal class by yielding the remaining set of non-dominated classes.

Let us focus now in particular on the test of dominance; let $\mathbf{x}_1 := (\mathbf{f}, c_1)$ and $\mathbf{x}_2 := (\mathbf{f}, c_2)$. We say that class c_2 is dominated by c_1 if and only if

$$P(\mathbf{x}_1|\mathbf{d}) > P(\mathbf{x}_2|\mathbf{d}) \ \forall P \in \mathcal{P},$$

or, equivalently,[4] if and only if

$$\frac{P(\mathbf{x}_1|\mathbf{d})}{P(\mathbf{x}_2|\mathbf{d})} = \frac{P(\mathbf{x}_1, \mathbf{d})}{P(\mathbf{x}_2, \mathbf{d})} > 1 \ \forall P \in \mathcal{P},$$

which, taking Eq. (2) into consideration, can be finally re-written as

$$\inf_{P \in \mathcal{P}} \frac{\sum_{g \in \mathcal{G}} P(\mathbf{x}_1|g, \mathbf{d}) P(\mathbf{d}|g) P(g)}{\sum_{g \in \mathcal{G}} P(\mathbf{x}_2|g, \mathbf{d}) P(\mathbf{d}|g) P(g)} > 1. \tag{11}$$

A procedure to determine all the non-dominated classes via pairwise comparisons is shown in Figure 1.

Prior credal set We are finally ready to define the prior credal set. Let us focus on p_i, that is, the probability that feature variable i is connected to the class. Remember that we want to model a condition of prior ignorance about the actual graph, among the 2^N possible ones, giving rise to the data. Since we are ignorant a priori, this means that for each feature variable, we ignore whether it is linked or not to the class. In turn, this means that our probability p_i for the related arc should lie in $[0, 1]$. We can therefore construct the credal set $\mathcal{P}$ by considering the set of all the mass functions defined as in (8) that are obtained when each p_i, $i = 1, \ldots, N$, is subject to the constraint $0 < p_i < 1$ (and, as before, $p_0 = 0$). However, it can be checked that this choice does not allow us to learn from data about the relative credibility of the models. Broadly speaking, this is a relatively well-known phenomenon (e.g., something similar was noticed in [8, Section 3] in the case of feature selection); the intuition here is that the modeled condition is of such deep ignorance a priori that no amount of data

[4] If the denominator is positive, which is always the case in this paper.

would be able to make us get out of such a state. For this reason, we need to consider a slightly smaller credal set as defined by the following constraints:

$$\begin{aligned} p_i = 0 &\text{ if } i = 0, \\ \epsilon \le p_i \le 1 - \epsilon &\text{ if } i \neq 0, \end{aligned} \tag{12}$$

where ϵ is a small number in $(0, 0.5)$, which we will set to 10^{-5} in our experiments.[5] By this simple consideration, we model a condition that is still close to ignorance but that at the same time enables us to learn.

Two final remarks are worth making. One is that when CMA is determinate, it returns the same class as BMA. This is the consequence of two facts: that (a) CMA returns a determinate output when a certain class dominates all the remaining ones, under all the priors of the credal set; and that (b), the credal set includes the flat prior adopted by BMA (remember that it is actually characterized by $p_i = 0.5$ for all the feature variables). Therefore, BMA and CMA achieve the same accuracy on the subset of instances determinately classified by CMA, and whose classification is prior-independent.

The second remark is that CMA will converge to BMA with increasing sizes of the learning set. This follows because all the precise priors in the prior credal set will converge towards a single posterior with more and more data. Therefore in the limit, CMA will yield a traditional classifier that always issues determinate classifications.

2.4 CMA computation

Recalling Eqs. (10) and (11), and letting $\mathbf{p} := (p_1, \ldots, p_N)$, the CMA test of dominance for classes c_1 and c_2 can be written as follows:

$$\min_{P \in \mathcal{P}} \frac{\sum_{g \in \mathcal{G}} P(\mathbf{x_1}|g, \mathbf{d}) P(\mathbf{d}|g) P(g)}{\sum_{g \in \mathcal{G}} P(\mathbf{x}_2|g, \mathbf{d}) P(\mathbf{d}|g) P(g)} = \min_{\mathbf{p}} \frac{\rho_{01}}{\rho_{02}} \prod_{i=1}^{N} \frac{\left[(1-p_i)\, \rho_{iK}^{\emptyset} + p_i \rho_{i1K}^{C}\right]}{\left[(1-p_i)\, \rho_{iK}^{\emptyset} + p_i \rho_{i2K}^{C}\right]}. \tag{13}$$

Note that the function in (13) can be globally minimized by minimizing independently each term of the product, i.e., by minimizing independently over each p_i. The minimization problem to be solved for a single p_i is

$$\min_{p_i} \frac{\left[(1-p_i)\, \rho_{iK}^{\emptyset} + p_i \rho_{i1K}^{C}\right]}{\left[(1-p_i)\, \rho_{iK}^{\emptyset} + p_i \rho_{i2K}^{C}\right]} = \min_{p_i} \frac{p_i m_1 + a}{p_i m_2 + a} = \min_{p_i} f(p_i), \tag{14}$$

where

$$a := \rho_{iK}^{\emptyset},\ m_1 := \rho_{i1K}^{C} - \rho_{iK}^{\emptyset},\ m_2 := \rho_{i2K}^{C} - \rho_{iK}^{\emptyset} \tag{15}$$

and subject to the constraint $\epsilon \le p_i \le 1 - \epsilon$. This is an easy problem because the derivative of the function in (14), which is

$$\frac{\partial f}{\partial p_i} = \frac{a(m_1 - m_2)}{(p_i m_2 + a)^2},$$

[5] We have not tried to optimize this parameter, we have chosen it very small once for all just to create a credal set close to that modeling ignorance.

is a ratio with positive denominator, as it follows from Eqs. (9) and (12). It follows that:

- if $m_1 > m_2$, the derivative is positive over the interval $(\epsilon, 1-\epsilon)$; the function is minimized by setting $p_i := \epsilon$;
- if $m_1 < m_2$, the derivative is negative over the interval $(\epsilon, 1-\epsilon)$; the function is minimized by letting $p_i := 1 - \epsilon$;
- if $m_1 = m_2$, the function is constant.

These three rules define the graph over which CMA will concentrate the prior probability when testing whether c_1 dominates c_2; hence, if $P(c_1|\mathbf{d})/P(c_2|\mathbf{d}) > 1$ under this prior, the same will happen under all the remaining priors of the credal set, and hence class c_2 can be safely dropped.

As a side remark, let us note that $(m_1 - m_2) = (\theta^*_{i1K} - \theta^*_{i2K})$. Hence, the architecture over which CMA concentrates the mass when testing whether c_1 dominates c_2 can be defined also in an alternate way, i.e., feature X_i is linked to the class node if and only if its addition decreases the ratio $P(c_1|\mathbf{d})/P(c_2|\mathbf{d})$ computed by the BMA summary network. It also interesting to note that CMA has the freedom to change architecture depending on the specific pair of classes that are compared.

Software availability The software implementing CMA has been realized in Java; we plan to release the package soon under the GNU GPL license. Sources, binaries and documentation (both user manual and sources documentation in javadoc format) will be available from the website `http://www.idsia.ch/~giorgio/jncc2.html`. Meanwhile, it is possible to obtain the software by contacting the authors by e-mail.

3 Experiments

We present the results obtained on 31 data sets from the UCI repository. The data sets cover a wide spectrum of conditions in terms of number of instances (min: 57, labor; max: 4601, spambase), number of feature variables (min: 3, haberman; max: 69, audiology) and number of classes (up to 24, audiology). On each data set, the classifiers have been evaluated via 10 runs of 10 folds cross-validation. Numerical features have been discretized via MDL-based discretization [9]; in each training/test experiment, the discretization intervals have been computed on the training set and then applied unchanged on the test set.

Some questions of interest are then: is CMA truly able to isolate instances which are hard to classify for BMA? How does BMA behave on the instances which are classified determinately and indeterminately by CMA? Are indeterminate classifications informative and reliable?

We start our analysis by measuring the accuracy of BMA on the instances classified determinately and indeterminately by CMA; these two indicators are

Dataset	Accuracies			
	BMA	BMA	BMA	CMA
	(Avg.)	(Cma D)	(Cma I)	Determ.
anneal	97.9%	98.7%	71.2%	97.3%
audiology	73.5%	99.5%	63.7%	27.0%
autos	66.7%	81.3%	31.7%	70.8%
balance-scale	72.8%	72.8%	n.a	100.0%
breast-cancer	74.9%	83.4%	68.0%	44.9%
c-14-heart-disease	83.0%	85.8%	60.9%	88.3%
cmc	50.3%	58.5%	39.4%	57.7%
credit-rating	85.5%	90.1%	51.8%	88.0%
german_credit	73.7%	87.1%	61.7%	46.9%
glass	71.1%	71.4%	60.3%	98.8%
haberman	71.8%	77.2%	50.8%	81.4%
heart-statlog	83.1%	85.1%	53.0%	93.6%
hepatitis	84.3%	95.6%	72.3%	51.4%
horse-colic	81.2%	86.2%	58.2%	82.1%
h-14-heart-disease	84.3%	85.6%	64.9%	94.2%
ionosphere	89.9%	89.9%	n.a	100.0%
iris	93.7%	93.7%	n.a	100.0%
kr-vs-kp	88.0%	93.7%	60.5%	82.9%
labor	86.9%	98.6%	82.3%	32.2%
liver-disorders	57.4%	60.0%	48.9%	80.1%
lymphography	81.1%	96.1%	73.5%	33.3%
pima_diabetes	75.7%	77.3%	37.2%	95.8%
segment	92.5%	92.5%	60.0%	99.9%
soybean	91.9%	95.8%	26.3%	94.2%
spambase	89.8%	89.8%	n.a.	100.0%
vote	90.2%	90.2%	75.0%	99.8%
wisc-breast-cancer	97.1%	97.1%	n.a	100.0%
yeast	57.2%	57.2%	29.5%	99.9%
zoo	96.4%	98.1%	65.1%	94.2%
primary-tumor	36.8%	83.4%	25.9%	19.0%
contact-lenses	87.3%	100.0%	85.7%	22.2%
average	79.6%	86.2%	54.7%	75.9%

Table 1. Comparison of BMA and CMA on 31 UCI data sets. Note that the indicator BMA(CMA I) is not available for those data on which CMA achieves 100% determinacy.

denoted respectively by BMA(CMA D) and BMA(CMA I). If CMA is able to recognize instances that are hard to classify, we should observe a significant drop of BMA accuracy between the former and the latter set of instances.

These results are shown in Table 1, which also reports, to complement the information, the average accuracy of BMA. On average, there is a drop of 32 points between BMA(CMA D) and BMA(CMA I); moreover, on *every* data set we clearly observe that BMA(CMA D) is strictly larger that BMA(CMA I); hence, we can safely state that CMA isolates instances that are hard to classify and where, as a consequence, BMA becomes less reliable.

On the other hand, CMA reacts to the hard instances by returning indeterminate classifications. Another point of interest is hence to evaluate the informative content of the indeterminate classification; this can be properly assessed only on data sets with at least three classes, since, on data sets with two classes, indeterminate classifications contain all the classes.[6] Excluding hence data sets with two classes from the analysis, we have measured on average that set-valued classifications return 35% of the classes of the data set, dropping hence 65% of them; therefore, they convey significant information. Moreover, set-valued classifications are very reliable; in fact, they contain the actual class in 90% of cases. Summing up, CMA is able to detect hard instances where the accuracy of BMA drops indeed; on these instances, indeterminate classifications preserve the reliability of CMA, by conveying reliable information, without however drawing too strong conclusions.

A further important indicator of performance is the determinacy of CMA, i.e., the percentage of instances over which CMA returns a single class; on average, CMA achieves 77% determinacy, i.e., it yields set-valued classification on 23% of instances. The data sets which lead to the largest indeterminacy are characterized by a small number of instances and a relatively high number of feature variables/classes; see for instance: primary-tumor (339 instances, 17 feature variables, 22 classes, determinacy: 19%), contact-lenses (24 instances, 4 feature variables, 3 classes, determinacy: 22%), audiology (226 instances, 69 feature variables, 24 classes, determinacy: 27%). However, the caution of CMA on these data sets is justified, as the drop between BMA(CMA D) and BMA(CMA I) is respectively of 57.5, 14.3 and 35.9 points. On the other hand, the determinacy of CMA quickly increases on data sets which contain more instances or less features variables.

3.1 BMA probabilities Vs. CMA set-valued classifications

We have shown that, thanks to imprecise probabilities, CMA delivers set-valued classifications on hard-to-classify instances, over which the accuracy of BMA clearly drops. In the following, we analyze the association between the posterior probabilities computed by BMA and the set-valued classifications returned by

[6] Nevertheless, we deem set-valued classifications to be valuable also in the case of data sets with two classes only, as they highlight that a certain classification is doubtful, thus preventing an over-confident use of the output of the model.

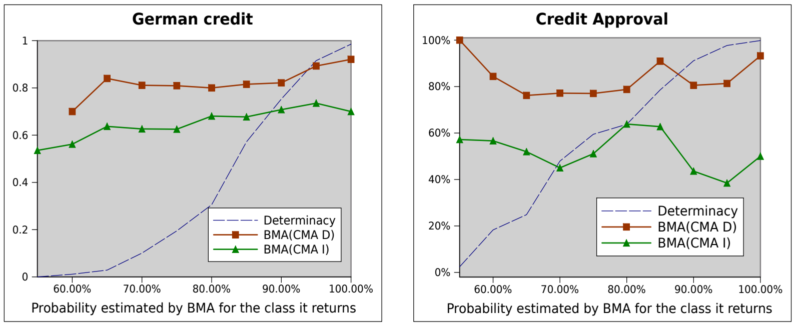

Fig. 2. Relationship between the posterior probabilities computed by BMA and the output of CMA.

CMA. To this purpose, we focus on the example of the German credit data set, which is made of 2 classes, 20 feature variables and 1000 instances. As the data set has two classes, it is easy to spot instances that are deemed doubtful according to BMA (they are classified with probability lower than, say, 55%) and to CMA (indeterminate classifications).

To perform our analysis, we consider four pieces of information for each instance: (i) the actual class, (ii) the class returned by BMA, (iii) its probability, and (iv) whether the instance has been classified determinately or indeterminately by CMA. The instances are then partitioned into subsets, according to the probability estimated by BMA for the returned class, i.e., instances for which BMA estimates a probability in the range 50–55%, 55–60%, and so on (i.e., we use a step of 5% in probability to define the subsets). On each subset of instances, we measure: (a) the determinacy of CMA; (b) BMA(CMA D) and (c) BMA(CMA I).

The results are reported in Figure 2. There is a positive association between higher posterior probabilities computed by BMA and higher determinacy; indeed the choice of the prior over the graphs is less likely to change the classification outcome when the probability computed by BMA for the most probable class increases. The output of CMA is indeterminate for all the instances as long as the probability estimated by BMA for the returned class is lower than 55%. Hence, on the instances classified with probability less that 55% by BMA, BMA and CMA convey a similar message, i.e., that of a doubtful classification: BMA by returning a low probability for the class, CMA by becoming indeterminate.

Moving on to greater probabilities, the determinacy of CMA rises progressively; however, CMA keeps returning a mix of determinate and indeterminate classifications, even on instances classified very confidently by BMA (for instance, with probability higher than 80%). The point is that the behavior of CMA is justified, since at any level of posterior probability estimated by BMA there is a clear drop of accuracy between BMA(CMA D) and BMA(CMA I).

Similar patterns have been observed on most of the data sets with two classes included in our list; we report for instance in Figure 2 also the results obtained on the credit approval data set.[7]

From Figure 2, one can also appreciate that the behavior of CMA cannot be mimicked by a BMA with threshold, i.e., a BMA which returns two classes unless the probability for the most probable class exceeds a fixed threshold t. In fact, a BMA with threshold would assume all instances classified with probability less than t to be hard; instead CMA identifies in a sensible way a mix of easy and hard instances between both the instances classified with probability greater or smaller than the threshold. Moreover, CMA is able to detect hard instances also among those classified with very high probability from BMA, something which would not be possible to accomplish with a BMA with threshold.

4 Conclusions

In this paper we have proposed an extension of Bayesian model averaging to imprecise probabilities that we have called credal model averaging. By CMA, we have tried to tackle one of the more serious challenges of BMA, which is related to the choice of the prior over the models: both the difficulty in defining such a prior, and the unavoidable arbitrariness that any choice entails. In our approach, prior beliefs model a condition close to ignorance about the models, thus trying to implement an objective-minded approach to model averaging. This naturally leads to a new form of averaging whose conclusions are robust to the definition of prior beliefs.

We have applied CMA in particular to problems of classification based on naive Bayes nets. Our empirical experiments over many data sets have confirmed that CMA leads to reliable inference. It leads, in particular, to create classifiers that can suspend the judgment when the conditions do not justify strong conclusions, and that we have called credal classifiers. What we have seen clearly from the experiments is that suspending judgment has been well motivated: the attempt of BMA-based classifiers to produce a determinate classification when CMA leads to suspend judgment, yields fragile classifications that heavily deteriorate the predictive performance of the former.

In summary, CMA approaches in an original way the controversial problem of setting the prior over the models for BMA; moreover, it performs well and reliably in classification problems.

CMA has been derived assuming to have a complete data set; it would be however interesting to extend CMA to deal also with missing data. If one assumes the missing data to be generated by a missing-at-random (MAR) process, realizing such an extension would be straightforward. However, the MAR assumption is not always met; therefore, a more sophisticated treatment of missing data

[7] To prevent ambiguities, we point out that German-credit and credit approval are two distinct data sets; the former has been donated by Prof. Hofmann, and contains 20 feature variables; the latter has been donated by Prof. Quinlan, and contains 15 feature variables.

should be developed, able to deal with missing data differently, depending on whether they are generated by a MAR or non-MAR missingness process. We have followed a similar avenue in [7]; however, incorporating such a treatment of missing data into CMA could be technically quite involved and therefore it needs careful investigation.

Acknowledgments Work for this paper has been partially supported by the Swiss NSF grants 200021-113820/1 and 200020-116674/1, and by the Hasler Foundation (Hasler Stiftung) grant 2233.

References

1. Hoeting, J., Madigan, D., Raftery, A., Volinsky, C.: Bayesian Model Averaging: a Tutorial. Statistical Science **14**(4) (1999) 382–417
2. Clyde, M., George, E.I.: Model Uncertainty. Statistical Science **19** (2004) 81–94
3. Boullé, M.: Compression-Based Averaging of Selective Naive Bayes Classifiers. Journal of Machine Learning Research **8** (2007) 1659–1685
4. Dash, D., Cooper, G.: Exact Model Averaging with Naive Bayesian Classifiers. Proceedings of the Nineteenth International Conference on Machine Learning (2002) 91–98
5. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)
6. Zaffalon, M.: The Naive Credal Classifier. Journal of Statistical Planning and Inference **105**(1) (2002) 5–21
7. Corani, G., Zaffalon, M.: Learning Reliable Classifiers from Small or Incomplete Data Sets: the Naive Credal Classifier 2. Journal of Machine Learning Research **9** (2008) 581–621
8. Abellán, J., Moral, S.: A New Score for Independence Based on the Imprecise Dirichlet model. In Cozman, F.G., Nau, R., Seidenfeld, T., eds.: ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Manno, Switzerland, SIPTA (2005) 1–10
9. Fayyad, U.M., Irani, K.B.: Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, San Francisco, CA, Morgan Kaufmann (1993) 1022–1027