Comments on "Imprecise probability models for learning multinomial distributions from data. Applications to learning credal networks" by Andrés R. Masegosa and Serafín Moral

Marco Zaffalon[a,*], Giorgio Corani[a]

[a]*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)*
Galleria 2, 6928 Manno (Lugano), Switzerland

Abstract

We briefly overview the problem of learning probabilities from data using imprecise probability models that express very weak prior beliefs. Then we comment on the new contributions to this question given in the paper by Masegosa and Moral and provide some insights about the performance of their models in data mining experiments of classification.

Keywords: Near-ignorance, learning probabilities, imprecise probability, imprecise Dirichlet model, credal classification.

1. Introduction and discussion

Learning probabilities from a sample of i.i.d. data is not an easy problem. Many solutions have been proposed in the literature, among which there are frequentist and Bayesian methods, and more recently methods from imprecise probability. The paper by Masegosa and Moral [1] focuses on the latter, and in particular on a generalization of Bayesian statistical methods to imprecision, and so do we in this short discussion paper.

A method that has attracted much attention in this respect is that based on the *imprecise Dirichlet model* (IDM) by Walley [2] (see also [3]). In the binomial case, the model was proposed also by Bernard [4] and by Walley himself in his seminal book [5, Section 5.3.2]. Details about the IDM are given in the paper by Masegosa and Moral, so we refrain from doing the same here. We rather focus on a specific aspect of the IDM, which is its being a model of near-ignorance a priori.

A model of prior beliefs has the property of *near ignorance* if it leads to vacuous lower and upper expectations for some gamble of interest f (a gamble is just a bounded random variable): $\underline{E}(f) = \inf f$ and $\overline{E}(f) = \sup f$. For instance, the IDM is such that for any event A (and the related indicator function I_A), we obtain that $\underline{E}(I_A) = \underline{P}(A) = 0$ and $\overline{E}(I_A) = \overline{P}(A) = 1$. Models of this type are hence good candidates to represent a state of prior ignorance about those gambles. This is useful in the attempt to stay with (generalized) Bayesian methods while getting as close as possible to the ideal of letting the data speak for themselves. On the other hand, the word 'near' is there to say that such a model is not yielding vacuous expectations a priori for all gambles. If that were actually the case, then no matter the amount of data we had, posterior beliefs would be vacuous as prior beliefs [5, Section 7.3.7]—thus making learning from data impossible. Learning is instead possible using near-ignorance models: in the case of the IDM, for instance, the difference between the upper and the lower posterior expectations decreases as more data are available.

Near-ignorance models thus originated hopes to realize a kind of 'objective-minded' approach to learning from data within the tradition of subjective statistics. However, an unexpected result has partly undermined those hopes: in 2009, Piatti et al. [6] have shown, in very common and general conditions, that any prior near-ignorance model for the multinomial case will remain vacuous a posteriori if the observations that we do (that is, the data that we see) are imperfect. This means that learning is most often impossible using prior near-ignorance models when there can be

*Corresponding author
Email addresses: zaffalon@idsia.ch (Marco Zaffalon), giorgio@idsia.ch (Giorgio Corani)

errors in the process of making the observations—no matter how unlikely these errors can be—or, more in general, when we cannot directly access data about the multinomial process we are after, but only about a different process that is (even very much) related to it. The fact that observations are arguably always imperfect in real settings has then casted doubts about the actual possibility to use prior near-ignorance models in practice.

Rather then considering this a defect of prior near-ignorance models, our view is that it is a natural—albeit disturbing—phenomenon: it says that there is a fundamental limit to the degree at which we can weaken our prior beliefs without creating a conflict with the possibility to learn. A basic question remains, though: is there a way to learn from data when there is no prior information?

This question is puzzling for many reasons. One in particular is that the IDM, as well as other near-ignorance models (e.g., see [7]), works relatively fine when one neglects that observations are possibly subjects to errors and feeds the IDM with them as if they were perfect: this procedure appears to be, even if paradoxically, useful in applications. Masegosa and Moral (right after Example 4), as well as Piatti et al. themselves [6, Section 5], argue that this approach is not tenable on the theoretical level. And yet it has some appeal; in fact, we believe that this aspect of near-ignorance is worth being analyzed in greater depth, as we feel that the paradox has not been fully clarified yet. Moreover, the research on prior near-ignorance models is still quite lively and actually yielding interesting outcomes. Thus it seems to be a matter of fact that near-ignorance models have still something to say, despite the mentioned problems.

An entirely different path to answer the question of learning from data in case of no prior information, consists in regarding the mentioned limitation as a hard constraint on prior models that requires them to be necessarily informative (that is, not near-ignorant). This is a very reasonable attitude in the light of such a limitation, and in fact we believe that first Moral [8] and then Masegosa and Moral, with their current paper, should be commended for having studied such an alternative path. In the present paper, they define in particular a 'learning principle' (Definition 1) that provides quite a general characterization of the models for which learning from data is possible even in case of imperfect observations. The learning principle is especially easy to appreciate in the form given by their Theorem 3. Such a form shows also that it is quite unlikely that the learning principle can be satisfied by a near-ignorance prior. Thus the learning principle is, in our view, an important contribution of Masegosa and Moral's paper.

The learning principle motivates the development of models, to be used in absence of prior information, that express very weak prior beliefs while being informative. The assumption of no prior information makes the definition of such models uncomfortable because every informative model expresses an initial bias that is arbitrary under the assumption; the struggle lies in defining the bias in a way that is as harmless as possible. Masegosa and Moral's attempt yields the *imprecise sample size Dirichlet model* (ISSDM).

The ISSDM is a set of symmetric distributions—each of them assuming a priori that the probabilities to be learned from data are uniform—using different equivalent sample sizes. It was proposed, in a special case, already by Walley [5, Section 5.4] in order to model situations of prior-data conflict. A prior-data conflict arises when the likelihood is concentrated in the tails of the prior; this makes the posterior inferences more uncertain than in less extreme cases. The increased uncertainty materializes in the ISSDM in the form of imprecision of those inferences.

Whether a model originally developed for a situation of prior-data conflict can be possibly used in a condition of ignorance is a matter of discussion. Walley, for instance, clearly distinguishes the two situations, in this way emphasizing that prior ignorance, on the one side, and prior-data conflict, on the other, are different possible sources of indeterminacy of a statistical model [5, Section 5.4].

In their paper, Masegosa and Moral make an attempt to address this question through their cautious proposal of 'strong symmetry' (see the discussion after their Example 5). To exemplify the situation, let us consider an urn with red, orange and green balls in some unknown proportions. Consider also a random, but fixed and unknown to us, permutation σ of the colors, say $\sigma(\text{red}) = \text{green}$, $\sigma(\text{orange}) = \text{red}$, $\sigma(\text{green}) = \text{orange}$. A person scrambles the balls and draws one of them without watching; say that the ball is red. We are then told that the color is green, according to σ. The ball is put back in the urn and the process is repeated. The question is whether or not the nature of our information about this process is different from that of the process where the random permutation is not employed. Masegosa and Moral argue that if for us the information is the same, then the models of prior beliefs for the two situations should be the same as well, which implies that they should be made of symmetric distributions; this would justify the use of the ISSDM in the case of no prior information. But is it actually the case that our information is the same in the two situations? There are reasons to doubt it, as also Masegosa and Moral point out. Our view is in fact that knowing that a random permutation is employed (even if we do not know which permutation) is a non-vacuous piece of information. Therefore we are no longer ignorant about the domain and for this reason we cannot justify the use of the ISSDM in

a situation of ignorance. This is consistent with the fact that the ISSDM yields uniform probabilities a priori, which cannot be regarded as modeling our ignorance about the domain (but just our indifference about the possible outcomes of the sampling process). Therefore we are not inclined towards taking up the strong symmetry principle as a way to justify using the ISSDM in a condition of ignorance.

Perhaps one should just drop the ambition to find a principled justification for using an informative model in a situation of ignorance. And perhaps then what one should rather do is just to analyze how sensibly the model performs in the type of application under study. Masegosa and Moral take this pragmatic stance in the second half of their paper, where they discuss the application of the ISSDM to learn graphical models and in particular credal classifiers. In the following we only focus on the classification experiments to keep the discussion short.

We recall that a classifier is a model of the relationship between the characteristics of an object, expressed by a set of *features*, and its category, which is also called its *class*. Classifiers are usually learned from data and later used as predictors: given the features of a new object they have to predict its class. Their predictive performance is measured also on data through some statistics. *Credal classifiers* are an extension of traditional classifiers that allow for set-valued predictions of classes: the idea is that a credal classifier aims at dropping the unlikely classes for a certain object and that this does not always lead to come up with a single class. The output set is made of classes that are all (potentially) optimal given the information available to the classifier; the size of the set is the smaller the greater the information in the data. Therefore the output set is typically larger (we say that the prediction is more *indeterminate*) when the data set is small or it contains many missing values. Credal classifiers have in fact been originally introduced with the aim of obtaining reliable classifications also in conditions of poor information in the data. We remark that it is not trivial to empirically compare different credal classifiers, as each of them typically implements a different trade-off between informativeness and robustness. However, utility-based indicators, of the type used in Masegosa and Moral's paper, provide a sensible approach to this end.

Masegosa and Moral introduce two new credal classifiers: the 'local' and the 'global' ISSDM-NCC. Both extend the well-known naive Bayes classifier (NBC) to imprecision through the ISSDM. Naive Bayes is a simple and yet powerful classifier based on the assumption of stochastic independence of the features given the class of the object under consideration. Thus the joint probability represented by NBC factorizes as the product of the marginal probability of the class and the conditional probability of the features given the class. To learn such probabilities in a Bayesian way one has to specify the equivalent sample size s, which expresses the strength of one's beliefs in the prior distribution.

The global ISSDM-NCC is a collection of NBCs learned using different values of s taken from a certain finite set $\mathcal{S}$ (we employ a different notation compared to that by Masegosa and Moral, in order to keep things simple in this short paper). It returns the union of the classes issued by the various NBCs.

The local ISSDM-NCC allows one to change the value of s when learning different marginal or conditional probabilities within the same naive structure. This leads to a much wider collection of NBCs; thus the inferences for the local model are computed via sampling, despite the use of the simple naive structure. In this respect the global ISSDM-NCC is preferable to the local model, as it leads to quick and exact inferences. Moreover, at present we are uncertain as to the rationale behind the idea of allowing the equivalent sample size vary within the same naive Bayes model; for this reason we would prefer to have some insight about, or a justification of, such an idea that characterizes the local approach. On the other hand, a nice property of the local model is that it naturally embeds an approach for removing irrelevant features. For a given feature, it draws the value of $s \in \mathcal{S}$ and decides on the basis of the Bayesian score whether the feature should be discarded or not. A similar approach is not present in the global model and perhaps it could be considered in that case as well: the idea could be to learn the most probable naive structure, for each of $s \in \mathcal{S}$, which possibly discards some features as irrelevant. Thus the global model would instantiate $|\mathcal{S}|$ NBCs, each different as for the value of s and potentially also as for the structure. Such model would perform credal classification while allowing to monitor the sensitivity of the naive structure to s, thus linking the ideas of the experiments of Sections 6.1 (learning the structure of credal networks) and 6.2 (credal classification). Having said this, we also note that the experiments in the paper do not show any great difference between the local and the global ISSDM-NCC. For this reason in the following we do not distinguish between those two variants, as simply refer to them as the ISSDM-NCC.

In particular, Masegosa and Moral compare the ISSDM-NCC against two alternative credal classifiers, NCC2 and CMA, on various data sets. An important common trait of the considered classifiers is that they all extend naive Bayes to imprecision. It is worth discussing briefly the difference among these extensions.

NCC2 is designed on the basis of the IDM to represent prior near-ignorance about the parameters of naive Bayes. Its

corresponding set of probabilities contains priors that are also very skewed; sometimes this leads to high indeterminacy in the predictions, as it has been detailed in previous works ([9, Section 6], [10]). The resulting classifier is hence very cautious as a consequence of the near-ignorance assumption: it implicitly takes into account also the worst-case scenarios in order to decide what is the best set-valued prediction to issue for a given object. For this reason, NCC2 may as well not stand out when the focus is on assessing the *average* behavior of a classifier on a number of data sets: for the worst-case scenarios that arise in some data sets are often more than counterbalanced by the (arguably many) other, less extreme, scenarios related to the remaining data sets. We conjecture that this could be the reason why more aggressive (i.e., less cautious) credal classifiers such as CMA and ISSDM-NCC achieve higher average utility than NCC2.

CMA [11] is based on a set of naive Bayes classifiers, each learned adopting a uniform prior on the parameters. Each NBC has a different feature set. CMA averages the posterior probabilities returned by the different NBCs, giving more weight to the NBCs that receive more support a posteriori. CMA assumes a condition close to prior ignorance about the weights of the NBCs; for this reason it shows the classical behavior of credal classifiers: it is quite indeterminate on small data sets and becomes more informative as the sample size increases. However CMA is more determinate than NCC2 since (i) each NBC is learned using the uniform prior on the parameters and (ii) the credal set on the coefficients of the mixture does not include extremely skewed priors.

The performance of ISSDM-NCC and CMA is quite similar on the original data sets but quite different when the same data sets are heavily downsampled. A striking peculiarity of ISSDM-NCC is its high determinacy when dealing even with very small data sets. Such a behavior is almost unique among credal classifiers. The reason is that ISSDM-NCC aims at detecting prior-data conflict rather than at modeling prior near-ignorance: if prior-data conflict is not detected, ISSDM-NCC returns determinate classifications in spite of the small sample size. In a number of cases such a conflict is absent, allowing ISSDM-NCC to behave aggressively, with few indeterminate classifications, and obtaining higher utility than CMA on very small data sets.

Our summary view about the experiments, in brief, is that they appear to support using the ISSDM-NCC as a very viable approach for the typical data mining task of analyzing many data sets while achieving a good average performance—often in the contest of some competition. On the other hand, we would still opt for using NCC2, or some other classifier based on prior beliefs close to near-ignorance, when the focus in on addressing a real problem about which we have very little information a priori and on the basis of which we have to take some sensible decision. The reason is that the high determinacy of ISSDM-NCC, even in case of poorly informative learning sets, might not work well together with some specific applications, as it could lead to an overconfident behavior.

We note also that the ISSDM-NCC might not be easily applicable for case studies in which experts can provide domain knowledge. Experts are often relatively at ease by providing intervals of probabilities (or some other type of qualitative assessments) regarding certain events. Such probabilities can easily be encoded by shrinking the set of priors of credal classifiers such as CMA and NCC; this seems more difficult to implement in the case of ISSDM-NCC.

Finally, an open problem is how to choose a specific instance of the ISSDM-NCC. Empirical or theoretical guidelines in this respect might be necessary, should the ISSDM-NCC become largely adopted.

2. Concluding comments

Masegosa and Moral have focused on what we believe to be a very important problem. We praise them in particular for having taken up a research direction that has been explored very little so far and that is certainly worth considering: the definition of weakly informative imprecise prior models to be used in a condition of ignorance. Their empirical analyses of the ISSDM appear to indicate that their model can be a valuable tool to learn sets of Bayesian networks from data as well as for pattern classification; it is also a model that does not seem to yield to an excess of caution in the inference, which is obviously important in applications. On the other hand, it is especially appealing to us that the ISSDM seems to work quite well in classification also in case of small learning samples. This could lead the ISSDM to become quite a competitive credal classifier for data mining applications.

The fact that the ISSDM yields uniform probabilities a priori can also somewhat be regarded as going back to the Bayesian tradition, which uses uniform priors to model lack of knowledge. Even though this is inconsistent with an actual lack of knowledge, it could contribute to give a chance to the ISSDM to be used in problems characterized by little prior information also outside the community of imprecise probability.

Acknowledgements

This work has been partly supported by the Swiss NSF grants Nos. 200021_146606 / 1 and 200020_137680 / 1.

References

[1] A. R. Masegosa, S. Moral, Imprecise probability models for learning multinomial distributions from data. Applications to learning credal networks, International Journal of Approximate Reasoning, in press. doi:10.1016/j.ijar.2013.09.019.

[2] P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society, Series B 58 (1996) 3–57, with discussion.

[3] J. M. Bernard, An introduction to the imprecise Dirichlet model for inference with multinomial data, International Journal of Approximate Reasoning 39 (2–3) (2005) 123–150.

[4] J. M. Bernard, Bayesian interpretation of frequentist procedures for a Bernoulli process, American Statistician 50 (1996) 7–13.

[5] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.

[6] A. Piatti, M. Zaffalon, F. Trojani, M. Hutter, Limits of learning about a categorical latent variable under prior near-ignorance, International Journal of Approximate Reasoning 50 (2009) 597–611.

[7] A. Benavoli, M. Zaffalon, A model of prior ignorance for inferences in the one-parameter exponential family, Journal of Statistical Planning and Inference 142 (7) (2012) 1960–1979.

[8] S. Moral, Imprecise probabilities for representing ignorance about a parameter, International Journal of Approximate Reasoning 53 (3) (2012) 347–362.

[9] M. Zaffalon, Statistical inference of the naive credal classifier, in: G. de Cooman, T. L. Fine, T. Seidenfeld (Eds.), ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, 2001, pp. 384–393.

[10] G. Corani, A. Benavoli, Restricting the IDM for classification, IPMU 2010: Proc. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods (2010) 328–337.

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