Expert Systems With Applications 46 (2016) 24–32

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Graded comparison of imprecise fitness values

Ignacio Montes, Susana Díaz, Susana Montes*

Department of Statistics and O. R., University of Oviedo, Spain

ARTICLE INFO

Keywords:
Vague data
Fitness value
Probabilistic relation
Fuzzy relation
Statistical preference

ABSTRACT

Genetic algorithms can be used to construct knowledge bases. They are based on the idea of "survival of the fittest" in the same way as natural evolution. Nature chooses the fittest ones in real life. In artificial intelligence we need a method that carries out the comparison and choice. Traditionally, this choice is based on fitness functions. Each alternative or possible solution is given a fitness score. If there is no ambiguity and those scores are numbers, it is easy to order individuals according to those values and determine the fittest ones. However, the process of assessing degrees of optimality usually involves uncertainty or imprecision.

In this contribution we discuss the comparison among fitness scores when they are known to be in an interval, but the exact value is not given. Random variables are used to represent fitness values in this situation. Some of the most usual approaches that can be found in the literature for the comparison of those kinds of intervals are the strong dominance and the probabilistic prior method. In this contribution we consider an alternative procedure to order vague fitness values: statistical preference. We first study the connection among the three methods previously mentioned. Despite they appear to be completely different approaches, we will prove some relations among them. We will then focus on statistical preference since it takes into consideration the information about the relation between the fitness values to compare them. We will provide the explicit expression of the probabilistic relation associated to statistical preference when the fitness values are defined by uniform and beta distributions when they are independent, comonotone and countermonotone.

1. Introduction

Expert systems traditionally involve a knowledge base (KB) containing the experience provided by expertise and a rule or inference engine, which derives solutions to particular situations from the facts and rules of the knowledge-base. The construction of the knowledge base is one of the major problems in this context (Merritt, 1989). Tools that allow to build and improve the population of KBs are necessary and genetic algorithms are a widespread option to optimize the KBs since many years ago (Baron, Achiche, & Balazinski, 2001; Payri, 1999; Xiong & Funk, 2006).

Genetic algorithms in artificial intelligence are supported by the same idea as the theory of evolution: Only those individuals that best fit survive in nature and this contributes to the improvement of species. In order to provide the best solution to a problem, alternatives are compared at each state and only the fittest ones survive to the next generation (next stage). Genetic algorithms have proved their ability in a very wide range of fields as generate and optimize fuzzy rule bases, create membership functions or tuning processes (see for example, Cordón, Herrera, Gomide, Hoffmann, & Magdalena, 2001a; Girgis, Sewisy, & Mansour, 2009; Jiménez et al., 2015). All these tasks can be considered as optimization or search processes.

In order to measure how well each individual fits the problem at each step fitness or objective functions are used. They usually assign a value to each element and those with highest fitness values are used to the next generation. Different fitness functions can be defined depending on the problem and it is a relatively easy task to provide a method or function that compares individuals and provides a best element when the information is precise. But imprecision appears very frequently in the context of expert systems (see, among many others, Armero, Artacho, López-Quílez, & Verdejo, 2011; Ładyżyński & Grzegorzewski, 2015; Palacios, Palacios, Sánchez, & Alcalá-Fdez, 2015). Tools that allow us to cope with this situation of incomplete or imprecise knowledge become necessary.

This is the purpose of genetic fuzzy systems (GFS) which apply genetic algorithms to design and improve fuzzy systems, where the data and/or the rules handled are vague. Genetic fuzzy systems have received a great deal of attention in the last years (Casillas & Martínez-López, 2009; Cordón, Herrera, Hoffmann, & Magdalena, 2001b; Elhag, Fernández, Bawakid, Alshomrani, & Herrera, 2015; Saniee, Mohamadi, & Habibi, 2011). The input data and the output solutions used to be crisp in the first contributions on this topic

* Corresponding author. Tel.: +34 985182133; fax: +34 985181902.
E-mail addresses: imontes@uniovi.es (I. Montes), diazsusana@uniovi.es (S. Díaz), montes@uniovi.es (S. Montes).

http://dx.doi.org/10.1016/j.eswa.2015.10.017
0957-4174/

(Herrera, 2008). However, since some years ago, some authors (see Sánchez, Couso, & Casillas, 2006; Sánchez & Couso, 2007; Sánchez, Couso, & Casillas, 2007; 2009) have dealt with imprecise data to learn and evaluate GFS.

The function that assesses the quality of a solution in the genetic algorithm, that is, the fitness function, is vague in this approach. There are different ways to model imprecision in the literature (random variables, fuzzy sets, …) and many contributions have been devoted to the comparison of fuzzy sets. We can cite, among the most recent ones, Ezzati, Allaviranloo, Khezerloo, and Khezerloo (2012), Wang (2015), Yu and Dat (2014), Zhang, Ignatius, Lim, and Zhao (2014).

Sánchez, Couso, and Casillas (2009) considered that the fitness values are unknown, but bounded in an interval and the imprecision is here modeled by a random variable. The traditional procedures to rank the fitness values in this case are either too strict, in the sense that they don't allow us to compare the intervals in many cases, as is the case of strong dominance (Limbourg, 2005), or they are based on estimating and comparing two probabilities (Sánchez et al., 2009).

In this contribution we consider a more general and flexible way to compare two intervals which is based on a probabilistic relation: statistical preference (De Schuymer, De Meyer, De Baets, & Jenei, 2003b; De Schuymer, De Meyer, & De Baets, 2003a). Firstly, we will see the connection between this new procedure and the classical ones. Later, we will focus on statistical preference and we will provide explicit expressions for its associated probabilistic relation. In particular, we will consider two cases. Firstly, we will assume that we have no information about what are the most likely values in any of the two intervals to be compared. Thus, the uniform distribution will be used. The assumption of a uniform distribution is not an artificial requirement and it can be considered in many situation as a consequence of lack of information (see, for instance, Sánchez et al., 2009; Teich, 2001). When this distribution is considered, we will obtain the specific expression of the associated probabilistic and fuzzy relations. Secondly, we will consider the situation in which the likelihood of the scores in the interval is increasing (resp. decreasing). Beta distributions model these situations. We will also study statistical preference in this case. Moreover, since statistical preference takes into account the possible dependence between the distributions compared, we will consider three different situations both for uniform and beta distributions: independence, comonotonicity and countermonotonicity.

This contribution is organized as follows: In Section 2 we introduce the problem and collect some different methods for comparing two fitness values proposed in the literature. In Section 3 we consider a generalization of these methods, based on a probabilistic relation and we study how this probabilistic relation can be equivalently represented by means of a fuzzy relation. In Section 4, the defuzzification of the probabilistic relation is considered and from here a total order is obtained. In particular, we prove the relationship between this new approach and some usual methods used in the literature for the comparison of fitness values. In the last part of this section we study the expression of the probabilistic relation in two cases: when the fitness values are modeled by uniform and beta distributions. Section 5 includes some final remarks and future work.

2. Usual methods of comparison

Let us consider two fitness values θ_1 and θ_2. In many situations, these fitness scores θ_1 and θ_2 are unknown, but we have some imprecise information about them. Thus, we cannot determine the value of θ_1 and θ_2, but we know two intervals where each of them is contained. These intervals can be obtained by means of a fuzzy generalization of the mean squared errors (for a more detailed explanation, see Sections 4 and 5 in Sánchez et al. (2009)) and they will be denoted by FMSE_1 and FMSE_2, respectively. The comparison of this two intervals is needed in order to choose the predecessor and the successor.

In this section we will introduce two of the most usual methods that can be found in the literature for the comparison of such intervals, the strong dominance and the probabilistic prior.

2.1. Strong dominance

The method of the *strong dominance* was considered in Limbourg (2005). In that case, if these two intervals are disjoint, then we have not any problem to determine the preferred interval and therefore the decision is trivial. The problem arises when the intersection is nonempty, since the intervals are incomparable. Thus, if $\text{FMSE}_1 = [a_1, b_1]$ and $\text{FSME}_2 = [a_2, b_2]$, it holds that:

- If $b_2 < a_1$, then θ_1 is preferred to θ_2 with respect to the strong dominance, denoted by $\theta_1 \succ_{\text{sd}} \theta_2$.
- If $b_1 < a_2$, then θ_2 is preferred to θ_1 with respect to the strong dominance, denoted by $\theta_2 \succ_{\text{sd}} \theta_1$.
- Otherwise, θ_1 and θ_2 are incomparable.

Thus, this method is too restrictive, since it can be used only in a very particular case. A tentative to solve this problem is to use the *stochastic order* (Levy, 1998; Müller & Stoyan, 2002; Shaked & Shanthikumar, 2002), introducing a prior knowledge about the probability distribution of the fitness. Given two random variables X and Y with associated cumulative distribution functions F_X and F_Y, X is stochastically preferred to Y if and only if $F_X \le F_Y$, with a strict inequality in at least one point x_0, and it is denoted by $X \succ_{\text{st}} Y$. In our context, the stochastic order can be formulated in the following way:

$\theta_1 \succ_{\text{st}} \theta_2 \Leftrightarrow P(\theta_1 \le x) \le P(\theta_2 \le x)$, for any $x \in \mathbb{R}$ and $P(\theta_1 \le x_0) < P(\theta_2 \le x_0)$, for some $x_0 \in \mathbb{R}$.

In particular, if we assume that the fitness value follows a uniform distribution (as in Teich, 2001), then:

$$\theta_1 \succ_{\text{st}} \theta_2 \Leftrightarrow \begin{cases} a_1 \ge a_2 \text{ and} \\ b_1 \ge b_2, \end{cases}$$

being at least one of the inequalities strict. In particular, if θ_1 strong dominates θ_2, then $\theta_1 \succ_{\text{st}} \theta_2$ regardless on the distribution of the fitness.

Nevertheless, the stochastic order does not solve all the problems of the strong dominance. For instance, incomparability is also allowed with respect to this method.

2.2. Probabilistic prior

Another method, called *the method of the probabilistic prior*, was proposed by Sánchez et al. (2009). As the stochastic order, it is based on a prior knowledge about the probability distribution of the fitness, $P(\theta_1, \theta_2)$. In that situation, a decision rule considered was to decide that $\theta_1 \succ_{\text{pp}} \theta_2$ if and only if

$$\frac{P((\theta_1, \theta_2) : \theta_1 > \theta_1)}{P((\theta_1, \theta_2) : \theta_1 \le \theta_2)} > 1. \tag{1}$$

Remark 1. If $P((\theta_1, \theta_2) : \theta_1 \le \theta_2) = 0$, the quotient of Eq. (1) is not defined, but it is assumed that $\theta_1 \succ_{\text{pp}} \theta_2$.

Although this method allows us to compare a class of random intervals wider than those compared by the method of the strong dominance, not every pair of intervals can be ordered. In particular, whenever $P((\theta_1, \theta_2) : \theta_1 = \theta_2) \ge 0.5$, θ_1 and θ_2 will be incomparable. Moreover, in this approach we consider a crisp order between the intervals, but if we are in a fuzzy context, with imprecise data, some kind of gradual comparison could be more appropriate as the starting point of the comparison.