A Multiobjective Genetic Fuzzy System with Imprecise Probability Fitness for Vague Data

Luciano Sánchez, Inés Couso and Jorge Casillas

Abstract— **When questionnaires are designed, each factor under study can be assigned a set of different items. The answers to these questions must be merged in order to obtain the level of that input. Therefore, it is typical for data acquired from questionnaires that each of the inputs and outputs are not numbers, but sets of values.**

In this paper, we represent the information contained in such a set of values by means of a fuzzy number. A fuzzy statistics-based interpretation of the semantic of a fuzzy set will be used for this purpose, as we will consider that this fuzzy number is a nested family of confidence intervals for the value of the variable. The accuracy of the model will be expressed by means of an interval-valued function, derived from a recent definition of the variance of a fuzzy random variable.

A multicriteria genetic learning algorithm, able to optimize this interval-valued function, is proposed. As an example of the application of this algorithm, a practical problem of modeling in marketing is solved.

I. Introduction

When acquiring data from interviews or questionnaires, it is a common practice to evaluate the level of a factor by examining the answers to a set of different questions. These questions may show different aspects of the problem, or be redundant, to ensure the coherence of the data. Therefore, it is typical for data acquired from questionnaires that each of the inputs and outputs are not single numbers, but sets of values. The classical conversion of such a set of items into a compound value [1] consists in replacing each one of them by a suitable, numerical characteristic value, say its mean or median. This solution might not be the best one, because it discards the information about the dispersion of the items.

In this paper, we will transform each set of inputs into a fuzzy interval, that contains information about the average value and the dispersion of the items, and then learn the model from the produced fuzzy data. We will use a novel interpretation of the semantics of a fuzzy set [2], that regards fuzzy sets as families of confidence intervals, and let the learning be grounded in the minimization of an imprecisely known function, as proposed in [3]. We also propose to estimate the distance between the imprecisely known output data and the response of the model by mean of a new definition of squared error, derived from that of the variance of a fuzzy random variable. This paper also details the modifications that must be done to the NSGA-2 algorithm [4] in order to use our criterion. By last, this algorithm is applied to a marketing problem, that of modeling the consumer behavior from data obtained by questionnaires, and their crisp and fuzzy implementations are compared.

This work is organized as follows. Section II introduces the fuzzy representation issues, and Section III describes the imprecise probabilities-based objective criteria. Section IV briefly describes the practical problem based on consumer behavior models. Section V shows some obtained experimental results. Finally, Section VI concludes.

L. Sánchez is with the Computer Science Department of the Oviedo University, Spain. Email: luciano@uniovi.es. Inés Couso is with Statistics Department, Oviedo University, Spain. Email: couso@uniovi.es. Jorge Casillas is with the Computer Science Department, Granada University, Spain. Email: casillas@decsai.ugr.es.

II. A fuzzy interpretation of the semantic of a list of values

A. Semantics of a Fuzzy Set

Under the imprecise probabilities framework, it makes sense to understand a fuzzy set as a set of tolerances, each one of them is assigned a confidence degree, being the lower degree the narrower tolerance [5]. In particular, the α-cuts of the fuzzy set can be regarded as confidence intervals with degree $1-\alpha$ [2].

This representation allows us to codify the information contained in a set of numbers by means of a fuzzy set. This will be made clear with the example that follows. Let us suppose that a variable X has associated the items valued

$$X = \{2, 1, 3, 3, 2, 2, 4\}. \tag{1}$$

The most immediate calculation of a summary value is the sample mean, which is 2.429. While this is a good compromise value, we are discarding information that might be relevant: there are some items as low as 1, and others as high as 4. To gain additional insight about the importance of the dispersion of the values, we will assume that the set of items X is a sample of a larger population, whose mean is unknown. Given the sample X, we can calculate confidence intervals for the value of this mean, at different degrees. If we want to simplify the calculations, we can assume that the sample was drawn from a normal population. Then, the α-cuts of the fuzzy set $\widetilde{X}$ that represent the value of the variable are the confidence intervals

$$\widetilde{X}_\alpha = 2.429 \pm 0.9759 \cdot \mathrm{qt}_6\left(1-\frac{\alpha}{2}\right). \tag{2}$$

where qt_6 is the quantile function for the t distribution. A graphical representation of the membership function of $\widetilde{X}$ is shown in Figure 1. Observe that we can approximate it by a triangular membership function without incurring large errors. The same procedure must be applied to all lists of input and