Informational Paradigm, management of uncertainty and theoretical formalisms in the clustering framework: A review

Pierpaolo D'Urso

Department of Social Sciences and Economics, Sapienza Università di Roma, P.za Aldo Moro, 5, 00185, Rome, Italy

ARTICLE INFO

ABSTRACT

Fifty years have gone by since the publication of the first paper on clustering based on fuzzy sets theory. In 1965, L.A. Zadeh had published "Fuzzy Sets" [335]. After only one year, the first effects of this seminal paper began to emerge, with the pioneering paper on clustering by Bellman, Kalaba, Zadeh [33], in which they proposed a prototypal of clustering algorithm based on the fuzzy sets theory. Starting from this paper, several uncertain clustering methods based on different theoretical approaches for modeling the uncertainty have been proposed. The present paper presents a systematic literature review of these clustering approaches. In particular, with respect to the Statistical Reasoning System, we first illustrate the connection between Information and Uncertainty from the perspective of the so-called Informational Paradigm, according to which Information is constituted by "Informational ingredients", specifically the "Empirical Information," represented by statistical data, and "Theoretical information" consisting of background knowledge and basic modeling assumptions. We then describe different kinds of uncertainty affecting the Information. Focusing on the uncertainty associated with a particular statistical methodology, i.e. Cluster Analysis, and adopting as theoretical platform the Informational Paradigm, we present a systematic literature review of different uncertainty-based clustering approaches -i.e. Fuzzy clustering, Possibilistic clustering, Shadowed clustering, Rough sets-based clustering, Intuitionistic fuzzy clustering, Evidential clustering, Credibilistic clustering, Type-2 fuzzy clustering, Neutrosophic clustering, Hesitant fuzzy clustering, Interval-based fuzzy clustering, and Picture fuzzy clustering. We thus show how all these clustering approaches are able of managing in different ways the uncertainty associated with the two components of the Informational Paradigm, i.e. the Empirical and Theoretical Information.

© 2017 Elsevier Inc. All rights reserved.

On the 50th anniversary of fuzzy clustering

In the knowledge discovery process, although the Statistical Reasoning System

may be efficient a halo of uncertainty always will permeate

the information and therefore the knowledge;

the only certainty is that there are no certainties.

E-mail address: pierpaolo.durso@uniroma1.it

1. Introduction

Statistical reasoning can be viewed as a specific instance of approximate reasoning, where uncertainty affects the various ingredients of the reasoning process, which is therefore characterized by "approximation."

In particular, Statistical Reasoning Systems embody two types of "informational" ingredients, the Empirical Information represented by the dataset, and the initial Theoretical Information, which includes basic modeling assumptions, previous knowledge, and other pieces of Theoretical Information concerning the processing assumptions and the cognitive conclusions of the knowledge acquisition process (the "informational gain" obtained by means of appropriate strategies of analysis applied in the above context).

All of the above informational ingredients are affected by some source of uncertainty. For example, the data may be imprecisely measured or vaguely defined (e.g. use of linguistic expressions); furthermore, they may only partially represent the universe of possible data describing the investigated phenomenon (for instance when they are sampled from a larger population). Moreover, the basic modeling assumptions may also be uncertain, and the same is true of the assumptions used for processing the data (in fact any particular specification of these assumptions involves uncertainty as to their validity in the given research framework). Finally, the results of the statistical analysis reflect the uncertainties associated with the various pieces of information used for drawing the conclusions. In this respect, we have to cope with an uncertainty propagation process matching a parallel information propagation process, within the same Statistical Reasoning System.

In the above framework, randomness, imprecision, vagueness, partial ignorance are different types of uncertainty requiring a specific treatment. Standard probability theory may not be sufficient for dealing with all of them. We argue that fuzzy sets theory, as well as other uncertainty theories -such as, e.g., Type-2 fuzzy sets theory, Intuitionistic fuzzy sets theory, Rough sets theory, Shadowed sets theory, Credal sets theory and Evidential theory, Possibility and Credibilistic theories, Neutrosophic sets theory, Hesitant sets theory or inferential logic based on conditional probability (seen as a function of the conditioning event)- can suitably integrate the traditional probability theory in order to deal with the complexity of statistical reasoning.

In this connection, the paper will focus on the specific area of Cluster Analysis to illustrate the different theoretical approaches used in the literature to manage the uncertainty in the clustering process.

Fifty years have gone by since the publication of the first paper on clustering, based on fuzzy sets theory. In 1965, L.A. Zadeh had published "Fuzzy Sets" [332]. After only one year, the first effects of this seminal paper began to emerge, with the pioneering paper on clustering by Bellman, Kalaba, Zadeh [33], in which they proposed a prototypal of clustering algorithm based on the fuzzy sets theory. Starting from this paper, several uncertain clustering methods based on different theoretical approaches for modeling the uncertainty have been proposed.

The present paper presents a systematic literature review of these clustering approaches.

In particular, the main aim of the paper is to show in an organic manner the impressive impact of the seminal papers on fuzzy clustering on different scientific communities– mathematicians, statisticians, computer scientists, and so on- in the last 50 years. In fact, as we can see below, a massive and diversified scientific production has characterized those fruitful years. To do this, we define a general theoretical platform, i.e. the so-called Informational Paradigm, to manage different kinds of information and uncertainty, organically interconnected, which characterize the Statistical reasoning methods and in particular the clustering processes. Thus, we analyze systematically and in detail the Informational Paradigm, the possible uncertainty affecting different kinds of information and the connected theoretical formalisms for managing in a different manner the uncertainty, focusing on the fuzzy set theory and on its more fruitful theoretical extensions and generalizations in a methodological point of view. Successively, we adopt the Informational Paradigm as theoretical platform for the clustering methodology, showing different approaches for managing the uncertainty in the classification process. In this way, we assume that the different uncertainty-based clustering approaches are defined on the basis of the Information Paradigm. Thus, in different sections of the paper, we review systematically and in detail the more relevant uncertainty-based clustering approaches proposed in the literature for classifying objects and explain the respective theories used for managing the uncertainty. For each clustering approach, we illustrate the chronology of the various theoretical and methodological contributes, showing, with respect to the Informational Paradigm, the information ingredients and the uncertainty measures connected to different clustering approaches. Furthermore, we compare in a chronological point of view the different uncertainty-based clustering approaches and the connected uncertainty theories, showing the different timing of the impacts of the various uncertainty theories from the respective clustering approaches and then the different metabolic process of the theoretical results used in the respective clustering methodologies.

The paper is organized as follows. Starting from the definition of Informational Paradigm (Section 2), we illustrate various non probabilistic formalisms for managing uncertainty in data analysis (Section 3), including fuzzy sets theory and its developments, and theories that manage imprecision and uncertainty in a different way. Focusing on the fuzzy sets theory and on some of its recent developments -i.e. the Type-2 fuzzy sets theory, Intuitionistic fuzzy sets theory, Rough sets theory, Shadowed sets theory, Credal sets theory and Evidential theory, Possibility and Credibilistic theories, Neutrosophic sets theory, Hesitant sets theory and Picture Fuzzy Sets- in Section 4, we present a review of clustering methods based on the various formalisms present in the literature. As we shall see, these formalisms manage in different ways the uncertainty associated with the two components of the Informational Paradigm, i.e. the Empirical and Theoretical Information. A summary and some conclusions are presented, respectively in Sections 5 and 6.