International Journal of Approximate Reasoning 52 (2011) 1272–1282

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Nonparametric criteria for supervised classification of fuzzy data

Ana Colubi [a], Gil González-Rodríguez [a,*], M. Ángeles Gil [a], Wolfgang Trutschnig [b]

[a] Department of Statistics, University of Oviedo, 33007 Oviedo, Spain
[b] Research Unit on Intelligent Data Analysis and Graphical Models, European Centre for Soft Computing, 33600 Mieres, Spain

ARTICLE INFO

Article history:
Available online 28 June 2011

Keywords:
Fuzzy data
Random experiment
Supervised classification
Kernel estimation
Nonparametric density

ABSTRACT

The supervised classification of fuzzy data obtained from a random experiment is discussed. The data generation process is modeled through random fuzzy sets which, from a formal point of view, can be identified with certain function-valued random elements. First, one of the most versatile discriminant approaches in the context of functional data analysis is adapted to the specific case of interest. In this way, discriminant analysis based on nonparametric kernel density estimation is discussed. In general, this criterion is shown not to be optimal and to require large sample sizes. To avoid such inconveniences, a simpler approach which eludes the density estimation by considering conditional probabilities on certain balls is introduced. The approaches are applied to two experiments; one concerning fuzzy perceptions and linguistic labels and another one concerning flood analysis. The methods are tested against linear discriminant analysis and random K-fold cross validation.

1. Introduction

In many random experiments, as for instance sociological surveys, ecological studies, etc., some characteristics of interest can be assessed in a more meaningful scale if the respondents or the experts are allowed to indicate the degree of precision/imprecision of their answers/judgments/perceptions by means of fuzzy sets (see, for instance, [3] and Section 5 for more details). Those experiments can be soundly modeled by means of random fuzzy sets [17,19]. The approach to be used in order to handle random fuzzy sets depends on the aim of the experiment. The outputs of such experiments may be either fuzzy sets *per se* or ill-known values of a real-valued random variable. In the latter case the aim may refer to either the fuzzy sets that can be observed or to the underlying real-valued random variable. Random fuzzy sets in Puri and Ralescu's sense [19] are used to model experiments where the statistical interest lies in the fuzzy sets (irrespectively of the possible existence of any underlying real-valued random variable). This is the perspective that will be considered in this work. When the attribute of statistical interest is the underlying real-valued random variable that cannot be observed or measured precisely, different approaches may be considered (see, for instance, [6]).

From a formal point of view random fuzzy sets can be identified with a special case of function-valued random variables, although with some particular features concerning the natural arithmetic and metric structure (see [10,12] for a deep discussion). Functional data analysis has become an important area of research during the last two decades (see, for instance, [9,14,20,23]) and, as suggested in [8,12], it is possible to take advantage of some of the results developed for functional data to analyze fuzzy data.

As a first step in classification problems concerning random fuzzy sets, some unsupervised approaches have been considered in the literature (see, for instance, [11]). In this paper we deal with the supervised classification problem. That is,

* Corresponding author. Tel.: +34 985458118; fax: +34 985458110.
E-mail addresses: colubi@uniovi.es (A. Colubi), gil@uniovi.es (G. González-Rodríguez), magil@uniovi.es (M.Á. Gil), wolfgang.trutschnig@softcomputing.es (W. Trutschnig).

0888-613X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2011.05.007

given a set of possible groups and a training sample of fuzzy data of each group, the goal is to predict the group membership of a new fuzzy datum.

In the context of supervised classification of fuzzy data a simple idea is to defuzzify the imprecise data (in one or several crisp features) and to apply any multivariate supervised classification criteria. In this line, Yang et al. [22] have proposed a procedure based on the so-called defuzzified Choquet integral with fuzzy-valued integrand and a GA-based adaptive classifier-learning algorithm. In this procedure, the fuzzy data are projected onto a real axis of virtual variables and the classification is made with the optimality condition that the total misclassification rate is minimized. Nevertheless, the aim of this paper is to develop a supervised classification technique globally based on the whole fuzzy information, instead of only on some features extracted from the dataset. As indicated in [7], different approaches to this problem can be found in the literature for the functional context. Although most of these approaches are based on modifications of linear discriminant analysis, some nonparametric methods have been proposed in order to avoid the inconveniences that frequently arise due to the presence of nonlinear class boundaries.

The first supervised classification approach considered here is inspired by Ferraty and Vieu [7]. It is based on kernel estimation of the density of the distances between the data to be classified and each group. It is well-known that the criterion based on distances is equivalent to the optimal one based on the densities of the original variables for $\mathbb{R}$-valued random elements. When data are high-dimensional, the optimal criterion cannot be applied due to the curse of dimensionality. On the contrary, the criterion based on distances can still be applied but, unfortunately, it is not equivalent to the optimal one. Additionally, kernel density estimation requires large sample sizes. To avoid these inconveniences, a simpler approach eluding the density estimation (by considering conditional probabilities on certain balls) is introduced.

The rest of the paper is organized as follows. In Section 2 we introduce notation and basic concepts to be dealt with. In Sections 3 and 4 some distance-based and ball-based classification approaches are discussed, respectively. Section 5 is devoted to empirical results, and finally in Section 6 we conclude with some remarks and open problems.

2. Preliminaries

Let $\mathcal{F}_c(\mathbb{R}^p)$ denote the *class of fuzzy sets* $A : \mathbb{R}^p \to [0, 1]$ for which the α-levels A_α are nonempty compact convex subsets of $\mathbb{R}^p$ for all $\alpha \in (0, 1]$, whereby $A_\alpha = \{x \in \mathbb{R}^p | A(x) \geq \alpha\}$.

Recently, a new class of metrics based on the generalization of mid-point and spread of an interval has been defined. These metrics are very intuitive and versatile and exhibit good properties for statistical analysis (see [21]).

The generalized mid-point and spread of a fuzzy set A have been introduced as an alternative way of describing $A \in \mathcal{F}_c(\mathbb{R}^p)$. The idea consists firstly in levelwise projecting A onto all directions of the p-dimensional unit sphere $\mathbb{S}^{p-1}$, and secondly in calculating the mid-point and spread of all resulting intervals.

Formally, let $\alpha \in (0, 1]$ and $u \in \mathbb{S}^{p-1}$, and calculate the lengths $\pi_u(A_\alpha)$ of all orthogonal projections of A_α on this direction, i.e.

$$\pi_u(A_\alpha) = [\underline{\pi}_u(A_\alpha),\ \overline{\pi}_u(A_\alpha)] = [\,-s_{A_\alpha}(-u),\ s_{A_\alpha}(u)],$$

whereby s stands for the support function of a nonempty convex compact set (that is, $s_{A_\alpha}(u) = \sup_{a \in A_\alpha} \langle u, a\rangle$, $\langle \cdot, \cdot \rangle$ denoting the usual inner product in $\mathbb{R}^p$). The *generalized mid-point* and *generalized spread* of the fuzzy set A are then defined as the functions $\mathrm{mid}_A, \mathrm{spr}_A : \mathbb{S}^{p-1} \times (0, 1] \to \mathbb{R}$ such that

$$\mathrm{mid}_A(u, \alpha) = \mathrm{mid}_{A_\alpha}(u) = \frac{1}{2}(s_{A_\alpha}(u) - s_{A_\alpha}(-u)),$$
$$\mathrm{spr}_A(u, \alpha) = \mathrm{spr}_{A_\alpha}(u) = \frac{1}{2}(s_{A_\alpha}(u) + s_{A_\alpha}(-u)).$$

Note that in the interval-valued case, the unit sphere $\mathbb{S}^0$ reduces to $\{-1, 1\}$. Thus, $\mathrm{mid}_A(1, \alpha) = -\mathrm{mid}_A(-1, \alpha)$ coincides with the mid-point or center of A_α and $\mathrm{spr}_A(1, \alpha) = \mathrm{spr}_A(-1, \alpha)$ coincides with the spread or radius of A_α for all $\alpha \in (0, 1]$. The generalized mid-point and spread are defined as functions identifying the 'central points' and the 'imprecision' in the different directions of the Euclidean space. In this way, it becomes a meaningful characterization of fuzzy sets in $\mathcal{F}_c(\mathbb{R}^p)$ alternative to the classical support function.

The class of distances in [21] is defined from the distances between the level sets as a generalization of the Bertoluzza et al. metric [1], by considering L_2-type distances between the mid-points and the spreads. Specifically, for each level $\alpha \in (0, 1]$, one can define

$$d_\theta^2(A_\alpha, B_\alpha) = \|\mathrm{mid}\, A_\alpha - \mathrm{mid}\, B_\alpha\|^2 + \theta \|\mathrm{spr}\, A_\alpha - \mathrm{spr}\, B_\alpha\|^2,$$

where $\|\cdot\|$ is the usual L_2-norm in the space of the square-integrable functions $L^2(\mathbb{S}^{p-1})$ with respect to the uniform surface measure ϑ_p on $\mathbb{S}^{p-1}$, and $0 < \theta \leq 1$ determines the relative importance of the squared distance between the spreads in contrast to the squared distance between the mids.