Restricting the IDM for classification

G. Corani and A. Benavoli

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
CH-6928 Manno (Lugano), Switzerland,
giorgio@idsia.ch, alessio@idsia.ch

Abstract. The *naive credal classifier* (NCC) extends *naive Bayes classifier* (NBC) to imprecise probabilities to robustly deal with the specification of the prior; NCC models a state of ignorance by using a *set* of priors, which is formalized by Walley's *Imprecise Dirichlet Model* (IDM). NCC has been shown to return more robust classification than NBC. However, there are particular situations (which we precisely characterize in the paper) under which the extreme densities included by the IDM force NCC to become very indeterminate, although NBC is able to issue accurately classifications. In this paper, we propose two approaches which overcome this issue, by restricting the set of priors of the IDM . We analyze both approaches theoretically and experimentally.

1 Introduction

The *naive Bayes classifier* (NBC) is often accurate, despite the unrealistic assumption of independence of the features given the class. However, especially on small data sets, NBC can happen to issue *prior-dependent* classifications, i.e., the most probable class varies depending on the adopted prior. This is acceptable if the prior can be carefully elicited to model domain knowledge; otherwise, prior-dependent classifications can be fragile. Usually, NBC is learned using a uniform prior, in the attempt of being *non-informative*. Yet, this solution is hardly satisfactory because the uniform prior models *indifference* rather than ignorance and anyway the choice of any single prior implies some arbitrariness.
The *naive credal classifier* (NCC) extends NBC to imprecise probabilities to robustly deal with the specification of the prior density; NCC models a state of ignorance by using a *set* of priors, which is formalized by Walley's *Imprecise Dirichlet Model* (IDM) (see [2] for a tutorial). IDM satisfies several properties desirable to model prior ignorance, such as the *representation invariance principle* (RIP) and the *likelihood principle* (LP) [2].
NCC turns the set of priors into a set of posteriors by element-wise application of Bayes rule; eventually, it returns all the classes which are *non-dominated*[1] within the set of posteriors. In fact, NCC returns a set of classes when faced with instances whose classification is prior-dependent; it issues weaker but more robust classifications than NBC. We call *determinate* the classifications made of

[1] The definition of dominance is given in Section 2.

a single class, and *indeterminate* the others.
Yet, there are particular situations (which we precisely characterize in the paper) under which the extreme densities included by the IDM force NCC to become very indeterminate; as pointed out in [7, Sec. 6], this behavior is correct in principle, since it shows that the classifications issued by NBC are prior dependent. Yet in such cases the large indeterminacy of NCC is questionable, as it is mostly due to extreme (namely, very skewed) priors; in fact, NCC becomes more determinate if such extreme priors are removed. Moreover, in such cases NBC (learned with uniform prior) achieves good accuracy on the instances indeterminately classified by NCC, which further shows an excessive indeterminacy of NCC. A way to increase the determinacy of NCC in such cases is to remove these extreme densities by restricting the IDM's set of priors by a small amount (an ϵ), in order to remove the boundary. In this paper we propose two approaches to cut off the extreme densities in the IDM; in both cases, the amount of densities removed from IDM is controlled by the parameter $\epsilon > 0$. The value of ϵ determines a trade-off between robustness and informativeness of the issued classifications: increasing ϵ increases informativeness, at a cost of some robustness. The setting $\epsilon = 0$ corresponds to the IDM, which is maximally robust but, at least in such particular cases, leads to a questionable high indeterminacy.

An alternative approach for restricting the credal set by modelling domain knowledge is given in [1], where only those priors that guarantee an improvement of the Mean Squared Error over the Maximum Likelihood Estimator are included in the credal set. In [1] it is also shown that removing extreme densities from the IDM is equivalent to express preferences among subregions of the parameter space; from this viewpoint, the two approaches proposed in this paper are informative; thus, as we show in Sec.2.1, they cannot satisfy at the same time both RIP and LP.

By experiments, we show under which conditions NCC, learned with Walley's IDM, can become unnecessarily indeterminate; we compare its behavior against that of an alternative credal classifier, CMA (credal model averaging) [3]. Then, we show that NCC becomes more determinate without compromising its reliability, when the two approaches for restricting the IDM are applied.

2 Naive credal classifier

NCC models prior near-ignorance by a set of priors; the set is formally defined by using Walley's Imprecise Dirichlet Model (IDM) [6]. NCC updates each prior with the observed likelihood, via element-wise application of Bayes' rule; in this way, NCC turns the set of priors into a set of posteriors. 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 have k features $F_1, \ldots, F_k$ taking generic values $[f_1, \ldots, f_k] = \mathbf{f}$ from the sets $\mathcal{F}_1, \ldots, \mathcal{F}_k$; the features are assumed to be *discrete*. We denote by $\theta_{c,\mathbf{f}}$ the real unknown probability (*chance*) that $(C, F_1, \ldots, F_k)$ equals $(c, \mathbf{f})$, by $\theta_{f_i|c}$ the chance that $F_i = f_i$ conditional on c and by $\theta_{\mathbf{f}|c}$ the chance of $(f_1, \ldots, f_k)$ conditional on c. Let N be the total number of samples; let $n(c)$ and $n(f_i\,|c)$ be the observed frequencies of class c and of $(f_i\,|c)$. NCC, like NBC, (naively) assumes the independence of the attributes given the class $\theta_{\mathbf{f}|c} = \prod_{i=1}^{k} \theta_{f_i\,|c}$. The likelihood function is:

$$L(\mathbf{n}|\theta) \propto \prod_{c\in\mathcal{C}} \left[\theta_c^{n(c)} \prod_{i=1}^{k} \prod_{f_i\in\mathcal{F}_i} \theta_{f_i|c}^{n(f_i|c)} \right], \tag{1}$$

where $\mathbf{n}$ denotes the vector of all the above frequencies. Observe that for all c and i, the observations satisfy the *structural constraints* $0 \leq n(f_i\,|c) \leq n(c)$, $\sum_c n(c) = N$ and $\sum_{f_i\in\mathcal{F}_i} n(f_i\,|c) = n(c)$. The prior density is expressed similarly to the likelihood function, except that frequencies $n(\cdot)$ are replaced everywhere by $st(\cdot) - 1$, i.e., the prior is a Dirichlet density with parameters $\alpha(\cdot) = st(\cdot)$. The parameter s is a positive real number which can be regarded as the number of *hidden samples*, in the common interpretation of conjugate Bayesian priors as additional sample units (the number can be fractional, though); the parameters $t(\cdot)$ can be regarded as the proportion of units of the given type; for instance, $t_{c'}$ is the proportion of hidden units having class c' in the hidden samples. Theoretical considerations suggest that s should lie between 1 and 2 [2], while the $t(\cdot)$ are usually set according to the *uniform* prior: $t(c) = \frac{1}{|\mathcal{C}|}$ and $t(a_i|c) = \frac{1}{|\mathcal{C}||\mathcal{F}|}$. By multiplying the prior density and the likelihood function, we obtain a posterior density of the same form as the likelihood, with $n(\cdot)$ replaced by $st(\cdot) + n(\cdot) - 1$. We estimate the posterior joint probability of class and features by taking expectation over the posterior probability of θ, i.e., $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ equal to:

$$P(c|\mathbf{n}, s, \mathbf{t}) \prod_{i=1}^{k} P(f_i|c, \mathbf{n}, s, \mathbf{t}) = \frac{n(c) + st(c)}{N + s} \prod_{i=1}^{k} \frac{n(f_i\,|c) + st(f_i\,|c)}{n(c) + st(c)}. \tag{2}$$

Equation (2) is the posterior probability densities of class and features returned by NBC. However, the specification of any single prior entails the risk of issuing fragile prior-dependent classifications. Walley's IDM overcomes this problem, by letting the parameters $\mathbf{t}$ vary within intervals instead of being fixed to precise values. In particular, $\mathbf{t}$ vary within the polytope $\mathcal{T}$, defined by the following constraints:

$$\mathcal{T} := \left\{ \sum_{c\in\mathcal{C}} t(c) = 1, \quad \sum_{f_i\in\mathcal{F}_i} t(f_i\,|c) = t(c), \;\; 0 < t(f_i\,|c) < t(c) \quad \forall (i, f_i, c) \right\}. \tag{3}$$

Thus, IDM takes into consideration all the priors densities which belong to the simplex $\mathcal{T}$. Notice that, the above constraints are necessary and sufficient conditions to ensure that all the densities, obtained by letting $\mathbf{t}$ vary in $\mathcal{T}$, are proper. Walley's IDM satisfies the *representation invariance principle* because the uncertainty about any event does not depend on refinements or coarsening of categories; the *likelihood principle*, because posterior inferences depend on the data through the likelihood function only. The specific approach used by NCC [7] to identify the non-dominated classes is called *maximality* [6]. Consider the

$1 - 0$ utility functions associated with the actions of choosing class c' or c''. The family of posterior probabilities $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ (obtained by letting $\mathbf{t}$ vary in $\mathcal{T}$) are used to determine the lower expected utility of deciding between c' or c''. Class c' dominates c'' if the expected utility w.r.t. $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ of choosing a class c' over c'' is strictly positive for each $\mathbf{t} \in \mathcal{T}$. In the case of NCC, c' dominates c'' if and only if [7]:

$$\inf_{\mathbf{t}\in\mathcal{T}} \frac{P(c', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)}{P(c'', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)} = \inf_{\mathbf{t}\in\mathcal{T}} \left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{k-1} \prod_{i=1}^{k} \frac{n(f_i\,|c'\,) + st(f_i\,|c'\,)}{n(f_i\,|c''\,) + st(f_i\,|c''\,)} > 1. \tag{4}$$

When faced with a prior-dependent instance, NCC identifies more non-dominated classes and issues an indeterminate classification, thus preserving reliability.

2.1 Restricting IDM

As already observed, by considering all prior Dirichlet densities such that $0 < t(c) < 1$ and $0 < t(f_i\,|c) < t(c)$ for all i and c, IDM excludes the extremes of the simplex $\mathcal{T}$, which correspond to improper densities. This means that the simplex $\mathcal{T}$ is obtained by restricting the set $0 \leq t(c) \leq 1$ and $0 \leq t(f_i\,|c) \leq t(c)$ by an *arbitrary small* ϵ. If $n(f_i\,|c) > 0$ and $n(f_i\,|c) < n(c)$ for each feature i and class c, the posterior densities corresponding to the extremes of the simplex $\mathcal{T}$ are proper for any choice of ϵ. Thus, in this case, we can let ϵ go to zero. This is proved in [7] for the optimization in (4). In particular, it is shown that the infimum of (4) is obtained by letting $t(f_i\,|c'\,) \to 0$ and $t(f_i\,|c''\,) \to t(c'')$, thus using extreme densities (i.e., $\epsilon = 0$). Then, problem (4) is solved by optimizing on $t(c'')$ only. Since function (4) is convex with respect to $t(c'')$, the minimization can be solved exactly and efficiently.

In some cases, the use of extreme densities in (4) generate what we call the *class problem* and the *feature problem*. The class problem, already observed in [7, Sec. 6], takes place when a class c'' is never observed in the sample; in this case, it is difficult for an alternative class c' to dominate c'': for any value f_i of any feature, there are no data for estimating $P(f_i|c'')$, which therefore under the IDM varies between 0 and 1 and is set to 1 during the minimization. As this behavior repeats for each feature, $P(\mathbf{f}_i|c') \ll P(\mathbf{f}_i|c'')$, thus often preventing an alternative class c' to dominate c''. In fact, c'' will be often identified as non-dominated. The *feature problem* happens instead when there are no observations of one or more values of a certain feature conditional on class c'. In this case, there are no observations for estimating $P(f_i|c')$, which goes to sharp to zero during the solution of (4); this leads to sharp 0 also $P(c', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)$, because of F_i alone, regardless the information coming from all the remaining features. When either the class or the feature problem happen, NCC can get very indeterminate, while at the same time NBC achieves good accuracy on the instances indeterminately classified by NCC; this can be seen as disappointing behavior of NCC.

Note that if $n(f_i\,|c'\,) = 0$ (feature problem), the choice of extreme prior density ($t(f_i\,|c'\,) = 0$) leads to improper posteriors. Although, in this case, IDM is still well-defined, we cannot let ϵ go sharp to zero in the optimization in (4). Therefore, the set of posteriors and the set of non-dominated classes will depend on the choice of ϵ. In this paper, we propose two approaches to remove the extreme densities from $\mathcal{T}$, in order to increase the NCC determinacy.

The first approach uses the following restricted set for $\mathbf{t}$:

$$\mathcal{T}_\epsilon = \left\{ t(c'), t(c'') \geq \epsilon, t(c') + t(c'') = 1, \epsilon \leq t(f_i|c') \leq t(c'), \epsilon \leq t(f_i|c'') \leq t(c'') \right\} \tag{5}$$

where $\epsilon \in (0, 0.5]$; we call NCC$_\epsilon$ the resulting classifier. Such an approach is appropriate to deal with the feature problem, as it guarantees $t(f_i|c') \geq \epsilon$ and therefore avoids sharp zeros in the computation of the numerator; however, it should not be too effective against the class problem, as the conditional probabilities at the denominator will nevertheless reach $1 - \epsilon$. Moreover, the credal set of (5) satisfies the RIP, as it is not dependent on the number of categories. However, this comes at a cost. In fact, Eq.(5) requires to adjust the boundary of the credal set on every different pairwise comparison. For instance, when comparing c' with c'', $t(c')$ and $t(c'')$ are lower-bounded by ϵ while $t(c) = 0$ for all the remaining classes; but when comparing c' against c''', $t(c')$ and $t(c''')$ are lower-bounded by ϵ, while $t(c'')$ (and all the remaining $t(c)$) goes to 0 . Therefore, such an approach does not respect the likelihood principle as the set of priors depends on the couple of classes under exam. Moreover, it cannot be guaranteed that if c' dominates c'' and c'' dominates c''', then also c' dominates c''' (*transitivity*), because the pairwise comparisons are in fact performed on credal sets having different boundaries. For this reason, this approach should be used with small values of ϵ, so to enable addressing the feature problem while only minimally perturbing the credal set of the IDM. Using a value of ϵ comprised between 0.01 and 0.1, transitivity has been however always satisfied in our experiments. When the priors are restricted to be in $\mathcal{T}_\epsilon$, the analytical optimization procedure described in [7] remains valid, because the derivatives of function are unchanged compared to [7].

The second approach is based on a ϵ-contamination of the uniform prior of NBC with the set of priors in $\mathcal{T}$, which results in the set:

$$\mathcal{T}_c := \left\{ \begin{array}{ll} \sum\limits_{c \in \mathcal{C}} t(c) = 1, \quad t(c) \in \left[\epsilon_0 \dfrac{1}{|\mathcal{C}|}, \epsilon_0 \dfrac{1}{|\mathcal{C}|} + (1 - \epsilon_0)\right], \\ \sum\limits_{f_i \in \mathcal{F}_i} t(f_i\,|c) = t(c), \quad t(f_i|c) \in \left[\epsilon_i \dfrac{t(c)}{|\mathcal{F}_i|}, \epsilon_i \dfrac{t(c)}{|\mathcal{F}_i|} + (1 - \epsilon_i) t(c)\right], \quad \forall (i, c) \end{array} \right. \tag{6}$$

where the ϵ_0 refers to the class variable, while for each feature a different parameter $\epsilon_i \in (0, 1)$ can be specified. We call NCC$_c$ the resulting classifier, where c stands for “contaminated”. This approach, unlike the previous one, satisfies LP (no dependence of the set of priors on the data) but not RIP, since the priors depend on the number of classes (through $1/|\mathcal{C}|$) and number of categories of the features (through $1/|\mathcal{F}_i|$). The minimization problem, has to be numerically approximated because the interval for $t(f_i\,|c)$ depends on $t(c)$ and function (4) is not convex in $t(c)$. When $\epsilon_0 \to 1$ and $\epsilon_i \to 1$ $\forall i$, the set of priors collapses to the uniform prior and thus the classifier coincides with the NBC. Instead, NCC$_\epsilon$ never reduces to a single prior; with $\epsilon = 0.5$ it uses a single prior to compare a couple of classes, but this prior changes with the couple of classes.

3 Credal Model Averaging

Let us consider NBC again: given k features, there are 2^k possible NBCs, each characterized by a different subset of features; we denote by $\mathcal{M}$ the set of such models and by m a generic model. By feature selection, one can identify a single best feature set and then work with a single NBC. An alternative approach is *Bayesian Model Averaging* (BMA), which instead averages over *all* the 2^k different NBCs, the weight assigned to each classifier being proportional to its posterior probability. The joint probability $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ is obtained by marginalizing m out:

$$P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}) \propto \sum_{m \in \mathcal{M}} P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}, m) P(\mathbf{n}|m, s, \mathbf{t}) P(m), \tag{7}$$

where $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}, m)$ is the posterior probability of $c, \mathbf{f}$ computed by m, $P(\mathbf{n}|m, s, \mathbf{t})$ represents the *likelihood* of model m and $P(m)$ the prior probability of m. Dash and Cooper [5] provide an exact and efficient algorithm to compute BMA over 2^k NBCs. This algorithm has been extended to imprecise probabilities in [3], giving rise to credal model averaging (CMA) . In particular, CMA specifies a **set** of prior over the models instead of adopting a single $P(m)$; in fact, CMA imprecisely averages over the 2^k NBCs. CMA is free from both the feature problem and the class problem, as its base classifiers are NBCs.

4 Comparing credal classifiers

In order to completely describe the performance of a credal classifier, 4 indicators are necessary: *determinacy*: i.e, the percentage of determinate classifications; *single accuracy*: the accuracy of the classifier when determinate; *set-accuracy*: the accuracy of the classifier when indeterminate; *indeterminate output size*: the average number of classes returned by the classifier when indeterminate. Instead, to compare credal classifiers we adopt two metrics which have been introduced in [4]. We refer to a classifier as *accurate* on a certain instance if its output includes the correct class, regardless how many classes it has returned; we refer to a classifier as *determinate* if its output contains only a single class. The *discounted-accuracy* is: $d\text{-}acc = \frac{1}{n}\sum_{i=1}^{n}(accurate)_i/|Z_i|$, where $(accurate)_i$ is a 0-1 variable, showing whether the classifier is accurate or not on the i-th instance; $|Z_i|$ is the number of classes returned on the i-th instance. Yet, there is no reason for *linearly* discounting the accuracy on the number of returned classes; an alternative non-parametric approach proposed in [4] removes this arbitrariness, being based on a *rank test.* The *rank test* is more robust than *d-acc*, as it does not encode any (arbitrary) functional form for discounting accuracy on the basis of the output size; yet, it uses less pieces of information than d-acc and can be therefore be less sensitive. Overall, a cross-check of both metrics is recommended.

5 Results

We presents results on 45 classification data sets; they are publicly available from the WEKA data sets page.[2] Over each data set, we perform 10 runs of 10-folds cross-validation, namely 100 training/test experiments. Numerical features have been discretized via the entropy-based discretization; within each training-test experiment, the bins are learned on the current training set and then applied unchanged on the current testing set. The comparison of BMA and NBC shows 11 wins for NBC, 25 ties, 9 wins for BMA (over each data sets, the accuracies measured during cross-validation have been compared with a t-test, with $\alpha = 5\%$). There is therefore a balance between the two classifiers. Instead, when considering credal classifiers, CMA clearly dominates NCC: according to the rank test [or the discounted accuracy], there are 23 [26] wins for CMA, 17 [14] ties and 5 [5] wins for NCC [3]. In particular, CMA has much larger determinacy than NCC (on average, 95% vs 76%) and also higher discounted accuracy (0.76 vs 0.70 on average). We must recall that CMA also includes an ϵ parameter, which controls the determinacy of CMA. Yet, even adopting different values of ϵ, CMA remains much more determinate than NCC. The scatter plots of determinacy and discounted accuracy for the two classifiers are in Fig. 1.

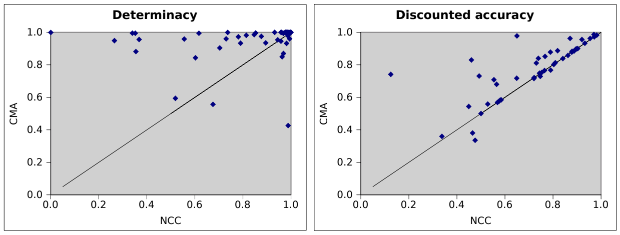

Fig. 1. Scatter plot of determinacy and the discounted-accuracy of CMA and NCC.

We focus on three data sets which highlight the consequences of the class and the feature problem; the characteristics of the data sets and the performance of the classifiers are shown in Tab.1. NCC has very low determinacy on these three data sets; this implies that *many* instances are classified in prior-dependent way by NBC. Yet, the classifications issued by NBC using the uniform prior are

[2] `http://www.cs.waikato.ac.nz/~ml/weka/index_datasets.html`

[3] On each data set, the values discounted accuracy measured for NCC and CMA during cross-validation have been compared via t-test, significance 5%.

quite accurate; in particular, they are much more accurate than simply returning the majority class, as shown in Table 1. The model of prior-ignorance which characterizes NCC is indeed theoretically sound, but in these cases its large indeterminacy appears questionable. On squash-stored, NCC suffers from the feature problem: the feature *fruit* has 22 states and requires to estimate 66 parameters for the conditional densities, from only 52 instances; removing this feature increases the NCC determinacy from 31% to 60%. Instead, NCC$_\epsilon$ properly deals with this feature: even using a small ϵ (0.01), determinacy increases from 32% to 42% and discounted-accuracy from 0.48 to 0.57, approaching that of CMA. Moreover single-accuracy (accuracy when determinate) also increases from 70% to 79%, showing that the feature problem prevents NCC to extract useful information from the remaining features. With NCC$_c$, it is instead necessary to use a larger ϵ (recall that NCC$_c$ lower-bounds the conditional probabilities in the numerator of Eq.(4) by $\frac{\epsilon}{|F_i||\mathcal{C}|}$); for instance, with ϵ=0.1, NCC$_c$ achieves determinacy 36% with discounted-accuracy of 0.53. Moreover, NCC$_c$ too has higher the single-accuracy than NCC. Note that for both NCC$_\epsilon$ and NCC$_c$, adopting increasing ϵ would steadily increase determinacy, as in fact it will reduce the credal set and thus the probability of the instance being prior-dependent. Instead, it cannot be foreseen how the discounted accuracy will vary when ϵ is increased, as this depends on the trade-off between determinacy and accuracy, which cannot be predicted in advance.

Data set	N	Feats	$\|\mathcal{C}\|$	Majority	Accuracy	
					NBC	BMA
primary-tumor	339	17	22	25%	46%	36%
audiology	226	69	24	25%	79%	73%
squash-stored	52	24	3	44%	66%	59%

Data set	Determ.		Disc-acc		NBC accuracy when	
	NCC	CMA	NCC	CMA	NCC det.	NCC ind.
primary-tumor	10%	88%	0.19	0.36	70%	43%
audiology	7%	95%	0.21	0.70	98%	78%
squash-stored	32%	84%	0.49	0.58	70%	63%

Table 1. Comparison of NCC and CMA on three data sets especially difficult for NCC. Majority is the percentage of instances belonging to the most frequent class in the data set.

The low determinacy of NCC on both audiology and primary-tumor is instead due to the class problem, as several classes are never observed, or observed only once or twice; in fact, removing these classes from the data set largely increases the NCC determinacy. However, NCC$_\epsilon$ does not address the class problem, as

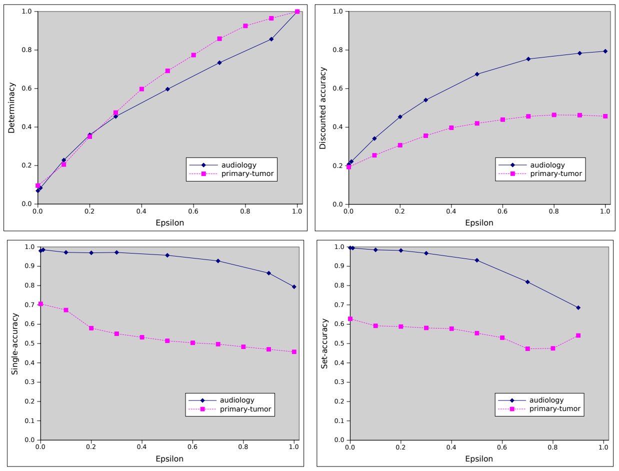

Fig. 2. Sensitivity of NCC$_c$ to the value of ϵ. For ϵ=1, NCC$_c$ corresponds to NBC; therefore, determinacy is 100% and set-accuracy is not measurable.

already pointed out; it is therefore more interesting analyse the behavior of NCC$_c$. In Fig.2, we show how the main indicators of performance of NCC$_c$ vary with different values of ϵ; for ϵ=1, the classifier corresponds to NBC. This plots highlight the trade-off between robustness and determinacy: increasing ϵ implies higher determinacy, which however comes generally at a cost of some accuracy, both on the instances determinately and indeterminately classified. Domain knowledge can suggest which is a reasonable choice of ϵ. As a last experiment, we have run NCC$_\epsilon$ and NCC$_c$, setting for both $\epsilon = 0.01$, on all the data sets. Overall, both classifiers achieve determinacy and discounted-accuracy which is significantly higher than that of NCC, although the impact is generally much lighter than in the three extreme examples previously analyzed.

6 Conclusions

We have presented two approaches to restrict the set of priors of the IDM, in order to overcome the large indeterminacy of NCC, when dealing with what we have called the feature problem and the class problem, discussing advantages and disadvantages of such approaches from a theoretical point of view. Then, by experiments, we have shown that on the data sets where such two problems heavily penalize the NCC determinacy, a small restriction of the set of priors allows to considerably increase the determinacy of the classifier without penalizing its reliability. This is particularly important on real problems, where a trade-off between informativeness and robustness is desirable. As future work, these two approaches could constitute a starting point to design a new classifier, which performs credal model averaging over NCCs characterized by different sets of features. In this case, restricting the imprecision could be a key-issue to manage the quantity of returned indeterminate classifications.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-121785/1, 200021-118071/1 and “Hydronet” project. We also thank the referees for their helpful remarks.

References

1. A. Benavoli and C.P. de Campos. Inference from multinomial data based on a MLE-dominance criterion. In *Proc. of the 10th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 22–33. Springer, 2009.
2. J.-M. Bernard. An introduction to the Imprecise Dirichlet Model for multinomial data. *Int. J. of Approximate Reasoning*, 39(2-3):123–150, 2005.
3. G. Corani and M. Zaffalon. Credal Model Averaging: An Extension of Bayesian Model Averaging to Imprecise Probabilities. In *Proc. of the 2008 European Conf. on Machine Learning and Knowledge Discovery in Databases*, pages 257–271, 2008.
4. G. Corani and M. Zaffalon. Lazy naive credal classifier. In *Proc. of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.
5. D. Dash and G.F. Cooper. Model Averaging for Prediction with Discrete Bayesian Networks. *J. of Machine Learning Research*, 5:1177–1203, 2004.
6. P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.
7. M. Zaffalon. Statistical inference of the naive credal classifier. In *ISIPTA ’01: Proc. of the Second Int. Symp. on Imprecise Probabilities and Their Applications*, pages 384–393, The Netherlands, 2001. Shaker.