International Journal of Approximate Reasoning 55 (2014) 1502–1518

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Statistical reasoning with set-valued information: Ontic vs. epistemic views

Inés Couso [a], Didier Dubois [b,*]

[a] Department of Statistics, Universidad de Oviedo, Spain
[b] IRIT, CNRS Université de Toulouse, France

ARTICLE INFO

Article history:
Available online 29 July 2013

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

ABSTRACT

In information processing tasks, sets may have a conjunctive or a disjunctive reading. In the conjunctive reading, a set represents an object of interest and its elements are subparts of the object, forming a composite description. In the disjunctive reading, a set contains mutually exclusive elements and refers to the representation of incomplete knowledge. It does not model an actual object or quantity, but partial information about an underlying object or a precise quantity. This distinction between what we call ontic vs. epistemic sets remains valid for fuzzy sets, whose membership functions, in the disjunctive reading are possibility distributions, over deterministic or random values. This paper examines the impact of this distinction in statistics. We show its importance because there is a risk of misusing basic notions and tools, such as conditioning, distance between sets, variance, regression, etc. when data are set-valued. We discuss several examples where the ontic and epistemic points of view yield different approaches to these concepts.

1. Introduction

Traditional views of engineering sciences aim at building a mathematical model of a real phenomenon, via a data set containing observations of the concerned phenomenon. This mathematical model is approximate in the sense that it is a simplified abstraction of the reality it intends to account for, but it is often precise, namely it typically takes the form of a real-valued function that represents, for instance, the evolution of a quantity over time. Approaches vary according to the class of functions used. The oldest and most common class is the one of linear functions, but a lot of works dealing with non-linear models have appeared, for instance and prominently, using neural networks and fuzzy systems. These two techniques for constructing precise models have been merged to some extent due to the great similarity between the mathematical account of fuzzy rules and neurons, and their possible synergy due to the joint use of linguistic interpretability of fuzzy rules and learning capabilities of neural nets [9]. While innovative with respect to older modeling techniques, these methods remain in the traditional school of producing a simplified and imperfect substitute of reality as observed via precise data.

Besides, there also exists a strong tradition of accounting for the non-deterministic aspect of many real phenomena subject to randomness in repeated experiments, including the noisy environment of measurement processes. Stochastic models enable to capture the general trends of populations of observed events through the use of probability distributions having a frequentist flavor. The probability measure attached to a quantity then reflects its variability through observed statistical data. Again in this approach, a stochastic model is a precise description of variability in physical phenomena.

* Corresponding author.
E-mail addresses: couso@uniovi.es (I. Couso), dubois@irit.fr (D. Dubois).

0888-613X/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ijar.2013.07.002

More recently, with the emergence of Artificial Intelligence, but also in connection with more traditional human-centered research areas like Economics, Decision Analysis and Cognitive Psychology, the concern of reasoning about knowledge has emerged as a major paradigm [42]. Representing knowledge requires a logical language and this approach has been mainly developed in the framework of classical or modal logic, due to the long philosophical tradition in this area. Contrary to the numerical modeling tradition, such knowledge-based models are most of the time tainted with incompleteness: a set of logical formulae, representing an agent's beliefs is seldom complete, that is, cannot establish the truth or falsity of any proposition. This concern for incomplete information in Artificial Intelligence has strongly affected the development of new uncertainty theories [32], and has led to a critique of the Bayesian stance viewing probability theory as a unique framework for the representation of belief that mimics the probabilistic account of variability.

These developments question traditional views of modeling as representing reality independently of human perception and reasoning. They suggest a different approach where mathematical models should also account for the cognitive limitations of our observations of reality. In other words, one might think of developing the epistemic approach to modeling. We call *ontic model* a precise representation of reality (however inaccurate it may be), and *epistemic model* a mathematical representation both of reality and the knowledge of reality, that explicitly accounts for the limited precision of our measurement capabilities. Typically, while the output of an ontic model is precise (but possibly wrong), an epistemic model delivers an imprecise output (hopefully consistent with the reality it accounts for). An epistemic model should of course be as precise as possible, given the available incomplete information, but it should also be as plausible as possible, avoiding unsupported arbitrary precision.

This position paper[1] discusses epistemic modeling in the context of set-based representations, and the mixing of variability and incomplete knowledge as present in recent works in fuzzy set-valued statistics. The outline of the paper is as follows. In Section 2, we discuss the use of sets for the representation of epistemic states as opposed to the representation of objective entities. Then in Section 3 we draw the consequences of this discussion in the theory of random sets, laying bare three approaches relying on the same mathematical tool. In Section 4, we show that the distinction drawn between epistemic and ontic random sets affects the practical relevance of formal definitions one can pose in the random set setting. It is shown that notions of conditioning, independence and variance differ according to the adopted point of view. The consequences of this distinction in the way interval regression problems can be posed are briefly discussed in Section 5.2. Finally, Section 6 carries the distinction between ontic and epistemic sets over to fuzzy sets, and, more briefly, to random fuzzy sets.

2. Ontic vs. epistemic sets

A set S defined in extension, is often denoted by listing its elements, say, in the finite case $\{s_1, s_2, \ldots, s_n\}$. As pointed out in a recent paper [33] this representation, when it must be used in applications, is ambiguous. In some cases, a set represents a real complex lumped entity. It is then a conjunction of its elements. It is a precisely described entity made of subparts. For instance, a region in a digital image is a conjunction of adjacent pixels; a time interval spanned by an activity is the collection of time points where this activity takes place. In other cases, sets are mental constructions that represent incomplete information about an object or a quantity. In this case, a set is used as a disjunction of possible items, or of values of this underlying quantity, one of which is the right one. For instance I may only have a rough idea of the birth date of the president of some country, and provide an interval as containing this birth date. Such an interval is the disjunction of mutually exclusive elements. It is clear that the interval itself is subjective (it is my knowledge), has no intrinsic existence, even if it refers to a real fact. Moreover this set is likely to change by acquiring more information. The use of sets representing imprecise values can be found for instance in interval analysis [54]. Another example is the set of models of a proposition in logic, or a propositional knowledge base: only one of them reflects the real situation; this is reflected by the DNF form of a proposition, i.e., a disjunction of its models, each of which is a maximal conjunction of literals.

Sets representing collections C of elements forming composite objects will be called *conjunctive*; sets E representing incomplete information states will be called *disjunctive*. A conjunctive set is the precise representation of an objective entity (philosophically it is a *de re* notion), while a disjunctive set only represents incomplete information (it is *de dicto*). We also shall speak of *ontic* sets, versus *epistemic* sets, in analogy with ontic vs. epistemic actions in cognitive robotics [43]. An ontic set C is the value of a set-valued variable X (and we can write $X = C$). An epistemic set E contains the ill-known actual value of a point-valued quantity x and we can write $x \in E$. A disjunctive set E represents the epistemic state of an agent, hence does not exist per se. In fact, when reasoning about an epistemic set it is better to handle a pair (x, E) made of a quantity and the available knowledge about it.

A value s inside a disjunctive set E is a possible candidate value for x, while elements outside E are considered impossible. Its characteristic function can be interpreted as a possibility distribution [77]. This distinction between conjunctive and disjunctive sets was already made by Zadeh [78] distinguishing between set-valued attributes (like the set of sisters of some person) from ill-known single-valued attributes (like the unknown single sister of some person). The study of incomplete

[1] An expanded version of a first draft (by the second author) that is part of the COST702 final report published by Springer as an edited volume "Towards Advanced Data Analysis by Combining Soft Computing and Statistics", in: C. Borgelt, et al. (Eds.), Studies in Fuzziness and Soft Computing, vol. 285, 2013.