Credal model averaging of logistic regression for modeling the distribution of marmot burrows

G. Corani
IDSIA (Switzerland)
giorgio@idsia.ch

A. Mignatti
Politecnico di Milano (Italy)
mignatti@elet.polimi.it

Abstract

Bayesian model averaging (BMA) weights the inferences produced by a set of competing models, using as weights the models posterior probabilities. An open problem of BMA is how to set the prior probability of the models. Credal model averaging (CMA) is a credal ensemble of Bayesian models, which generalizes BMA by substituting the single prior over the models by a set of priors. The base models of the ensemble are learned in a Bayesian fashion. We use CMA to ensemble base classifiers which are Bayesian logistic regressors, characterized by different sets of covariates. CMA returns indeterminate classifications when the classification is prior-dependent, namely when the most probable class depends on the prior probability assigned to the different models. We apply CMA for modelling the presence and absence of marmot burrows in an Alpine valley in Italy and show that it compares favorably to BMA.

Keywords. Bayesian model averaging, credal model averaging, logistic regression, classification, ecological modeling.

1 Introduction

Over the last years, classifiers based on imprecise probabilities have been mostly developed by extending probabilistic graphical models (see [30] for a pioneering work and [7] for a recent review) or decision trees (see [1] and the references therein). Alternatively, extension of the k nearest neighbors have been also proposed [11].

In this paper we consider the idea of credal model averaging (CMA) [8, 6], which can be described as a credal ensemble of Bayesian classifiers. In other words, the parameters of the base models are learned in a Bayesian way. The ensemble of the base models is instead carried out in an imprecise way, modelling a condition of ignorance about the prior probability of the different models.

Model uncertainty is the problem of many models being consistent with the available data. In this condition, there is substantial uncertainty about which model should be chosen for drawing inferences or computing predictions. Choosing a single model and then ignoring the substantial uncertainty of the model selection leads to overconfident inferences [3]. Bayesian model averaging (BMA) is a sound approach to deal with model uncertainty, based on the key idea of averaging the inferences produced by a set of different models, using the models' posterior probabilities as weights.

However, BMA requires to specify the prior probability of each model. This is a critical issue, as it is recognized in the BMA literature [5]. To tackle this issue, some authors repeat the BMA analysis assigning different prior probabilities to the models [21, 28]. From the viewpoint of the credal classification, it is well-known the relying on a single prior implies unavoidable arbitrariness, which entails the risk of drawing prior-dependent classifications.

CMA generalizes BMA, overcoming the problem of the prior specification by adopting a *set* of prior over the models. As a result, the posterior probability of the models lies within an interval rather being a punctual value. Moreover, CMA

automatically detects the instances which are prior-dependent, namely whose most probable class varies depending on the prior probability assigned to the different models. On such instances, CMA suspends the judgment by returning more than one class, thus automating the sensitivity analysis. So far, CMA has been proven effective in ensembling probabilistic graphical models [8, 6].

We develop CMA for ensembling logistic regression models characterized by different feature sets. Indeed, BMA of logistic regressors was already used to model presence or absence of ecological populations [21, 28]; we then compare BMA and CMA on the case study of predicting the presence of marmot burrows in an Alpine valley.

2 Bayesian model averaging (BMA)

Let us consider a logistic regression model for predicting the value of a binary class C, with classes c_0 and c_1. The set of *covariates* (or *features*) is $\mathcal{X} = \{X_1, X_2, \dots X_k\}$; in a generic instance, the value of the covariates is $\mathbf{x} = x_1, \dots, x_k$. We denote $\pi_0 = P(C = c_0|\mathbf{x})$ and $\pi_1 = P(C = c_1|\mathbf{x})$. The logistic regression model is

$$y = \text{logit}(\pi_0) = ln\frac{\pi_0}{1-\pi_0} = ln\frac{\pi_0}{\pi_1} = \\ = \beta_0 + \sum_{j=1}^{j=k} \beta_j x_j \qquad (1)$$

where x_j is the observation of X_j.

Given k covariates, the model space $\mathcal{M}$ is composed of 2^k possible model *structures*. Each model structure includes a specific set of covariates. We denote by m_i the i-*th* structure. The model size is defined as the number of covariates included in the structure.

Feature selection is the problem of identifying the supposedly best set of covariates for the model. The traditional feature selection approach is to assess the significance of each covariate through hypothesis tests [10]. More modern approaches for feature selection are instead based on the so-called Information Criteria [3], such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC)[1]. Information Criteria have been recognized to be more effective than repeated hypothesis tests for the purpose of feature selection [3]. Yet, even adopting Information Criteria one could face the problem of *model uncertainty*. If, for instance, different models obtain a similar value of BIC, a substantial uncertainty underlies the choice of a single model. The subsequent inferences are hence overconfident if this uncertainty is disregarded.

BMA addresses model uncertainty by combining the inferences of multiple models, using as weights the posterior probability of the models. We denote by D the available dataset, by $P(m_i|D)$ the posterior probability of model m_i and by $P(Y|D)$ the entire posterior distribution of Y given D, from which posterior probabilities $P(y|D)$ of a specific value y can be obtained. The posterior of Y under BMA is [5]:

$$P(Y|D) = \sum_{m_i \in \mathcal{M}} P(Y|m_i, D)P(m_i|D) \qquad (2)$$

where:

$$P(m_i|D) = \frac{P(m_i)P(D|m_i)}{\sum_{m_k \in \mathcal{M}} P(m_k)P(D|m_k)}$$

$$P(D|m_i) = \int P(D|\boldsymbol{\beta_i}, m_i)P(\boldsymbol{\beta_i}|m_i)d\boldsymbol{\beta_i},$$

having denoted by $P(m_i)$ the prior probability of model m_i, $\boldsymbol{\beta_i}$ the vector of its parameters and $P(D|m_i)$ its marginal likelihood, which in the linear case can be exactly computed [25]. Equation (2) requires an extensive summation over 2^k models, which is usually carried out by sampling the model space . Only for small k it is possible to exhaustively treat the model space.

As a result of averaging across different models, $P(Y|D)$ is given by a sum of distributions and thus has a multi-modal shape. Inferences about other quantities of interest such as the parameter of the models can be obtained by averaging over the models as in Eq(2).

BMA requires to set a precise prior over the parameters and over the models. As a prior distribution on the parameters $P(\boldsymbol{\beta_i}|m_i)$ we adopt Zellner's g-prior [13], setting g equal to the number of observations. As for the prior over the

[1] The BIC provides a simple but effective approximation of the posterior probability of a given model [24].

models, we adopt the binomial prior [25, 13]; namely, every covariate has the same prior probability θ of being included in the model; moreover, the probability of inclusion of each covariate is *independent*. Thus, the prior probability of model m_i, which includes a number k_i of covariates, is:

$$P(m_i) = \theta^{k_i}(1-\theta)^{k-k_i}. \quad (3)$$

Once the prior probability of each possible model is specified according to Eq.3, it can be analyzed the prior distribution of the random variable constituted by the *model size*, namely the number of covariates included in the model. The model size follows a binomial distribution with mean θk and variance $\theta(1-\theta)k$ [19], where k is the total number of available covariates. An easy way to elicit the prior distribution over the models is to ask the expert his beliefs about the model size.

3 Credal Model Averaging(CMA)

CMA generalizes BMA by substituting the *single* binomial prior over the models by a *set* of binomial priors: thus, the prior probability of inclusion of each covariate varies within the range $[\underline{\theta}, \overline{\theta}]$; thus, the mean model size a priori varies within the range $[\underline{\theta}k, \overline{\theta}k]$. Thus, CMA allows eliciting from the expert an *upper* and a *lower* model size. If no expert is available, one can model a situation of ignorance a priori, by setting $\underline{\theta} = \epsilon$ and $\overline{\theta} = 1 - \epsilon$. In our experiments we adopt this approach, setting ϵ=0.05.

Each model of the ensemble is learned in a Bayesian fashion, using a *precise* prior over the parameters. Instead, the prior probability of the models is imprecisely modelled. Hence CMA is a *credal ensemble of Bayesian models*. Because of imprecision, CMA computes for the logit the interval $[\underline{y}, \overline{y}]$ rather than a point value as in traditional logistic regression. The length of such interval varies instance by instance, showing the sensitivity of the prediction on the priors which has been set over the models, namely how much the BMA prediction would vary as a consequence of θ varying between $\underline{\theta}$ and $\overline{\theta}$. No coverage probability can be assigned to the CMA intervals. To compute $\overline{y}$ and $\underline{y}$, CMA solves a maximization and a minimization problem on each instance. Since the prior probability of inclusion θ is equal for all covariates, the optimization problem involves only a single variable.

Let us focus on the minimization case. We denote by a hat the estimated values. Given the data set D and the observation $\mathbf{x} = x_1, \ldots, x_k$ of the covariates, we denote the prediction of model m_i as $\hat{y}_i = \beta_0^i + \sum_{j=1}^{j=k_i} \beta_j^i x_j$, where β_0^i and β_j^i denote the parameters of m_i (the previous formula assumes, with no loss of generality, that for model m_i the covariates have been re-ordered, so that the first k_i covariates are those included in the model). For simplicity of notation we do not indicate the dependence of $\hat{y}_i$ on D and $\mathbf{x}$. The lower bound $\underline{\hat{y}}$ of the CMA interval is computed as:

$$\underline{\hat{y}} = \min_{\theta\in[\underline{\theta},\overline{\theta}]} \sum_{m_i\in\mathcal{M}} \hat{y}_i P(m_i|D) =$$
$$\min_{\theta\in[\underline{\theta},\overline{\theta}]} \sum_{m_i\in\mathcal{M}} \hat{y}_i \frac{P(D|m_i)P(m_i)}{\sum_{m_j\in\mathcal{M}} P(D|m_j)P(m_j)} =$$
$$= \min_{\theta\in[\underline{\theta},\overline{\theta}]} \frac{\sum_{m_i\in\mathcal{M}} \hat{y}_i P(D|m_i)\theta^{k_i}(1-\theta)^{k-k_i}}{\sum_{m_j\in\mathcal{M}} P(D|m_j)\theta^{k_j}(1-\theta)^{k-k_j}}$$
$$:= \min_{\theta\in[\underline{\theta},\overline{\theta}]} h(\theta)$$

Let us define the k sets $\mathcal{M}_1 \ldots \mathcal{M}_k$ which include all the models containing respectively $\{1, 2, \ldots, k\}$ covariates. For instance, $\mathcal{M}_2$ contains all the models which include two covariates. To address the optimization problem it is useful noting that all the models contained in the set $\mathcal{M}_j$ have the same prior probability $\theta^j(1-\theta)^{k-j}$. We introduce $Z_j = \sum_{m_v\in\mathcal{M}_j} \hat{y}_v P(D|m_v)$ and $L_j = \sum_{v\in\mathcal{M}_j} P(D|m_v)$ and then rewrite function $h(\theta)$ as:

$$h(\theta) = \frac{\sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} Z_j}{\sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} L_j} \quad (4)$$

In the interval $[\underline{\theta}, \overline{\theta}]$, the maximum and minimum of $h(\theta)$ should lie either in the boundary points $\theta = \overline{\theta}$ and $\theta = \underline{\theta}$, or in an internal point of the interval in which the first derivative of $h(\theta)$ is 0. Let us introduce $f(\theta) = \sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} Z_j$ and $g(\theta) = \sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} L_j$. The first derivative $h'(\theta)$ is:

$$h'(\theta) = \frac{f'(\theta)g(\theta) - f(\theta)g'(\theta)}{g(\theta)^2}, \quad (5)$$

where $g(\theta)$ is strictly positive because L_j is a sum of marginal likelihoods. We can therefore search the solutions looking only at the numerator $f'(\theta)g(\theta) - f(\theta)g'(\theta)$, which is a polynomial of degree $k(k-1)$ and thus has $k(k-1)$ solutions in the complex plain. We are interested only in the *real* solutions that lie in the interval $(\underline{\theta}, \overline{\theta})$. Such solutions, together with the boundary solutions $\theta = \overline{\theta}$ and $\theta = \underline{\theta}$, constitute the set of *candidate solutions*. To find the minimum and the maximum $h(\theta)$, we evaluate $h(\theta)$ in each candidate solution point, and eventually we retain the minimum and maximum among such values.

Having determined the upper and lower logit values $\underline{y}$ and $\overline{y}$, we obtain the upper and lower posterior probabilities of the two classes by inverting Eq.(1):

$$\overline{\pi}_0 = \frac{\exp(\overline{y})}{1 + \exp(\overline{y})}$$
$$\underline{\pi}_0 = \frac{\exp(\underline{y})}{1 + \exp(\underline{y})}$$
$$\overline{\pi}_1 = 1 - \underline{\pi}_0$$
$$\underline{\pi}_1 = 1 - \overline{\pi}_0$$

CMA adopts the criterion of *interval-dominance* [27] to take decisions: class c_1 is returned if $\underline{\pi}_1 > \overline{\pi}_0$, namely if $\underline{\pi}_1 > 1/2$. Conversely, class c_0 is returned if $\underline{\pi}_0 > 1/2$. In these cases the instance is *safe* because the rank between the two classes is the same regardless the prior probabilities assigned to the competing models. If instead the intervals of the posterior probability of the two classes overlap, the judgment is suspended. The instance is *prior-dependent*, since the rank among the classes changes when different prior probabilities are assigned to the competing models.

A final note regards the relation between the logit computed by BMA and CMA. If the value of θ used to induce BMA is included in the interval $[\underline{\theta}, \overline{\theta}]$ used to induce CMA, the logit computed by BMA is included within the the logit interval computed by CMA. Thus when CMA returns a single class, this is the same class returned by BMA.

4 Case study

The study area is located in the Italian Alps, near the Stelvio National Park. The valley has an altitude comprised between 2100 and 3100 m above sea level. The field surveys identified the position of the Alpine marmot burrows and the characteristics of their surrounding territory. The censuses were carried out in the summers 2010 and 2011; three different areas of the valley were investigated. To develop the species distribution model we divide the area into cells of 100m^2; the censused area is overall of about 95 ha (9500 cells). Presence of burrows has been detected in about 4.5% of the cells. Each cell is then labelled as presence or absence.

The considered covariates are altitude, slope, aspect (the direction in which the slope faces) topographic ruggedness index (TRI) [26], hillshade, curvature, soil temperature and soil cover. For the aspect, we did not directly use the angle from North, but we divided the information into two sub-variables that we called *northitude* and *eastitude*. The *northitude* is calculated as the cosine of the angle from North, while the *eastitude* is calculated as the sine of the same quantity. While the former represents the attitude of the marmot to select sunny slopes, the latter represents the preference to have a sunny territory during the sunrise and the morning rather than during the sunset and the evening. To build the soil temperature map we relied on five different meteorological stations (altitude comprised between 1800 and 2600 m a.s.l.) located in the surroundings, which provide the data of air temperature and snow depth. The soil temperature is a mean yearly value and was calculated starting from the DEM (digital elevation model) and the data of air temperature and snow depth, using the model developed by [14]. Finally, we express the soil cover as the percentage of cells with debris and outcrops cover in the buffer area (see later for an explanation of the buffer area).

As a pre-processing step we removed some highly cross-correlated ($|\rho| > .8$) covariates: more precisely the soil temperature (anti-correlated with the altitude), the TRI (anti-correlated with the slope) and the hillshade, correlated with the *northitude*.

The Alpine marmot is a mobile species, which uses a huge territory for its activities. Thus, we supposed that the decision to dig a burrow in a given cell does depend also on the environmental conditions of surrounding cells. For

this reason, the value associated to each cell (for each environmental variable) is calculated as the mean of the values of the variable in a surrounding of the same cell. We refer to this area with the term *buffer area*, and, in our case, it has a pseudo-circular shape, since we considered the cells within a circular area built around the given cell. The home range of the Alpine marmot ranges between 1 and 3 ha [23, 17]. We considered buffer areas of size 1 ha, 2 ha and 3 ha. Since the results are quite consistent when different buffer areas are considered, in the following we present results referring only to a buffer area of 2 ha.

5 Results

To gain understanding of the data and to investigate the role of the different covariates, we develop a BMA model using the entire dataset. The prior probability of inclusion of the covariates is set to 0.5, corresponding to a uniform prior probability over the models.

Under BMA the posterior probability of inclusion of a covariate is calculated as the sum of the posterior probability of the models in which the covariate is included. In Table 1 we report the posterior probability of inclusion of the covariates, the expected values and the standard deviations of the parameters of the models, obtained using the standardized values of the variables. The expected values and the standard deviations of the coefficients are calculated averaging over the models which do include the covariate.

Variable	2ha		
	p.inc.	EV	SD
altitude	1	-1.050	0.158
slope	1	0.491	0.067
curvature	0.02	0.001	0.011
northitude	1	-1.381	0.010
eastitude	1	-0.553	0.056
% of outcrops and debris cover	0.97	-0.399	0.122

Table 1: Posterior probability of inclusion of the covariates (p.inc), expected values (EV) and standard deviations (SD) of the model parameters.

The most important variables are the altitude, the slope, the *eastitude* and the *northitude*. The signs of the parameters confirm, for most of the variable, what is reported in literature. The coefficient of the altitude has a negative value, and the valley altitude ranges from ca. 2200 m a.s.l. and 3000 m a.s.l.. The suitable altitude for the marmot is approximately between 1650 m a.s.l. and 1950 m a.s.l. [4, 2] with maximum altitudes around 3000 m a.s.l.. Since the valley is above the optimal altitude range of the marmot, the fact that the suitability of the valley decreases with the altitude confirm the past results. The slope positively influences the presence of burrows. In this case, we have conflicting results reported in literature, with an optimal slope that varies from 0 to 60°[22]. The *northitude* negatively influences the presence of burrows, so that the marmot preference is for southerly exposed slopes, as previously reported in several studies [2]. The *eastitude* negatively influences the presence of burrows, contrary to what is reported in literature [2], with a preference for the westerly exposed slopes in the valley. This preference is probably due to the fact that, in the valley, the areas located at a higher elevation and with a low suitability, are mainly westerly exposed. This result seems therefore to be mainly due to the valley shape. A high percentage of outcrops and debris cover negatively influences the presence of marmot burrows, showing the importance of the alpine meadows for the species, as reported by [2, 22].

5.1 Comparing BMA and CMA

We compare BMA and CMA using training data sets of varying sample size. For comparing BMA and CMA, we downsample the original data set, generating training sets of size $n \in \{30, 60, 90 \ldots, 300\}$. For each sample size, we build 30 different training sets. The instances not contained in the training set constitute the test set. The training sets contain the same prevalence (fraction of presence data) of the entire dataset, namely 4.6%. For CMA we assume a situation of substantial ignorance a priori, setting $\overline{\theta} = 0.95$ and $\underline{\theta} = 0.05$.

CMA can be seen as dividing the instances into two groups: the *safe* ones, for which a single class is returned, and the *prior-dependent* ones, for which instead the judgment is suspended and both classes are returned. For the prior-dependent instances, presence or absence is more probable depending on the prior probability of the competing models.

The most common measure of performance in classification is the *accuracy*, defined as the proportion of instances correctly classified. To evaluate the effectiveness of CMA, we assess the accuracy of BMA on the safe and on the prior-dependent instances. As can be seen in Fig.1, BMA undergoes a sharp drop of accuracy on the instances indeterminately classified by CMA.

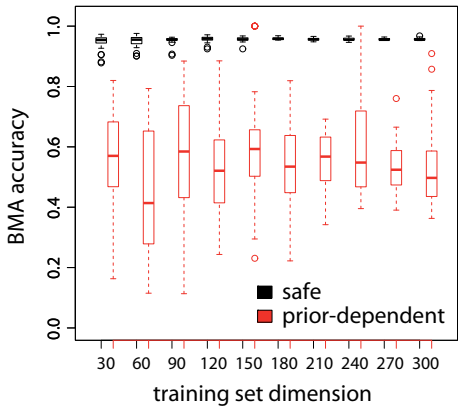

Figure 1: The accuracy of BMA drops on the prior-dependent instances. For each sample size, the boxplot refers to 30 experiments.

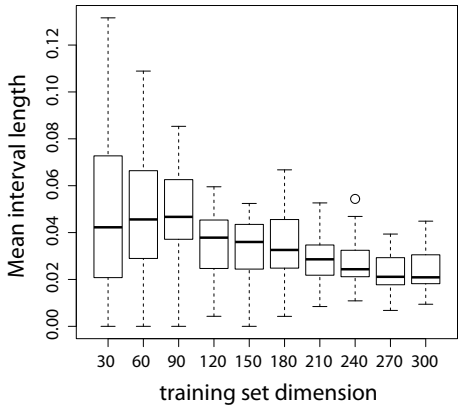

Figure 2: The length of the CMA interval $(\underline{\pi}_0, \overline{\pi}_0)$ decreases with the sample size. For each sample size, the boxplot refers to 30 experiments.

The length of the logit interval $[\underline{y}, \overline{y}]$ of CMA decreases with the dimension of the training set as shown in Figure 2: the larger the sample size, the less influential the prior probability of the models.

5.2 Credal Classification and Reject Option

Traditional classifiers can be equipped with a *reject option* [16], thus refusing to classify an instance if the posterior probability of the most probable class is below a certain threshold. To adopt the reject option, it is necessary setting the *rejection cost* which is incurred into when rejecting an instance. When classifying an instance, the *expected cost* [12] associated to decision of returning each class is computed. The instance is rejected if the expected classification cost of each class is higher than the rejection cost. This corresponds to rejecting all the instances in which the posterior probability p^* of the most probable class is below a threshold t [16].

However, the behavior induced by the reject option is quite different from that of a credal classifier. On a *large* data set the posterior probability of the classes is *not* sensitive on the choice of the prior; a credal classifier will generally return a single class. On the other hand, the determinate classifier could reject even a considerable number of instances, if the rejection cost is small. To fairly compare a traditional classifier equipped with rejection option against a credal classifier, it would be necessary making the credal classifier aware of the rejection cost. This point we leave for future research.

However, applying a *rejection option* to BMA does in general yield a behavior which is quite different from that of CMA. The point is that on the prior-dependent instances the BMA predictions are *not tightly* distributed around a 50% posterior probability; instead, there are many prior-dependent instances in which BMA estimates a posterior probability larger than 60-70% for the most probable class: see for an example Figure 3. Thus, BMA equipped with rejection option would reject only part of the prior-dependent instances. Conversely, it will instead reject some instances which are not prior-dependent.

5.3 Utility-discounted accuracy

To further compare CMA and BMA we adopt the utility-discounted accuracy introduced in [29]. We briefly summarize here the idea underlying this approach. The starting point is the *discounted accuracy*, which rewards a prediction

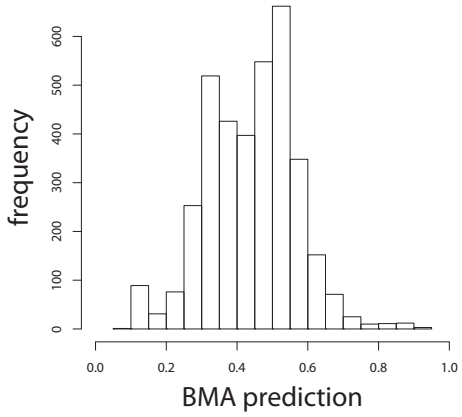

Figure 3: Distribution of the posterior probability associated by BMA to the most probable class in the *prior-dependent* instances. The figure refers to a training set of dimension n=210.

containing m classes with $1/m$ if it contains the true class, and with 0 otherwise. Within a betting framework based on fairly general assumptions, discounted-accuracy is the only score which satisfies some fundamental properties for assessing both determinate and indeterminate classifications; thus, the discounted accuracy of a credal classifier can be compared to the accuracy achieved by a determinate classifier. Yet discounted-accuracy has severe shortcomings. Consider two medical doctors, doctor *random* and doctor *vacuous*, who should diagnose whether a patient is *healthy* or *diseased*. Doctor *random* issues uniformly random diagnosis; doctor *vacuous* instead always returns both categories, thus admitting his/her ignorance. Let us assume that the hospital profits a quantity of money proportional to the discounted-accuracy achieved by its doctors at each visit. Both doctors have the same *expected* discounted-accuracy for each visit, namely 1/2. For the hospital, both doctors provide the same *expected* profit from each visit, but with a substantial difference: the profit of doctor vacuous has no variance. Any risk-averse hospital manager should thus prefer doctor vacuous over doctor random: under risk-aversion, the expected utility increases with expectation of the rewards and decreases with their variance [18]. To model this fact, it is necessary to apply a utility function to the discounted-accuracy score assigned to each instance. The utility function is designed as follows in [29]: the utility of a correct and determinate classification (discounted-accuracy 1) is 1; the utility of a wrong classification (discounted-accuracy 0) is 0. Therefore, the utility of a traditional determinate classifier corresponds to its accuracy. The utility of an accurate but indeterminate classification consisting of two classes (discounted-accuracy 0.5) is assumed to lie between 0.65 and 0.8. Two quadratic utility functions are then derived corresponding to these boundary values, and passing respectively through $\{u(0) = 0, u(0.5) = 0.65, u(1) = 1\}$ and $\{u(0) = 0, u(0.5) = 0.8, u(1) = 1\}$, denoted as u_{65} and u_{80} respectively. Since $u(1) = 1$, utility and accuracy coincide for determinate classifiers; therefore, utility of credal classifiers and accuracy of determinate classifiers can be directly compared. Interestingly, the u_{65} and u_{80} functions provides score which are numerically close to respectively the F_1 and F_2 metric, which have been used to score indeterminate classifications in [9], adopting an approach based on information retrieval.

In Figure 4 we compare the CMA utility (calculated using the u_{80} utility function) and the BMA accuracy. The utility produced by CMA is slightly higher on average than that of BMA; however the most striking feature of Fig.4 is that the CMA boxplots are much tighter than the BMA ones. This means that the utility yielded by CMA is not only higher on average, but also much more stable and predictable than that of BMA. The result do not change substantially if the u_{65} utility function is considered instead, apart from a slight shift downwards of the CMA boxplots.

5.4 The cost-sensitive setup

The classes of our problem are strongly skewed: about 4.5% and 95.5% of the instances are respectively presence and absence. It is unlikely that the two different kind of errors (false presence and false absence) have identical costs, as it is assumed by both the classification accuracy and the utility-discounted accuracy. To make the assessment more realistic, it is thus worth considering a cost-sensitive setup.

A simple measure of performance which accounts for costs is the AUC [20], namely the area under the *receiver operating characteristic* (ROC)

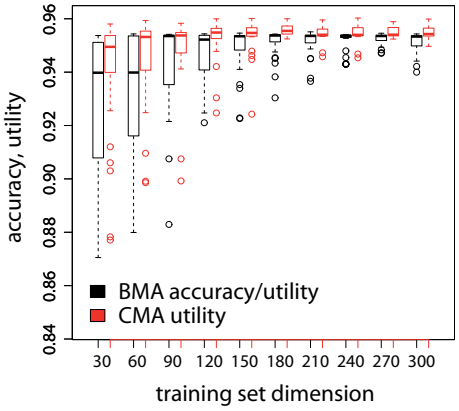

Figure 4: CMA utility compared to BMA accuracy, using the u_{80} utility function.

curve. Figure 5 shows that BMA achieves much higher AUC on the safe instances (determinately classified by CMA) than on the prior-dependent ones (indeterminately classified by CMA). This is a further favorable result for CMA.

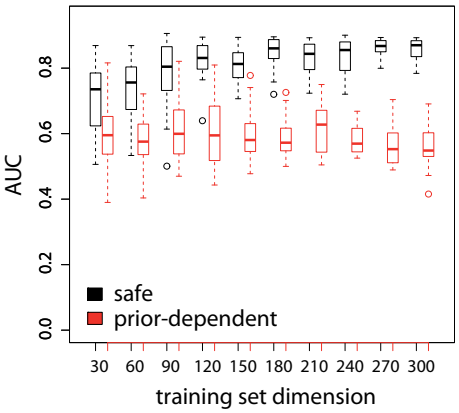

Figure 5: The AUC of BMA drops on the prior-dependent instances.

Yet, the AUC summarizes into a single scalar the whole area under the ROC curve, mixing the performance obtained under very different cost scenarios [20]. To provide a more detailed picture of the behavior of the classifier in the cost-sensitive setup, we then follow the approach of [12]. We introduce the *cost matrix*; in particular, we denote by $d(c_i, c_j)$ the cost of predicting class c_i when the actual class is c_j. The cost matrix is 2x2 since the problem has two classes (presence and absence), as shown in Table 2. Let us assume that the model is used to predict the presence/absence of burrows in a territory that has not yet been censused. If the model predicts the presence of a burrow in a given cell, an operator is sent to search for burrows, incurring the cost κ (this is a simplification, since the cost could vary for instance with the position of the cell to be surveyed). If a burrow is found, a gain ζ is obtained; overall, the negative cost (namely the reward) for having correctly predicted the presence is $\kappa - \zeta < 0$. If absence is predicted no survey is organized; thus, no costs are incurred regardless whether the considered cell contains or not a burrow.

	Actual	
Predicted	Absence	Presence
Absence	0	0
Presence	κ	$\kappa - \zeta$

Table 2: Cost matrix.

In the cost-sensitive setup, the classifier should return the class with the lowest expected cost rather than the most probable class. The expected cost of predicting class c_i is $\sum_{c_j \in \mathcal{C}} \pi_j d(c_i, c_j)$, where $\mathcal{C}$ denotes the set of classes and π_j is the posterior probability of class c_j, computed according to the logistic regression model. Given the above cost matrix, the expected cost of predicting absence is 0. Thus presence is predicted if the expected cost of doing so is negative:

$$\begin{aligned} \text{Expected cost (predicting presence)} < 0 \Leftrightarrow \\ \pi_1(\kappa - \zeta) + \pi_0(\kappa) < 0 \Leftrightarrow \\ \kappa - \pi_1 \zeta < 0 \end{aligned}$$

In other words, presence is predicted if its posterior probability is higher than the threshold $t = \kappa/\zeta$. Dealing with CMA, in some instances the posterior probability of presence might fluctuate below and above the threshold t depending on the prior probability assigned to the competing models. In this case, the decision should be suspended since the evidence coming from data is not strong enough to take a decision. However, we want CMA to take a decision. To this purpose, we consider the Γ-maximin approach [27], namely worst-case optimisation; this implies returning a prediction of *absence* on the prior-dependent instances. We also consider the opposite approach Γ-maximax, namely optimization

of the best case; this implies returning a prediction of *presence* on the prior-dependent instances.

We perform experiments with different values of the threshold $t = \kappa/\zeta$. Moreover, to compare the results obtained with different t, we fixed $\zeta = 1$; it is indeed easy to prove that ζ is only a multiplicative factor in the computation of the total cost, so that its value does not influence the quality of the results. In Figure 7 we report the results for the case t=0.5 ($\zeta = 2\kappa$) and t=1/23 ($\zeta = 23\kappa$). The latter value, in which the threshold equals the marginal probability of presence, is referred to as Kolmogorov-Smirnov statistic in [15]. Given the rarity of presence, we do not consider values of ζ smaller than 2κ, namely $t > 0.5$. Figures 6 and 7 show the results for the prior-dependent instances only; on the instances which are not prior-dependent, BMA and CMA take the same decisions and thus incur the same costs. Given the cost matrix of Table 2, the Γ-maximin strategy incurs a cost of 0 on the prior-dependent instances. In the case $t = 1/2$ (Figure 6), Γ-maximin incur lower costs than if the decision is taken according to the single posterior probability computed by BMA. The highest costs are instead incurred adopting the Γ-maximax strategy. However, the situation is reversed in the case t=1/23 (Figure 7): Γ-maximax incurs the lowest costs, followed by BMA; Γ-maximin incurs instead the highest costs. Interestingly the differences among the costs incurred by the various policies generally decrease with the size of the training set. For the case $t = 1/5$ (not shown) the costs of all policies are almost equivalent, lying close to 0.

It cannot be predicted whether deciding according to either Γ-maximin or Γ-maximax will eventually incur lower or higher total costs, for the prior-dependent instances, than deciding according to BMA. Our viewpoint is that on the prior-dependent instances taking a decision should be preferably avoided, trying instead to acquire new information.

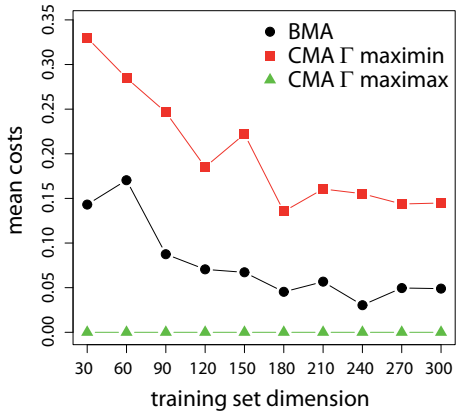

Figure 6: Mean costs incurred on the *prior-dependent* instances ($t = 1/2$).

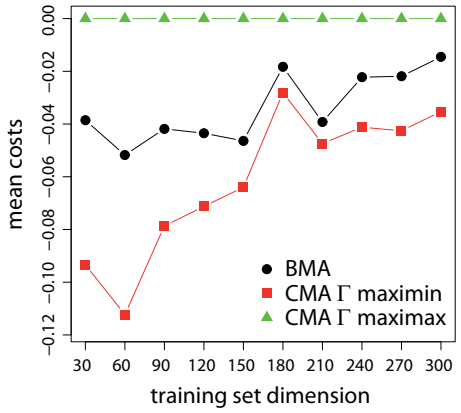

Figure 7: Mean costs incurred on the *prior-dependent* instances ($t = 1/23$).

6 Conclusions

CMA has proven effective on the real-world case study of predicting the presence of the Alpine marmot. Some future extensions can be foreseen. The first is adopting maximality rather than interval-dominance for detecting the prior-dependent instances; this should decrease the number of instances indeterminately classified without compromising the robustness of the classifications. Secondly, one could allow the prior probability of inclusion of each covariate to vary within a different interval; this would however imply solving a more complex optimization problem to detect the upper and lower bounds of the logit interval. Eventually the current algorithms could be extended to deal with more than two classes; for this purpose, the base classifiers to be ensembled should be polytomous (rather than dychotomous) logistic regressors.

Acknowledgments

The research in this paper has been partially supported by the Swiss NSF grants no. 200020-132252. The work has been performed during Andrea Mignatti's PhD, supported by Fondazione Lombardia per l'Ambiente (project SHARE- Stelvio). We thank the anonymous reviewers for their valuable insights.

References

[1] J. Abellán, R. Baker, F. Coolen, R. Crossman, and A. Masegosa. Classification with decision trees from a nonparametric predictive inference perspective. *Computational Statistics & Data Analysis*, in press, doi=10.1016/j.csda.2013.02.009, 2013.

[2] A Borgo. Habitat requirements of the Alpine marmot Marmota marmota in reintroduction areas of the Eastern Italian Alps. Formulation and validation of habitat suitability models. *Acta Theriologica*, 48(4):557–569, 2003.

[3] Kenneth P Burnham and David R Anderson. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, 2002.

[4] M. Cantini, C. Bianchi, N. Bovone, and D. Preatoni. Suitability study for the alpine marmot (marmota marmota marmota) reintroduction on the Grigne massif. *Hystrix, the Italian Journal of Mammalogy*, 9(1-2), 1997.

[5] M. Clyde and E. George. Model Uncertainty. *Statistical Science*, 19:81–94, 2004.

[6] G Corani and A Antonucci. Credal ensembles of classifiers. *Computational Statistics & Data Analysis*, in press, doi=10.1016/j.csda.2012.11.010, 2012.

[7] 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. Springer, 2012.

[8] G. Corani and M. Zaffalon. Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities. *Proc. ECML-PKDD 2008 (Eur. Conf. on Machine Learning and Knowledge Discovery in Databases)*, pages 257–271, 2008.

[9] J. Del Coz and A. Bahamonde. Learning nondeterministic classifiers. *The Journal of Machine Learning Research*, 10:2273–2293, 2009.

[10] Alfred DeMaris. A tutorial in logistic regression. *Journal of Marriage and the Family*, pages 956–968, 1995.

[11] S. Destercke. A k-nearest neighbours method based on lower previsions. In *Proc. IPMU 2010 (Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods)*, pages 129–138. Springer, 2010.

[12] C. Elkan. The foundations of cost-sensitive learning. *Proc. Int. Joint Conference on Artificial Intelligence (IJCAI - 01)*, pages 973–978, 2001.

[13] C. Fernandez, E. Ley, and M. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

[14] M. Guglielmin. Permaclim: a model for the distribution of mountain permafrost, based on climatic observations. *Geomorphology*, 51(4):245–257, April 2003.

[15] D. Hand. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Statistics in medicine*, 29(14):1502–10, 2010.

[16] Radu Herbei and Marten H Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721, 2006.

[17] D. Lenti Boero. Long-term dynamics of space and summer resource use in the alpine marmot (Marmota marmota L.). *Ethology Ecology & Evolution*, 15(4):309–327, 2003.

[18] Haim Levy and Harry M Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, 1979.

[19] E. Ley and M.F.J. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

[20] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 519–524. Morgan Kaufmann Publishers Inc., 2003.

[21] W.A. Link and R.J. Barker. Model weights and the foundations of multimodel inference. *Ecology*, 87(10):2626–2635, 2006.

[22] B.C. López, I. Figueroa, J. Pino, A. López, and D. Potrony. Potential distribution of the alpine marmot in Southern Pyrenees. *Ethology Ecology & Evolution*, 21(3-4):225–235, 2009.

[23] C. Perrin and D. Berre. Socio-spatial Organization and Activity Distribution of the Alpine Marmot Marmota marmota: Preliminary Results. *Ethology*, 93:21–30, 1993.

[24] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

[25] A.E. Raftery and D. Madigan. Bayesian model averaging for linear regression models. *Journal of the American Statistical*, 92(437):179– 191, 1997.

[26] S. J Riley, S.D. DeGloria, and R. Elliot. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of sciences*, 5(1-4):23–27, 1999.

[27] M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[28] B. Wintle, M. McCarthy, C. Volinsky, and R. Kavanagh. The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17(6):1579–1590, 2003.

[29] M. Zaffalon, G. Corani, and D. Maua. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282 – 1301, 2012.

[30] Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.