Credal Classification for Mining Environmental Data

M. Zaffalon [a]

[a]IDSIA
Galleria 2
6928 Manno, Switzerland
zaffalon@idsia.ch

Abstract: Classifiers that aim at doing reliable predictions should rely on carefully elicited prior knowledge. Often this is not available so they should be able to start learning from data in condition of prior ignorance. This paper shows empirically, on an agricultural data set, that established methods of classification do not always adhere to this principle. Common 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 represent states of ignorance. We show that with credal classification, conditions of ignorance may limit the power of the inferences, not the reliability of the predictions.

Keywords: Credal classification; imprecise probability; naive credal classifier; imprecise Dirichlet model; agricultural data.

1 INTRODUCTION

Classification is a very important methodology for knowledge discovery in databases (see Duda et al. [2001]). It permits learning, from data alone, the relationship between an object, described by a set of *features*, and its pre-defined *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.

Severe limitations to applying classification arise when the database contains scarcely or vaguely informative data. This is the case of small or 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 modeling environmental problems. He argues that 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 every strong assumption may severely bias the results, producing unreliable predictions. Reichert identifies such models with *imprecise probability* methods (Walley [1991]). 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 publicly available real agricultural data. 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 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 reliable predictions even under such difficult conditions. Credal classification is closely related to imprecise probability, being based on sets of probability distributions. Credal classifiers are more general than common classifiers because 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 they give rise to a set of alternative classes. In the exper-

iments we use the naive credal classifier (NCC, see Zaffalon [2001, 2002b]), which extends the well-known naive Bayes classifier (NBC) to credal sets. The NCC copes with small or incomplete data sets in such a way that the classifications are robust to all the possible unknown prior states of knowledge and to all the possible mechanisms generating the missingness. To date, the NCC is the only classifier with these characteristics.

By analyzing the results of the classification under several viewpoints, this work clearly shows that the prior assumptions needed by Bayesian models lead to unjustified conclusions for the presented case. It also shows that the weaker requirements of the NCC provide more reasonable, though less precise, answers also when only little information is available. The evidence suggests that credal classifiers are more apt to cope with domains where knowledge is possibly imprecise.

2 METHODS

2.1 Credal Classification

A classifier is an algorithm that allocates new objects to one out of a finite set of previously defined groups (or classes) on the basis of observations on several characteristics of the objects, called attributes or features (see Duda et al. [2001]). Credal classification, introduced by Zaffalon [1999] and discussed more widely in Zaffalon [2002b], sustains the viewpoint according to which a more general framework than common classification is needed to tackle real data sets, which may be small or incomplete. A credal classifier is defined as a function that maps an object into a set of classes. Credal classification assumes that the knowledge hidden in data does not always allow the classes to be completely ranked: in this case, only the classes that are dominated in the partial order can be discarded from consideration, producing the set of *undominated* classes as output (the more knowledge, the narrower the set in general).

Credal classification is closely related to the theory of *imprecise probabilities* (see Walley [1991]) in that it does not require that probability values be precise: they can be intervals or, more generally, uncertainty can be modeled by a set of distributions. For example, Zaffalon [2002a] shows that when we do not assume anything about the mechanism generating the missingness, missing data give rise to a set of possible distributions modeling the domain. From this, it is immediate to obtain a credal classifier that is robust to every possible mechanism of missingness, i.e. to all the possible replacements of missing data with known values.

Credal classification is a promising field and new proposals have already been advanced to build credal classifiers (see Fagiuoli and Zaffalon [2000]; Abellán and Moral [2001]; Nivlet et al. [2001]), although the NCC is still the only one that models both prior ignorance and vagueness due to missing data.

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

The naive credal classifier used in this paper is an extension of the discrete naive Bayes classifier (see Duda and Hart [1973]) to sets of distributions, with which it shares the assumption that the attributes are mutually independent conditional on the class. The NBC is learnt by inferring a distribution $P(C, \mathbf{A})$ from data, by which the NBC classifies a new vector of attributes by choosing the class $c^* = \arg\max_{c \in \mathcal{C}} P(c|\mathbf{a})$. In order to understand the classification procedure of the NCC, we must consider the way in which it is learnt from data (for details, refer to Zaffalon [2001]).

The NCC assumes that data are generated by a multinomial process, prior to the intervention of the missingness mechanism. The NCC is inferred from a complete sample by a special version of the *imprecise Dirichlet model* proposed by Walley [1996]. This models prior ignorance about the chances of the multinomial distribution by a set of Dirichlet densities of equal weight s, a parameter usually chosen in the real interval $[1, 2]$ and interpreted as a degree of caution of the inferences. Posterior inferences are obtained by combining the prior densities with the observed likelihood function; the resulting model is coherent in the strong sense of Walley [1991] (Section 7.8).

The posterior densities produce a set $\mathcal{P}$ of possible distributions $P(C, \mathbf{A})$. Given a new object, a class c is in the related output set of classes if there is no class $c' \in \mathcal{C}$ so that $P(c|\mathbf{a}) < P(c'|\mathbf{a})$ for each $P \in \mathcal{P}$. In this case c is said to be *credally undominated*. Learning from incomplete samples is achieved by applying the above method considering all the likelihood functions originated by replacing the missing data with known values in all the possi-

ble ways. $\mathcal{P}$ accounts now both for the imprecision due to the prior ignorance about the multinomial chances and for the imprecision originated by the ignorance concerning the missingness mechanism. Let us stress that using the NCC involves no approximations, that inferring the NCC and the NBC has the same computational complexity, and that the NCC classification complexity is $O(k\,|\mathcal{C}|^2)$.

2.2 The Data Set

The agricultural data set used in this analysis was donated by R. J. Townsend, from Lincoln, New Zealand. It is publicly available through the web page of Weka (http://www.cs.waikato.ac.nz/ml/weka/), 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 (low, average, high, veryhigh).

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

3 Experimental Analyses

The first step of the analysis was discarding the attributes "strip" and "pdk" from consideration. These were deemed irrelevant to predicting the class, by the discretization utility of MLC++ (see Kohavi et al. [1994]) used, with default options, to the extent of converting them to nominal attributes.

3.1 The NBC Performs Well

We firstly show that the NBC performs well with this data set, compared to other classifiers. We ran seven classifiers in Weka, using the empirical scheme of 10-folds cross-validation (see Kohavi [1995]) to evaluate their *prediction accuracy* (i.e., the relative number of correct predictions) on unseen data. The classifiers involved in the comparison are: 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 are all used with default options. (See Witten and Frank [1999] for a thorough description of the above classifiers.) From the comparison in Table 1, it appears that the data set carries only limited information about the domain. Indeed all the classifiers do not capture

Table 1: The cross-validated prediction accuracy for several classifiers available in Weka on the grass grub data. The accuracies are given as percentages $\pm$ 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	$\mathbf{49.03} \pm 4.01$
OneR	45.16 ± 3.99
PART	36.77 ± 3.87
SMO	40.64 ± 3.94

strong relationships between the features 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 the best candidate for the data set.

3.2 NBC vs NCC

From now on, we focus on the NBC and its extension to credal sets, the NCC (with caution parameter s=1). We 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 renders the NBC overconfident. In contrast, the NCC, being able to model prior ignorance, starts from much weaker assumptions, and is able to suspend the judgment on the instances for which the information does not allow strong conclusions to be drawn.

We ran 10-folds cross-validation using both the NCC and the NBC. The NCC produced a precise classification (i.e., a single class) for about the 59.55% of the 155 instances, with an accuracy C_1=52.01%. In the remaining 40.45% (S) of instances, it produced an average of Z=2.36 classes 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 on about 40% of the instances, the available knowledge is not sufficient to produce a single class, but only a set of possible alternative classes.

Table 2: Experimental results for the NBC. Each row reports the result for an NBC inferred according to a different "noninformative" prior distribution. The columns report percentage 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 distribution. There are three cases, according to three well-known proposals to model prior ignorance within the precise probability framework. These are the Perks, Uniform and Jeffreys priors (see Zaffalon [2001] for details). 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 facts, the NBC is overconfident, in a way that its predictions are not reliable on 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 on which robust predictions are possible (C_1). Also, instead of predicting at random on the remaining instances, the NCC produces a set of classes 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 being true.

3.3 A Deeper View

Now we analyze the behavior of the NCC and the NBC from another angle. Let us consider the process of sequential learning: the data set is read instance by instance; each time the new instance is classified and then, together with its actual class, it is used to incrementally update the classifiers' knowledge. This is a very natural learning process when the data to be classified are available sequentially.

When learning sequentially, there is initially very little knowledge available to make reliable 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. (Note the short notation for the classes, 'l'=low, 'a'=average, etc.) 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 col-

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

umn 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 columns "loss" reports the *logarithmic score* related to the NBC on the instance, i.e. the negated logarithm in base 2 of the probability in the fourth column, measured in *bits*. 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.

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 "low" should only appear 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 reliable 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 predictor. Cowell et al. [1993] (Section III.A) suggest evaluating a classifier through the logarithmic scoring rule by comparing it with an alternative predicting system. As alternative system, we choose 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, it would quickly tend to 1.

4 CONCLUSIONS

Real domains are very often subject to imprecise prior knowledge and imprecision in the available data. Environmental problems seem to be in such a category, and the present paper shows that this is true for a specific example of agricultural problem.

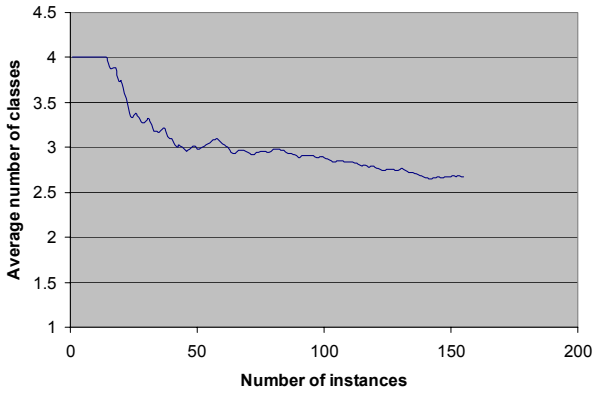

Figure 1: The average number of classes produced by the NCC as a function of the number of instances used to infer the classifier.

The evidence suggests that we should use tools able to cope with imprecision in a reliable way, instead of producing stronger conclusions than possible. Imprecise probability is identified as the mathematical framework to model imprecision in a sound way. In particular, real applications can benefit from the inherent reliability of credal classification also under severe conditions of scarce information.

ACKNOWLEDGMENTS

I would like to thank Peter Walley for enlightening discussions and for encouraging me to develop credal classification.

REFERENCES

Abellán, J. and S. Moral. Building classification trees using the total uncertainty criterion. In de Cooman, G., Fine, T., and Seidenfeld, T., editors, *ISIPTA'01*, pages 1–8, The Netherlands, 2001. Shaker Publishing.

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

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

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

Fagiuoli, E. and M. Zaffalon. Tree-augmented naive credal classifiers. In *IPMU 2000: Proceedings of the 8th Information Processing and Management of Uncertainty in Knowledge-Based Systems Conference*, pages 1320–1327, Spain, 2000. Universidad Politécnica de Madrid.

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

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

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

Nivlet, P., F. Fournier, and J.-J. Royer. Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition. In de Cooman, G., Fine, T., and Seidenfeld, T., editors, *ISIPTA'01*, pages 284–292, The Netherlands, 2001. Shaker Publishing.

Reichert, P. On the necessity of using imprecise probabilities for modelling environmental systems. *Water Science and Technology*, 36(5):149–156, 1997.

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

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

Witten, I. H. and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

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

Zaffalon, M. Statistical inference of the naive credal classifier. In de Cooman, G., Fine, T., and Seidenfeld, T., editors, *ISIPTA'01*, pages 384–393, The Netherlands, 2001. Shaker Publishing.

Zaffalon, M. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002a.

Zaffalon, M. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002b.