International Journal of Approximate Reasoning 55 (2014) 1594–1596

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Discussion

Learning imprecise probability models: Conceptual and practical challenges

Fabio Gagliardi Cozman

Escola Politecnica, Universidade de Sao Paulo, Av. Prof. Mello Moraes 2231 - Cidade Universitaria, Sao Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 18 November 2013
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 28 April 2014

Keywords:
Imprecise probabilities
Multinomial learning
Credal networks

ABSTRACT

The paper by Masegosa and Moral, on "Imprecise probability models for learning multinomial distributions from data", considers the combination of observed data and minimal prior assumptions so as to produce possibly interval-valued parameter estimates. We offer an evaluation of Masegosa and Moral's proposals.

1. Introduction

The paper by Masegosa and Moral [7], on "Imprecise probability models for learning multinomial distributions from data", not only deals with technically demanding issues but also with several complex conceptual questions concerning the nature of statistical models. On top of that, they also consider "applications to learning credal networks." Hence the reader cannot have doubts that Masegosa and Moral have set their eyes on ambitious goals.

Their paper does contribute with several novel ideas, and the balance of their results is certainly positive and welcome. The point of the present commentary is to examine some of their questionable assumptions and proposals, so as to investigate ways in which they can be improved, and to suggest issues that deserve further consideration.

2. Learning with few prior assumptions: the learning principle and the ISSDM

The theme of Masegosa and Moral's paper is the construction of a data generating model from observed data. One might argue that the construction of a model should assume nothing a priori. But what constitutes "nothing a priori" is a difficult matter. The Bayesian strategy is to express "ignorance" about parameters through a prior distribution. To indicate that nothing is assumed a priori, a Bayesian may adopt flat priors, reference priors, maximum-entropy priors [2]. There is no real agreement on how to encode ignorance in such a way, and some of these efforts seem more desperate than heroic.

Another possibility is to encode the absence of a priori assumptions by adopting a *set* of prior distributions. The larger the set, the less is assumed about the parameters of interest. This is exactly the strategy followed by Masegosa and Moral.

The first part of their paper looks at the specific case of multinomial distributions. Thus we have to estimate parameters $\{\theta_i\}_{i=1}^k$, all nonnegative, such that $\sum_i \theta_i = 1$. Each parameter θ_i is interpreted as the probability of a particular value of a variable X; that is, θ_i means $P(X = x_i)$. The natural conjugate prior is given by a Dirichlet distribution; that is, a distribution proportional to $\prod_{i=1}^k \theta_i^{st_i-1}$, where s and $\{t_i\}_{i=1}^k$ are parameters. Masegosa and Moral dismiss methods that try to encode

DOI of original article: http://dx.doi.org/10.1016/j.ijar.2013.09.019.

http://dx.doi.org/10.1016/j.ijar.2014.04.016
0888-613X/

ignorance by a single Dirichlet distribution, and instead move to models based on sets of prior Dirichlet distributions. As noted in the previous paragraph, their motivation is clear: ignorance cannot be translated into a single prior; rather, ignorance should be expressed by a set of priors.

However, it is not easy to select a set of prior distributions. Suppose one adopts a maximal set of prior distributions (a set containing *every possible* distribution over a given space), and Bayes rule is applied elementwise. Then the set of posterior distributions is also maximal. That is, if we use a maximal set of distributions to encode a priori ignorance, nothing can ever be learned — *ex nihilo nihil fit*, mocks the Bayesian. One way to start from complete ignorance and to reach useful inferences is to abandon Bayes rule, for example by adopting Dempster rule [3, Section 6]; but if we are to stay with Bayes rule, we must judiciously construct our sets of prior distributions. Such sets must be large enough to properly encode ignorance, but not so large that they lead to vacuous inferences.

A popular way to combine Dirichlet distributions with sets of prior distributions is to employ the Imprecise Dirichlet Model (IDM). Here we take a set of Dirichlet distributions where parameter s is fixed but parameters t_i are left free [15]. The IDM displays the interesting feature that each one of values x_i is assigned the maximal probability interval $[0, 1]$. That is, very little is assumed a priori. The IDM also embodies a principle of symmetry: all values of interest are treated equally. Additionally, the IDM embodies the "representation invariance principle" (RIP): if we change the number k of values by coarsening or refining the set of values, the inferences about values do not change.

The IDM is quite attractive, but as noted by the authors it has its own drawbacks. The authors mention the fact that the IDM often prevents learning [10]; that is, the set of prior distributions is too large, leading to vacuous posterior inferences.

What to do? Masegosa and Moral wish to focus on a new "imprecise Dirichlet scheme" that goes by the mind-bending acronym ISSDM, for Imprecise Sample Size Dirichlet Model. To justify their strategy, they propose a "learning principle" that essentially requires any event of interest to have positive prior probability. This is an old idea, for instance going back to Shimony's strict coherence, where probability zero is only attached to the empty event [12]. The constraint that any possible event must receive positive probability, however minute, often referred to as *regularity* [6], and has received considerable attention and sizeable criticism [5]. For one, the mathematics of regularity is not simple for general spaces [1], and a discussion of possible solutions based on infinitesimal probabilities would take us into complicated matters [4,13,17].

Masegosa and Moral wish to stay with real positive numbers, and they smartly couch their learning principle so as to apply only to some measurable events, thus avoiding several mathematical difficulties (a similar strategy is contemplated by Skyrms [13]). However, we are still left with the old debate on the justification of strict-coherence/regularity, and while the jury is still unsure on this issue, it does not seem that past history speaks much in favor of the learning principle.

All of this suggests that the learning principle should be viewed with some healthy skepticism. Perhaps the authors can offer more arguments for it, and spend some future effort in connecting their proposal with the existing literature.

In any case, even if the learning principle may be questionable, the ISSDM is an interesting idea, first suggested by Walley [14, Section 5.4]. Actually, the learning principle by itself does not solve our problem, which is how to represent uncertainty with minimal assumptions. There is tension here: can we really move away from maximal prior sets, while still keeping assumptions at a minimum?

Indeed, the ISSDM does bring its own stock of concerns. For one, it does not satisfy the representation invariance principle; that principle was once used by Walley to argue for imprecise probabilities, but now it is found deficient. If RIP is inadequate, does it mean that the IDM loses some of its appeal, or even that imprecise probabilities lose some of their appeal? Apparently the authors think that RIP can be dismissed while imprecise probabilities are desirable to model ignorance. But even as we let RIP rest in peace, we find that the ISSDM does present some surprising behavior as it typically leads to dilation: that is, we may start with precise probabilities, only to find that, after observation and inference, we are left with more imprecise conclusions [11]. There is debate around dilation [9,16], and some may ask whether they should pay *not* to collect data in some circumstances... Interestingly, for the ISSDM, dilation is in a sense informative as it is a marker for conflict between data and prior assumptions. Do the authors feel that dilation is a regrettable feature, an illuminating phenomenon, or an inevitable pain?

The authors also suggest a generalized version of the ISSDM that consists of a set of ISSDMs. The advantage is increased flexibility, but the practitioner now faces a number of parameters to set, so some of the initial beauty of almost-no-assumption-models like the IDM cannot be fully maintained. Finding easy and intuitive ways to specify generalized ISSDMs is an important task for the future.

3. Learning credal networks

If several variables are considered at once, then it is likely that many combinations of their values are never observed in a reasonably-sized dataset. Hence prior distributions have increased importance in producing estimates. And exactly because many variables are present, we must find ways to specify prior distributions without too many controlling parameters, lest we are to spend all available time fiddling with priors. Thus it is appropriate that Masegosa and Moral have chosen to apply their ISSDMs to the construction of multivariate models, in particular to models that encode independence relations.

A credal network is a structure consisting of a graph where each node is a variable, and where edges encode independences. Each variable is assumed independent of its nondescendants given its parents. In short, a credal network can be understood as a set of Bayesian networks [8]. Most applications of credal networks in the literature adopt the assumption that all Bayesian networks represented by a given credal network share the same edges (hence they share the same inde-

pendences). Another common assumption is that the conditional distributions associated with a particular variable do not restrict in any way the conditional distributions associated with any other variable; the credal network is then called *separately specified*. Masegosa and Moral depart from these assumptions: their priors introduce restrictions between conditional distributions, and, more importantly, then consider a generalized kind of credal network that may encode many different independence relations at once.

There are several challenges in learning this sort of credal network from data, and to do so the authors put together an array of techniques: scores, discretization methods, and Monte Carlo schemes for inference. The latter idea is an interesting contribution, as existing algorithms that operate with credal networks usually focus on optimization, while here the focus is on sampling. The only concern is, again, that the initial beauty of almost-no-assumption-models is blurred by the machinery required for practical operation.

However, in the middle of this battle, Masegosa and Moral have found a true gem. It turns out that by clever selection of parameters, one can guarantee that a single graph is built from data, but a graph where some edges are *necessary* in that they are selected for all choices of parameters, while others are *ambiguous* in that they are sensitive to the choice of parameters. This sort of graphical structure is rather intuitive and informative, and should be quite useful not only when learning networks from data, but also when building networks from opinions of a single expert, or from opinions of a sets of experts. The authors must be congratulated for this contribution.

Finally, a word on the evaluation of credal networks. This is always a difficult matter, well discussed in the literature (as noted by the authors). Their experiments reveal interesting features of their methods, and it seems that more practical experience is needed. For instance, how can we really compare a Naive Bayes classifier with a classifier that sometimes outputs set of classes? The authors argue that "it should be better to produce an imprecise result" when a precise answer will be sensitive to parameters, but some may disagree — in practice, one might argue that a precise answer is always better. Thus comparisons with strategies that always yield a precise answer (for instance, minimax strategies) may be useful in the future.

4. Final comments

The problems tackled by Masegosa and Moral are complex and sometimes perplexing; their proposals are bold and it is only fair that they raise further difficult questions. The learning principle surely deserves more discussion. The ISSDM family of models encodes important ideas that should be pursued in earnest, not only by theoretical study but also by practical evaluation. Finally, the insights by Masegosa and Moral concerning credal network learning, particularly in connection with ambiguous links, are quite welcome and merit practical use.

I am grateful to have had the opportunity to comment on this valuable contribution, and can only hope that the authors continue producing similarly stimulating material.

References

[1] Thomas E. Armstrong, William D. Sudderth, Locally coherent rates of exchange, Ann. Stat. 17 (3) (1989) 1394–1408.
[2] James Berger, The case for objective Bayesian analysis, Bayesian Anal. 1 (3) (2006) 385–402.
[3] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (1967) 325–339.
[4] K. Easwaran, Regularity and hyperreal credences, Philos. Rev. 123 (1) (2014) 1–41.
[5] Alan Hájek, Is strict coherence coherent? Dialectica 66 (3) (2012) 414–424.
[6] D. Lewis, Philosophical Papers, Volume II, Oxford University Press, 1986.
[7] A.R. Masegosa, S. Moral, Imprecise probability models for learning multinomial distributions from data. Applications to learning credal networks, Int. J. Approx. Reason. 55 (7) (2014) 1548–1569, http://dx.doi.org/10.1016/j.ijar.2013.09.019 (in this issue).
[8] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, California, 1988.
[9] A.P. Pedersen, G. Wheeler, Demystifying dilation, Erkenntnis (October 2013), http://dx.doi.org/10.1007/s10670-013-9531-7, in press.
[10] A. Piatti, M. Zaffalon, F. Trojani, M. Hutter, Limits of learning about a categorical latent variable under prior near-ignorance, Int. J. Approx. Reason. 50 (2009) 597–611.
[11] T. Seidenfeld, L. Wasserman, Dilation for sets of probabilities, Ann. Stat. 21 (9) (1993) 1139–1154.
[12] A. Shimony, Coherence and the axioms of confirmation, J. Symb. Log. 20 (1955) 1–28.
[13] B. Skyrms, Strict coherence, sigma coherence, and the metaphysics of quantity, Philos. Stud. 77 (1) (1995) 39–55.
[14] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.
[15] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. R. Stat. Soc. B 58 (1) (1996) 3–57.
[16] R. White, Evidential symmetry and mushy credence, in: Oxford Studies in Epistemology, vol. 3, 2010, pp. 161–186.
[17] T. Williamson, How probable is an infinite sequence of heads? Analysis 67 (2007) 173–180.