International Journal of Approximate Reasoning 55 (2014) 1570–1574

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On various ways of tackling incomplete information in statistics

Didier Dubois

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 2 April 2014
Accepted 3 April 2014
Available online 24 April 2014

Keywords:
Incomplete information
Random sets
Fuzzy sets
Evidence theory
Imprecise probability
Possibility theory

ABSTRACT

This short paper discusses the contributions made to the featured section on Low Quality Data. We further refine the distinction between the ontic and epistemic views of imprecise data in statistics. We also question the extent to which likelihood functions can be viewed as belief functions. Finally we comment on the data disambiguation effect of learning methods, relating it to data reconciliation problems.

1. Introduction

The set of position papers gathered in the special section on low quality data proposes various ways of handling incomplete information in statistics. Imprecision may pervade the chosen model or the observed data. Moreover, prior information is generally poor. Two of these contributions focus on imprecise data, two other ones on the lack of prior information. On the one hand, one question is whether set-valued data can be handled just like any other kind of complex data. On the other hand, there is the problem of choosing a formal framework for handling incomplete information. The paper cosigned by this discussant [3] shares with the paper on fuzzy random variables [21] the use of multiple-valued mappings and random sets, but the way proposed to exploit set-valued data is radically different, as further discussed in the next section. Another pair of strikingly different papers dealing with related issues is formed by Denoeux and Masegosa–Moral papers that deal with the role of likelihood functions when prior information is poor and cannot be modelled by a unique probability distribution on the parameter space. In [14], what is proposed is essentially a form of sensitivity analysis over Bayesian inference, where the likelihood function alone is considered as totally insufficient to allow for any form of learning. However, Denoeux [4] argues to the converse, namely by exploiting an idea originally proposed by Shafer in his book [16]: he claims that the likelihood function does inform us to some extent on the value of the parameter of a model, when an observation becomes available. It is then possible to handle fuzzy data as well. Finally the paper by Huellermeier [12], even if it does not use likelihood functions explicitly (but he shows how they can be laid bare), suggests that the chosen class of models may help reducing the imprecision of the data. In the following we briefly comment these contributions.

2. On the distinction between ontic and epistemic data

In our position paper [3], we made the distinction, also endorsed by Huellermeier [12], between ontic and epistemic views of fuzzy set-valued data. The impressive set of statistical methods developed by the Oviedo SMIRE team [21] considers fuzzy set-valued data as precise entities belonging to a space of functions [11], equipped with suitable operations in order to preserve the fuzzy set semantics of such functions (especially fuzzy arithmetic operations). As a consequence, while the

http://dx.doi.org/10.1016/j.ijar.2014.04.002
0888-613X/© 2014 Elsevier Inc. All rights reserved.

mean value in this setting is a fuzzy interval, the variance is precise (based on sophisticated distances between fuzzy sets). In this sense, the view of fuzzy data advocated by this group is clearly ontic. However, the applications they developed elsewhere (such as human perception of length, and flood prediction [2]) handle low quality human-originated data on numerical quantities; to quote them [21]: "variables or attributes [that] can only be observed imprecisely". However the ontic view of set-valued data is primarily devoted to natural entities that take the form of fuzzy sets and that are tainted with variability: observations of regions in an image, time intervals during which some activities take place, blood vessel snapshots, vectors of performance ratings across a population of candidates, etc.

But things are not that simple. In fact, the ontic–epistemic distinction does not correspond to the objective–subjective distinction exactly, that is, ontic set-valued data may not just reflect sets that occur as such in the nature. There are circumstances when epistemic set-valued data, even if they are imprecise descriptions of otherwise point-valued variables can be treated as ontic entities. Here are two examples:

- Suppose one imprecisely measures a precisely defined attribute a number of times, say via a number of different observers (e.g. human testimonies on the value of a quantity), but the actual aim of the statistics is to model the variability in the imprecision of the observers. In other words, while such fuzzy data are subjective descriptions of an otherwise objective quantity, they can be considered as ontic with respect to the observers. In particular, if agreeing observers provide nested set-valued estimates of a constant but ill-known value (with various degrees of confidence), one may consider that the various levels of precision correspond to a form of variability, which justifies the use of a scalar distance between such sets in the computation of the variance. But it is the variance of the imprecision levels of the observer responses that is obtained. This variance says little about the properties of the objective quantity on which observers report.
- Sometimes human perceptions refer to a complex matter that cannot be naturally represented by a precise numerical value. For instance, ratings in a dish tasting experiment are verbal rather than numerical. Imprecise terms then refer to no clear objective feature of the phenomenon under study. For instance, the taste of a dessert does not directly describe the objective ingredients of the dish. So, the collected imprecise data can be considered ontic, because you want to know if people will like the dessert, not how much butter or sugar it contains. Here again, the human perceptions can be handled as ontic entities. However, the question is then to figure out whether the collected subjective data in this kind of situation is liable of a representation by means of a fuzzy set over a numerical scale: indeed, the very reason why a precise numerical estimate is inappropriate in this kind of situation is because a one-dimensional numerical scale does not make sense. Then, the statistician is not better off when representing human perceptions by means of fuzzy sets (let alone trapezoidal ones) on a meaningless numerical scale.

In summary, a set-valued statistic is ontic if the set representation captures the essence of the issue under study; it is epistemic if the purpose is to provide some information on the precise entity that could not be precisely observed because of the poor quality of the knowledge. Adopting an ontic approach to the statistics of human perceptions of otherwise objective quantities yields a description of the observer behaviour, not of the natural phenomenon on which this observer reports.

3. On likelihoods in the contexts of belief functions, possibility, and imprecise probability theories

In his position paper, Denoeux [4] argues in favour of the use of likelihood functions for building data-driven belief functions, assuming the contour function of the belief function should be taken as proportional to the likelihood function, thus following a suggestion made in Shafer's book [16].

This point of view is strengthened by the formal result proved in the paper, namely that this approach is the consequence of the likelihood principle, the consistency with Bayes rule in case a probabilistic prior is available, and the minimal commitment principle: there exists a unique minimally committed belief function whose contour function is proportional to the likelihood function $L(\theta) = P(x|\theta)$, where x is the observation, and $\theta \in \Theta$ is the parameter of the distribution generating x, and this belief function is consonant. In other words, it is a necessity measure based on the possibility distribution

$$\pi(\theta) = \frac{L(\theta)}{\max_{\tau \in \Theta} L(\tau)}.$$

This result looks compelling. However, it relies on an assumption that may sound questionable, namely that the relative information contained in belief functions is evaluated on the basis of their commonalities. This view has been advocated quite early by Smets [18], but there are alternative definitions of relative information to the comparison of commonalities, such as specialisation (random set inclusion) and the comparison of plausibility (or belief) functions [8,20]. Interestingly, in [20], Smets advocates the latter approach as the basis for the minimal commitment, not the comparison of commonalities. He also spends some time discussing the concept of specialisation, but seems to have given up using commonality functions. It suggests that the issue of choosing a proper information comparison technique for belief functions was not quite settled in his mind. In the following discussion, we focus on the choice between the comparison of plausibilities, and of commonalities on which Denoeux relies.