Credible classification for environmental problems

Marco Zaffalon [1]

IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland

Abstract

Classifiers that aim at doing credible predictions should rely on carefully elicited prior knowledge. Often this is not available so they should start learning from data in condition of near-ignorance. This paper shows empirically, on an agricultural data set, that established methods of classification do not always adhere to this principle. Traditional ways to represent prior ignorance are shown to have an overwhelming weight compared to the information in the data, producing overconfident predictions. This point is crucial for problems, such as environmental ones, where prior knowledge is often scarce and even the data may not be known precisely. Credal classification, and in particular the naive credal classifier, are proposed as more faithful ways to cope with the ignorance problem. With credal classification, conditions of ignorance may limit the power of the inferences, not the credibility of the predictions.

Key words: Credal classification; imprecise probabilities; naive credal classifier; imprecise Dirichlet model; agricultural data.

1 Introduction

Classification is one of the most important techniques for knowledge discovery in databases (Duda et al., 2001). It permits learning, from data, the relationship between a set of *attributes* (or *features*), describing an object, and the object's predefined *class*. Classifiers are used for predicting the unknown class of new objects, with applications that range from recognition to diagnosis, and forecasting. The methods, being largely independent of the domain, impact on nearly every field where a convenient database is available (see Section 2.1

[1] *E-mail address*: zaffalon@idsia.ch.

for a brief introduction to classification, and Section 2.4 for an introduction to empirical methods used in classification).

Severe limitations to applying classification arise when the database contains scarcely or vaguely informative data. This is the case of small and incomplete data sets (i.e., data sets with missing values), which are unfortunately a commonplace of real applications. In particular, Reichert (1997) raises some concerns that are very relevant to the present discussion, about the difficulty of modelling environmental problems (see also Kriegler and Held (2003)). He argues that many environmental problems are characterized both by vague prior knowledge and by imprecise knowledge of the data. In these conditions, there is the need of models capable of relying on weaker assumptions than common models (e.g., Bayesian models), because strong assumptions may severely bias the results, producing unreliable predictions. Reichert identifies such models with *imprecise probability* methods (Walley, 1991). Imprecise probability is a generic term for the many mathematical or statistical models which measure chance or uncertainty without sharp numerical probabilities. The present work uses sets of probability distributions (or *credal sets*, after Levi (1980)), a very general imprecise probability model.

This paper presents an empirical analysis of real agricultural data in Section 4. The machine learning objective is to qualitatively predict the grass grub quantity (grass grubs are one of the major insect pests of pasture in Canterbury, New Zealand) based on characteristics of the paddock and on farming practice. The data set (Section 3) contains 155 complete observations and its being small is shown to pose difficult problems for common machine learning techniques.

This work proposes the new paradigm of *credal classification* to obtain credible predictions even under such difficult conditions (Section 2.2). Credal classification is closely related to imprecise probability, being based on sets of probability distributions. Credal classifiers are more general than common classifiers in that an object can be assigned to more than one class: they recognize that the available knowledge may not justify the choice of a single class, and in this case they give rise to a set of alternative classes. In the experiments I used the naive credal classifier (NCC, see Zaffalon (2001, 2002b)), which extends the well-known naive Bayes classifier (NBC, see Duda and Hart (1973)) to credal sets (see Section 2.3). The NCC copes with small and incomplete data sets in a way that the classifications are robust to a wide set of unknown prior states of knowledge and to all the possible mechanisms responsible for the missing data. To date, the NCC is the only classifier with both these characteristics.

By empirically analyzing the results of the classification from several viewpoints, this work points out that the traditional prior assumptions are strong, and lead to unjustified conclusions for the presented case. It also shows that

the weaker requirements of the NCC provide more reasonable, though less determinate, answers when only little information is available. This evidence suggests that credal classifiers are more suitable to cope with domains where knowledge is imprecise.

2 Background

2.1 Classification

Each object under study in a classification problem is characterized by a vector of *attribute variables* $(A_1, \ldots, A_k)$ and by a *class variable* C. The generic variable A_i takes values in a set of *attributes* $\mathcal{A}_i$. C takes values in a set of classes $\mathcal{C}$. In this paper attribute variables are assumed to be *categorical* (or *discrete*), i.e. the sets $\mathcal{A}_i$ $(i = 1, \ldots, k)$ have finitely many elements. C must be categorical for the problem to be of classification. The purpose of attribute variables is to describe objects, while C serves the purpose of grouping (i.e., categorizing) objects.

To make this description more concrete, consider two possible applications. In a medical application, objects could be identified with patients. Their attributes would report information about the patient (such as age, gender, life style, etc.) and the results of medical tests. The patients could then be grouped according to the status of a given disease (e.g., "no disease", "moderate", "severe"). In the environmental domain, objects could be some type of plants under study, the vector of attributes describing characteristics of the plants (e.g., sepal/petal length and width), and the classes describing which plants are considered (e.g., iris setosa, iris versicolor, iris virginica).

Classification is related to the issue of *learning* (or *inference*, in statistical terms) in the following respect. In the typical classification setting, some objects are entirely known, which means that for each of them both the attributes and the class variables are in a known state. The problem is then to infer, from the known objects, what relates the attributes to the classes, in order to be able to predict the class of new objects for which only the attributes are known. The set of entirely known objects is called *learning set*. In the preceding example on plants, from observing the differences in sepal/petal length and width, across the classes in the learning set, one might infer a rule that allows a new plant to be placed in the right category (usually with a certain probability of error) only on the basis of its specific values of sepal/petal length and width. Similarly, in the medical example above, past examples of patients correctly diagnosed can be regarded as implicitly representing knowledge on the diagnostic process. This knowledge, once made explicit, could be used to diagnose

new patients.

Strated in a different way, classification methods are algorithms that take data (the learning set) in input and output a model of the relationship between the attributes and the classes. Such a model is called *classifier*: a classifier is a function that maps a vector of attributes to a class. There exist many possible classifiers, such as Bayesian models (of which Bayesian networks are an important special case), neural networks, support vector machines, classification trees, and many others (Duda et al., 2001).

Let us stress that the usefulness of classification and its wide application to a variety of real domains is basically due to the availability of methods to infer classifiers by looking only at the learning set. This is also what makes classification different from other scientific investigations, i.e. that the modelling of a phenomenon heavily relies on algorithms. This nevertheless, the role of the human analyst is very important. Prior to the inference, the analyst selects the attribute and the class variables, and preprocesses the data in order to prepare the learning set. At this stage the analyst also selects the type of classifier to be used, and incorporates possible prior knowledge on the phenomenon under study. After the inference, the analyst's tasks involve doing sensitivity analysis and model validation (usually by testing the model on new data). In many cases the analyst also decides to repeat the entire process by changing attributes, classifiers, and other parameters, until the final model produced is acceptable. The pre- and post-inference phases will be described in some detail in Section 2.4.

In the following the focus will be on the inferential part with respect to a special Bayesian model called naive Bayes classifier, and to its extension to sets of probability distributions.

2.2 Imprecise probabilities and credal classification

In some cases, the task of inferring a classifier must be achieved without substantial knowledge on a phenomenon, because data are the only source of information one has. In other cases, one may simply want to "let the data speak for themselves," because other forms of knowledge are too weak or vague to be useful, or because they are difficult to incorporate in the model.

Observe that in these conditions, being able to model ignorance in the most appropriate way becomes a fundamental issue. Basically, ignorance arises for two reasons. When data are the only source of information, one starts learning from them in a state of near-ignorance. Modelling *prior ignorance* is a long-standing and difficult problem. The Bayesian literature (Bernardo and Smith, 1996) proposes models based on so-called noninformative prior distributions

(or *priors*). This method, however, is very controversial (Walley, 1991, pp. 226–235). Many people use noninformative priors because the effects of a given prior are severe only for small samples and will disappear in the limit of an infinite sample, no matter which prior is used. However, data sets are often small. Furthermore, "small" and "large" are always relative to the sample space; also a data set with, say, 10^6 records (or *units*) is small if the sample space is sufficiently complex.

The second type of ignorance is related to incomplete samples. These are data sets in which some values have been turned into missing data. In such cases the mechanism responsible for the missing data should not be ignored if one aims to obtain credible conclusions, unless data are subject to a condition known as "missing at random" (Little and Rubin, 1987). Unfortunately, such condition cannot be tested statistically (Manski, 1993, pp. 73–74) and hence it does not seem to be suited when data are the only source of information. More generally speaking, missing data prevent us from having complete knowledge of the likelihood, i.e. there exists partial ignorance about the likelihood (or *likelihood ignorance*, for short). Of course, likelihood ignorance can produce effects for any size of the data set. Likelihood ignorance is a serious problem for knowledge discovery applications, among others, for which established methods do not provide a widely accepted solution.

Recently, there has been a great development of innovative proposals to model ignorance. **?** presents strong arguments that support the use of sets of prior densities to model prior ignorance. A number of contributions show that also likelihood ignorance can be modelled satisfactorily by sets of measures (Horowitz and Manski, 1998, 2001; Manski, 2003; Ramoni and Sebastiani, 2001b; Zaffalon, 2002a). The underlying idea of these modern approaches to prior and likelihood ignorance is the same: the body of all the possible states of knowledge *is* the model of ignorance. For example, all possible mechanisms responsible for the missing data, taken as a whole, are a model for likelihood ignorance. Overall, sets of probability distributions appear as a well-founded framework suited to model ignorance. Sets of probability distributions belong to the theory of *imprecise probabilities* (Walley, 1991) (see http://www.sipta.org for up-to-date information).

Two major consequences stem from adopting imprecise probabilities in classification problems. On the one hand, there is much greater modelling flexibility and realism. On the other, classifications can be partially indeterminate: an object is generally mapped to a set of classes. In general, the output classes should be all interpreted as candidates for the actual class, since there is no way to rank them. In other words, the different types of ignorance may limit the strength of the conclusions. Precisely determinate classifications, i.e. strong conclusions, are a special case achieved only when the conditions justify precision.

This more general way to address classification problems is called *credal classification*. Credal classification was introduced in Zaffalon (1999) and discussed more widely in Zaffalon (2002b). A *credal classifier* is defined as a function that maps a vector of attributes to a set of classes. A credal classifier is not only a new classifier, it implements a new way to perform classification.

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

2.3 The naive credal classifier

This section introduces the credal classifier called naive credal classifier. This is an extension of the well-known naive Bayes classifier to sets of probability distributions. For details about the model, please refer to Zaffalon (2001).

Definition of the model. As before, let C denote the classification variable and $(A_1, \ldots, A_k)$ denote the attribute variables. Let the classes and attributes be denoted by lowercase letters.

Let us assume the N units of the learning set, each with known values of the attributes and the class (for the moment, consider the case of complete data), are generated from an unknown multinomial process. 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 $n(c)$ and $n(a_i, c)$ be the observed frequencies of class c and of the joint state (a_i, c) in the N observations, respectively.

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:

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

Based on this assumption and imposing a Dirichlet prior over the chances, it is possible to obtain a Dirichlet posterior distribution:

$$P(\theta|\mathbf{n}) \propto \prod_{c \in \mathcal{C}} \left[\theta_c^{st(c)+n(c)-1} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{st(a_i,c)+n(a_i,c)-1} \right], \tag{2}$$

where $\mathbf{n}$ is the sample, $t(c)$ and $t(a_i, c)$ are hyperparameters corresponding to $n(c)$ and $n(a_i, c)$, respectively, and $s > 0$ is a constant representing the prior weight (also known as the number of virtual units).

So far I have presented a traditional Bayesian learning approach for the NBC. The extension to imprecise probabilities and the NCC is achieved by modelling prior ignorance by a set of Dirichlet prior densities. Consider the set of all Dirichlet priors, and, consequently, posteriors of the form (2), that are obtained by letting the t-hyperparameters vary in the following region:

$$\sum_{c} t(c) = 1 \tag{3}$$

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

$$t(a_i, c) > 0 \qquad \forall (i, a_i, c). \tag{5}$$

These constraints resemble the structural constraints to which the counts $n(c)$ and $n(a_i, c)$ naturally obey. The model obtained in this way is a special version of the *imprecise Dirichlet model* (**?**) and is coherent in the strong sense of Walley (1991, Section 7.8). In this framework s is interpreted as a degree of caution. Recall that the choice of the weight of the Bayesian prior is arbitrary, as it happens usually with Bayesian models. The NCC inherits this characteristics from the IDM, though Walley gives reasonable motivations to choose s in the interval $[1, 2]$. The effect of s on the NCC classifications follows easily under the interpretation of s as caution parameter: the larger s, the larger in general the output sets of classes for a given instance. Notably, the sets related to larger s's will always include those following from smaller ones.

Classification procedure. Let $E[U(c)|\mathbf{a}, \mathbf{n}, \mathbf{t}]$ denote the expected utility with respect to (2) from choosing class c, given $\mathbf{a}$, the previous data $\mathbf{n}$ and a vector $\mathbf{t}$ of hyperparameters. Since $\mathbf{t}$ belongs to a region, there are many such expected utilities for every class c, so that we cannot always compare

two classes: generally, there is a partial order on the classes that only allows us to discard the dominated ones. Indeed, the output of a credal classifier is the set of classes that are not dominated. Note that the partial order depends on the chosen dominance criterion. I use credal dominance, defined below.

The class c' is said to *credal-dominate* class 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 imprecise model.

Credal dominance is a special case of *strict preference* justified by Walley (1991, Sect. 3.7.7) on the basis of rationality (behavioral) arguments. It was previously proposed by Seidenfeld in the commentary of Kyburg (1983, p. 260, P-III′).

In the following 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 $E[U(c)|\mathbf{a}, \mathbf{n}, \mathbf{t}] = P(c|\mathbf{a}, \mathbf{n}, \mathbf{t})$. With the NCC, credal dominance reduces itself to the following nonlinear optimization problem:

$$\inf \left\{ \left[\frac{n(c'') + st(c'')}{n(c') + st(c')} \right]^{k-1} \prod_i \frac{n(a_i, c')}{n(a_i, c'') + st(c'')} \right\} \tag{6}$$

$$t(c') + t(c'') = 1 \tag{7}$$

$$t(c'), t(c'') > 0. \tag{8}$$

The optimum value of the problem is greater than 1 if and only if c' credal-dominates c''.

The extension to incomplete samples is straightforward. Let us limit ourselves to the common case when the class is never missing. Everything is unchanged in the classification procedure except that the counts $n(a_i, c')$ and $n(a_i, c'')$ in (6) are replaced by $\underline{n}(a_i, c')$ and $\overline{n}(a_i, c'')$, respectively. These are the minimum value of $n(a_i, c')$ and the maximum value of $n(a_i, c'')$ achieved by replacing the missing values of the attribute variable i with the values in $\mathcal{A}_i$ in all the possible ways.

Implementation of the naive credal classifier. This section is intended as a brief guide for interested readers to make their own implementation of the naive credal classifier. [2] Please refer to Zaffalon (2001) for proofs and further details.

The basic issue is the implementation of a procedure called "CD" that tests credal dominance by solving the optimization problem in (6)–(8). The procedure takes as input the bivariate counts [3] $n(a_i, c')$ and $n(a_i, c'')$ for each

[2] To date, there are no freely available implementations of the naive credal classifier.
[3] As obtained from the learning set.

$i = 1, \ldots, k$ ($k \geq 1$), and the counts $n(c')$ and $n(c'')$. Here a_i denotes the value of the i-th attribute for the particular instance to classify, and c' and c'' are two classes. The output of the procedure is a Boolean value indicating whether or not c' credal-dominates c''.

In order to show how to solve the optimization problem, let us first re-write it by the following notation, for short: $\alpha_i = n(a_i, c'), \beta_i = n(a_i, c''), \alpha = n(c'), \beta = n(c'')$ and $x = st(c'')$. The problem becomes:

$$\inf h(x) = \inf \left\{ \left[\frac{\beta + x}{\alpha + s - x} \right]^{k-1} \prod_i \frac{\alpha_i}{\beta_i + x} \right\} \tag{9}$$

$$0 < x < s. \tag{10}$$

The first and second logarithmic derivatives of $h(\cdot)$ are $(\ln h(x))' = \frac{k-1}{\beta+x} + \frac{k-1}{\alpha+s-x} - \sum_i \frac{1}{\beta_i+x}$, and $(\ln h(x))'' = -\frac{k-1}{(\beta+x)^2} + \frac{k-1}{(\alpha+s-x)^2} + \sum_i \frac{1}{(\beta_i+x)^2}$, respectively.

The following algorithm computes the global minimum of $h(x)$, subject to $0 < x < s$.

(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 $(\ln h(x))'|_{x=0} := -\infty$, else compute $(\ln h(x))'|_{x=0}$.
(3) Compute $(\ln h(x))'|_{x=s}$.
(4) If $(\ln h(x))'|_{x=0} \geq 0$, let $\inf h(x) := h(0)$. Stop.
(5) If $(\ln h(x))'|_{x=s} \leq 0$, let $\inf h(x) := h(s)$. Stop.
(6) If $(\ln h(x))'|_{x=0} < 0$ and $(\ln h(x))'|_{x=s} > 0$, approximate the minimum numerically. Stop.

On the basis of the value $\inf h(x)$ as computed by the algorithm above, the procedure "CD" outputs 0 if $\inf h(x) \leq 1$ and 1 in the opposite case.

With reference to the numerical approximation mentioned in point (6) above, I recall that $h(\cdot)$ is convex, whence any steepest descent algorithm will find its global minimum. One of the fastest options in this respect appears to be Newton-Raphson's method because the first and second derivatives are available. There is a problem of convergence with the basic Newton-Raphson algorithm, but the algorithms always converges when it is combined with bracketing, as in Press et al. (1993, p. 366). [4]

[4] 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 dominated in this case (this follows naturally by treating the zero of the machine as the actual

Now recall that the output of the NCC is the set of classes that are not dominated. To produce this set, it suffices to take the classes one by one, and test each of them by "CD" against all the others. If any of the latter classes dominates the former, this is discarded. The classes left at the end of this procedure are those that are not dominated. Note that "CD" is invoked $|\mathcal{C}|(|\mathcal{C}|-1)$ times at most. "CD" works in time linear in the number of attributes, so the overall computational complexity to credal-classify an instance is $O(k|\mathcal{C}|^2)$.

The extension to incomplete samples is straightforward by feeding the procedure "CD" with the lower and upper counts $\underline{n}(a_i, c')$ and $\overline{n}(a_i, c'')$, for each $i = 1, \ldots, k$, in the place of $n(a_i, c')$ and $n(a_i, c'')$, as mentioned in the preceding section. The frequencies $\underline{n}(a_i, c')$ and $\overline{n}(a_i, c'')$ can be computed in linear time in the size of the data set, so that the extension to incomplete samples does not increase the computational complexity to learn the NCC with respect to the NBC.

Comparison with other models. The NCC allows us to model prior and likelihood ignorance under very weak assumptions, so that the classifications are inherently robust to small sample sizes and missing data. These are benefits of the innovative ideas brought by credal classification. Credal classification is a promising field and new credal classifiers have already been proposed (Abellán and Moral, 2001, 2003; Zaffalon and Fagiuoli, 2003; Nivlet et al., 2001), although the NCC is still the only one that deals with both prior and likelihood ignorance. It is also important to mention the *robust Bayes classifier* from Ramoni and Sebastiani (2001a). Although developed independently, the NCC and the robust Bayes classifier share many characteristics. Both are generalizations of the NBC to sets of distributions, and the treatment of likelihood ignorance of the robust Bayes classifier is much in the same spirit of the NCC. However, the robust Bayes classifier uses the traditional model of non-informative priors for prior ignorance. This makes it generally overconfident when only small data sets are available, a problem that is more successfully addressed by the NCC implementation of the imprecise Dirichlet model (see Zaffalon et al. (2003) for a thorough comparison).

2.4 Experimental methods

Classification is focused on the core problem of inferring a classifier from a data set. However, the pre- and post-inference phases, mentioned in Section 2.1, also play a very important role so that the overall process is successful. Here I briefly summarize some main tasks that can be carried out in the two phases (I consider the case of prior ignorance). Please refer to Witten and Frank (1999) for more details.

zero).

Pre-inference phase. The purposes of the pre-inference phase are to prepare the learning data for the analysis and to select a classifier. The following list reports typical activities prior to the inference.

- Attribute selection. The data preparation involves the selection of the attribute variables. Often, all the attribute variables that are deemed to be relevant to predict the classes are initially included in the analysis. Some effort is usually done then, to select a subset of attribute variables that is best suited to predict the class variable. This activity is called *feature selection*. Feature selection employs different techniques to help the analyst to select the subset of features.
- Data cleaning. Real data often contain wrong information, duplicated units, inconsistencies, etc. The purpose of this activity is to correct the data as much as possible in order to produce a clean data set.
- Data transformation. It can be useful to transform the data, by changing measurement units, grouping values, applying mathematical functions, etc.
- Discretization. Some classifiers can only deal with categorical variables, therefore it is important to be able to turn continuous (or numerable) attributes to categorical ones, by a *discretization* method. A common choice in this respect is the entropy-based discretization of Kohavi et al. (1994).
- Choice of the classifier. There are several reasons that can favor the application of a classifier to a particular domain. Yet, it is often the case that there is no real preference and many classifiers are tested in order to select the best of them for the application under consideration.

Post-inference phase. After the classifier has been inferred, it should be validated. This is usually done empirically, by testing the classifier on new data (i.e. the so-called *test set*). The classes predicted by the classifier for the test set are compared with the actual ones, providing indexes of predictive performance. A common index is the *prediction accuracy*: i.e., the relative frequency of successfully predicted classes. Another method that tests the ability of a classifier to predict *probabilities* of classes is described in Section 4.3.

There are different empirical schemes to test the classifier on new data. A popular method is *tenfold cross-validation* (Kohavi, 1995). In this case $\mathcal{D}$ represents all the data we have; part of them will have to be used for learning, and the rest for testing. According to tenfold cross-validation, $\mathcal{D}$ is partitioned at random into ten subsets (the folds) $\mathcal{D}_1, \ldots, \mathcal{D}_{10}$ of approximately equal size. The classifier is inferred and tested ten 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 prediction accuracy, are computed each time on the test set $\mathcal{D}_t$, and are collected over the ten folds. In order to produce better estimates, tenfold cross validation is also repeated ten times and the results are averaged over the repetitions.

I only mention that the above methods can be elaborated in many ways, for example by considering costs on wrong classifications other than 0-1, by providing more robust estimates of performance via confidence intervals, or by comparing classifiers by statistical tests.

3 The data set

The agricultural data set used in this work was made publicly available by R. J. Townsend, from Lincoln, New Zealand (see the web page of Weka, [5] which is a free software for machine learning).

The data set describes the relationship between grass grub population and pasture damage levels, in order to provide objective estimates of the annual losses caused by grass grubs. Grass grubs can indeed cause severe pasture damage and economic loss. Grass grub populations are often influenced by biotic factors (diseases) and farming practices (such as irrigation and heavy rolling). The machine learning objective is to find a relationship between grass grub numbers, irrigation and damage ranking for the period between 1986 to 1992.

The data sets contains 155 complete instances. The attributes are the following (the possible values are in parentheses).

- Year_zone: the years of the period under consideration, divided into three zones, f, m, c (6f, 6m, ..., 2c).
- Year: the years of the period under consideration (86, 87, ..., 92).
- Strip: a strip of paddock sampled (integer).
- Pdk: a paddock sampled (integer).
- Damage_rankRJT: R. J. Townsend's damage ranking (0, 1, ..., 5).
- Damage_rankALL: other researchers' damage ranking (0, 1, ..., 5).
- Dry_or_irr: indicates if the paddock was dry or irrigated (d: "dryland", o: "irrigated overhead", b: "irrigated border dyke").
- Zone: position of the paddock (f: "foothills", m: "midplain", c: "coastal").
- GG_new: class variable, based on grass grubs per square metre (l: "low", a: "average", h: "high", v: "very high").

The empirical distribution of the classes is (0.316, 0.264, 0.297, 0.123), for "low", "average", "high" and "very high", respectively.

[5] http://www.cs.waikato.ac.nz/ml/weka/

4 Experimental analysis

The following sections present the experimental analysis carried out on the agricultural data set presented in Section 3.

4.1 Preprocessing

Initially the data set was discretized by the default entropy-based discretization utility of MLC++ (Kohavi et al., 1994). As a by-product of this step, the attributes "strip" and "pdk" were deemed irrelevant to predicting the class and were discarded. The remaining six attributes were tested for relevance in predicting the class by the feature selection option of Weka, and all of them were eventually kept for the analysis.

Subsequently, I tested seven classifiers on the data set. The classifiers were inferred and tested using Weka, with tenfold cross-validation. The classifiers involved in the comparison were: Decision Table, IB5 (an *instance-based* classifier), J48 (an implementation of Quinlan's *C4.5*), Naive Bayes, OneR (a *one-rule* classifier), PART (a *rule-induction* classifier) and SMO (an implementation of *support vector machines*). These were all used with default options. (See Witten and Frank (1999) for a thorough description of the above classifiers.)

Table 1
The cross-validated prediction accuracy for several classifiers available in Weka on the grass grub data. The accuracies are given as percentages ± their standard deviations. The NBC achieves the best performance.

Classifier	Accuracy %
Decision Table	40.00 ± 3.93
IB5	45.16 ± 3.99
J48	42.58 ± 3.97
Naive Bayes	**49.03** ± 4.01
OneR	45.16 ± 3.99
PART	36.77 ± 3.87
SMO	40.64 ± 3.94

Table shows 1 the result of the comparison. It appears that the data set carries only limited information about the domain. Indeed all the classifiers do not capture strong relationships between the attributes and the class. However, some predictions are significantly higher than what the simple majority rule

achieves (i.e., 31.6%). In particular, the NBC appears to be a good candidate method for the data set.

4.2 NBC vs NCC

From now on, I focus on the NBC and its extension to credal sets, the NCC (with caution parameter s=1). The following discussion will show that, despite the good performance, the NBC makes random predictions for a large fraction of the instances. This is due to the overwhelming weight of the precise prior distribution over the knowledge carried by the data, which makes the NBC overconfident. In contrast, the NCC starts from much weaker assumptions, and is able to suspend the judgment on the instances for which the information in the data does not allow strong conclusions to be drawn.

I ran tenfold cross-validation using both the NCC and the NBC. The NCC produced a precise classification (i.e., a single class) for about 60% of the 155 instances, with an accuracy 52.01%=:C_1. In the remaining 40%=:S, it produced 2.36=:Z classes, on average, out of the possible 4. This set of classes contained the actual class with probability 0.82=:Cs. The most relevant output here is S: the NCC states that in about 40% of the instances, the available knowledge is not sufficient to produce a single class, but only a set of alternative classes.

Table 2
Experimental results for the NBC. Each row reports the result for an NBC inferred according to a different noninformative priors. The columns report per cent accuracies.

	N	Ns	Rs
Perks	48.21	42.74	44.47
Uniform	48.83	44.24	44.47
Jeffreys	48.58	43.65	44.47

Table 2 reports the results related to the NBC. Each row refers to an NBC inferred according to a different prior. There are three cases, according to three well-known proposals to model prior ignorance within the precise probability framework. These are the Perks (Perks, 1947), Uniform (Laplace, 1812) and Jeffreys priors (Jeffreys, 1983). The column N is the accuracy of the NBCs on the entire test set. Ns is the accuracy of the NBCs on the subset of instances (S) for which the NCC produces more than one class. Finally, Rs is the prediction accuracy of a random predictor on the same subset of instances (S). The random guesser randomly chooses one of the classes in the subset of classes produced by the NCC.

The comparison of Ns and Rs shows that every NBC is simply doing random predictions on the subset related to S, their performance being almost identical. This is an empirical proof that the NCC is correct in partially suspending the judgment on such instances. In fact, the NBC is overconfident, in a way that its predictions are not reliable in a large fraction of cases (40%). This fact is hidden when only the overall prediction accuracy of the NBC is considered.

We can appreciate the behavior of the NCC by also noting that the NCC isolates a subset of instances in which robust predictions are possible (C_1). Also, instead of predicting at random on the remaining instances, the NCC produces a set of classes (Z) with a high probability (Cs) of including the actual class: in other words, we can be confident that the discarded classes have low chance of containing the actual one.

4.3 A deeper view

Now I analyze the behavior of the NCC and the NBC from another angle. Let us consider the process of sequential learning, as explained at the end of Section 2.2. When learning sequentially, there is initially very little knowledge available to make reliable determinate predictions. It is therefore interesting to compare the behaviors of the NBC (the uniform prior is used for the experiments below) and the NCC.

Table 3 reports the results of the experiment on the first 15 instances of the data set. The first column reports the instance number. The second column reports the actual class of the instance. The column "NBC" shows the classes produced by the NBC, i.e. all the classes with maximum posterior probability for a given instance. The next column contains the posterior probability that the NBC assigns to the actual class (the probabilities are displayed with an approximation at the second decimal digit). The column "loss" reports the *logarithmic score* (measured in bits) related to the NBC on the instance, i.e. the negated logarithm in base 2 of the probability in the fourth column. The NCC column reports the classes produced by the NCC. Finally, the last column reports the lower and the upper posterior probabilities assigned by the NCC to the actual class (these probabilities have been approximated numerically).

Cowell et al. (1993) propose the logarithmic scoring rule as a way to evaluate and compare classifiers based on the probability that they assign to the actual class. The higher the probability, the smaller the loss, with the limit of zero loss when the class is judged to be certain. By this rule, it is easy to see that the NBC produces very unreliable predictions and consequently large losses for the examined cases. For example, the second instance produces a loss of 4.64 bits since the NBC deems that the actual class "high" should only appear

Table 3
Results of the sequential learning on the first 15 instances of the data set.

#	c	NBC	$P(c\|a)$	loss	NCC	$\underline{P}(c\|a)$, $\overline{P}$ $(c\|a)$
1	l	lahv	0.25	2.00	lahv	0.00,1.00
2	h	l	0.04	4.64	lahv	0.00,1.00
3	h	l	0.31	1.67	lahv	0.00,1.00
4	h	h	0.75	0.41	lahv	0.06,0.94
5	l	h	0.05	4.32	lahv	0.00,0.63
6	l	h	0.20	2.33	lahv	0.00,0.68
7	h	lh	0.49	1.02	lahv	0.00,1.00
8	l	h	0.30	1.76	lahv	0.14,0.67
9	a	h	0.02	5.51	lahv	0.00,1.00
10	a	a	0.53	0.92	lahv	0.00,1.00
11	l	a	0.30	1.75	lahv	0.00,1.00
12	h	l	0.32	1.64	lahv	0.00,1.00
13	h	h	0.40	1.33	lahv	0.00,1.00
14	h	h	0.75	0.42	lahv	0.00,1.00
15	v	h	0.00	8.38	ahv	0.00,0.96

4 times out of 100. Units 5 and 9 present similar situations. The last unit is even worse, with a loss of 8.38 bits, i.e. the actual class should appear 3 times out of 1000.

As far as the NCC is concerned, we see that it suspends the judgment for all the instances except for the last one, where the amount of past examples starts turning total indeterminacy (i.e., when all the classes are possible alternatives) into partial indeterminacy. In fact the class "low" is not considered plausible for the last instance. By this behavior, the NCC informs us that the knowledge available in the data does not allow us to make any reliable prediction in the first 14 instances, and only a weak prediction for the last one. This appears to be a very reasonable way to act when the information is very scarce, as is certainly more credible than giving strong judgments, not justified by the evidence. We can also note that the uncertainty about the actual class is confirmed by the large, sometimes complete, indeterminacy (i.e., the difference between the upper and the lower probability) of the intervals in the last column.

Similarly to the preceding section, we can show that the empirical evidence supports the behavior of the NCC by showing that the NBC acts as a random

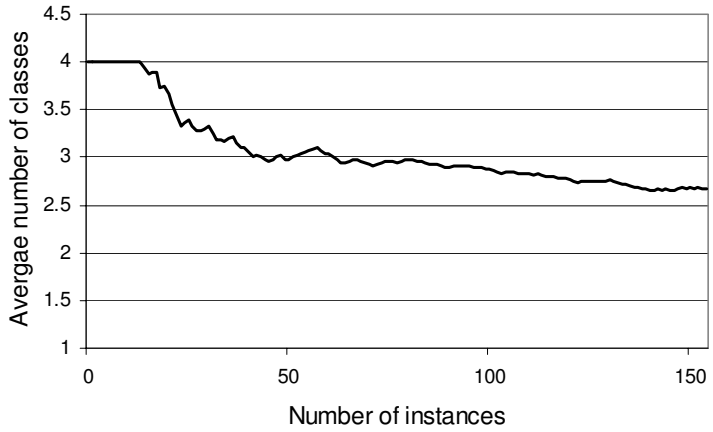

Fig. 1. The average number of classes produced by the NCC as a function of the number of instances used to infer the classifier.

predictor. Cowell et al. (1993, Section III.A) suggest evaluating a classifier by comparing it with an alternative predicting system. As alternative system, I used a random guesser, i.e. the classifier that each time assigns uniform probability to the classes, irrespectively of the attribute values. This, assigning probability 0.25 to the actual class, produces a loss of 2 bits for each instance, with an overall loss of 30 bits. By summing the losses in the fifth column of Table 3, we see that the total loss of the NBC is 38.09. The NBC predicts probabilities even worse than the random predictor.

Finally, we can have an idea of how credal classification works in the rest of cases by examining Figure 1. This reports the average number of classes produced by the NCC as a function of the number of available instances in the sequential learning. Initially, when the NCC is fed with very few past examples (as in Table 3), the output indeterminacy is very high: the NCC tends to produce completely indeterminate classifications (4 classes). As more data accumulate, the average number of classes decreases. This value is close to 2.5 when all the instances have been read. By reading more data, the average would tend to 1.

5 Conclusions

In his recent book, Manski (2003, p. 1) states the *law of decreasing credibility*: "the credibility of inference decreases with the strength of the assumptions maintained." There appears to be a wide agreement on this statement, but for it to be put to practice, we need models able to produce results under very weak assumptions. Imprecise probability methods are good candidates in this

respect. Walley (1991) developed a very general theory of uncertainty based on rationality arguments (much like the Bayesian theory), though relaxing the assumptions that probabilities have to be known precisely. Relaxing the assumption of precision is the key to allow weak assumptions to be used.

Indeed, one advantage of permitting imprecision in probability is that states of ignorance can be modelled very faithfully. This characteristic appears to be particularly important in the environmental domain, where knowledge on a phenomenon can be highly vague. With regard to classification, this means that prior and likelihood ignorance can be frequent conditions, and the present paper shows that this is true for a specific example of agricultural data.

Credal classifiers deal with prior and likelihood ignorance by incorporating ignorance in the model, using imprecise probabilities, so that they can be more credible models for environmental problems. In the application under study, the naive credal classifier is shown to provide reliable predictions also with a small learning set. Reliability is maintained by weakening the predictions (i.e., by providing set-based classifications) in the most difficult cases. This is a logical consequence of the poor knowledge jointly available from the learning data and the chosen assumptions. Of course we should aim at having deep knowledge of a phenomenon, as this would make us draw stronger conclusions. But this is not always possible. In these case, it is important to let scarce knowledge show us what its logical implications are.

Acknowledgments

I would like to thank Peter Walley for enlightening discussions and for encouraging me to develop credal classification. This research was partially supported by the NSF grant 2100-067961.02.

References

Abellán, J., Moral, S., 2001. Building classification trees using the total uncertainty criterion. In: de Cooman et al. (2001). pp. 1–8.

Abellán, J., Moral, S., 2003. Maximum of entropy in credal classification. In: Bernard et al. (2003). pp. 1–15.

Bernard, J.-M., Seidenfeld, T., Zaffalon, M. (Eds.), 2003. ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications. Carleton Scientific, Canada, Proceedings in Informatics 18.

Bernardo, J. M., Smith, A. F. M., 1996. Bayesian Theory. Wiley, New York.

Cowell, R. G., Dawid, A. P., Spiegelhalter, D., 1993. Sequential model criticism in probabilistic expert systems. PAMI 15 (3), 209–219.

de Cooman, G., Cozman, F. G., Moral, S., Walley, P. (Eds.), 1999. ISIPTA '99: Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications. The Imprecise Probability Project, Universiteit Gent, Belgium.

de Cooman, G., Fine, T. L., Seidenfeld, T. (Eds.), 2001. ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications. Shaker, The Netherlands.

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

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification. Wiley, 2nd edition.

Horowitz, J. L., Manski, C. F., 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. Journal of Econometrics 84, 37–58.

Horowitz, J. L., Manski, C. F., 2001. Imprecise identification from incomplete data. In: de Cooman et al. (2001). pp. 213–218.

Jeffreys, H., 1983. Theory of Probability. Clarendon Press, Oxford, 3rd edition.

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

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

Kriegler, E., Held, H., 2003. Climate projections for the 21st century using random sets. In: Bernard et al. (2003). pp. 345–360.

Kyburg, H. E. J., 1983. Rational belief. The behavioral and brain sciences 6, 231–273.

Laplace, d. P. S., 1812. Théorie Analytique des Probabilités. Courcier, Paris.

Levi, I., 1980. The Enterprise of Knowledge. MIT Press, London.

Little, R. J. A., Rubin, D. B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Manski, C., 1993. The selection problem in Econometrics and Statistics. In: Rao, C. R., Maddala, G. S., Vinod, H. (Eds.), Handbook of Statistics, Vol. 11: Econometrics. North-Holland, Amsterdam, pp. 73–84.

Manski, C. F., 2003. Partial Identification of Probability Distributions. Springer-Verlag, New York.

Nivlet, P., Fournier, F., Royer, J.-J., 2001. Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition. In: de Cooman et al. (2001). pp. 284–292.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1993. Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univer-

sity Press, Cambridge, 2nd edition.
Ramoni, M., Sebastiani, P., 2001a. Robust Bayes classifiers. Artificial Intelligence 125 (1–2), 209–226.
Ramoni, M., Sebastiani, P., 2001b. Robust learning with missing data. Machine Learning 45 (2), 147–170.
Reichert, P., 1997. On the necessity of using imprecise probabilities for modelling environmental systems. Water Science and Technology 36 (5), 149–156.
Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York.
Witten, I. H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.
Zaffalon, M., 1999. A credal approach to naive classification. In: de Cooman et al. (1999). pp. 405–414.
Zaffalon, M., 2001. Statistical inference of the naive credal classifier. In: de Cooman et al. (2001). pp. 384–393.
Zaffalon, M., 2002a. Exact credal treatment of missing data. Journal of Statistical Planning and Inference 105 (1), 105–122.
Zaffalon, M., 2002b. The naive credal classifier. Journal of Statistical Planning and Inference 105 (1), 5–21.
Zaffalon, M., Fagiuoli, E., 2003. Tree-based credal networks for classsification. Reliable Computing 9 (6), 487–509.
Zaffalon, M., Wesnes, K., Petrini, O., 2003. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. Artificial Intelligence in Medicine 29 (1–2), 61–79.