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.

The metric D_θ^φ between fuzzy sets is defined as a weighted mean (w.r.t. a weighting measure φ formalized by means of a square-integrable absolutely continuous probability measure with support $[0, 1]$) of the distances between the level sets as follows:

$$D_\theta^\varphi(A, B) = \left(\int_{(0,1]} d_\theta^2(A_\alpha, B_\alpha) d\varphi(\alpha) \right)^{1/2}.$$

The weighting measure φ enables to enhance the intuitive (or subjective) interpretation of fuzzy sets – for instance one may treat every α-level as equally important (and therefore use the Lebesgue measure as weighting measure φ), or assign more mass either to α-levels close to 1 or to α-levels close to 0.

If $\|\cdot\|_{\tilde{2}}$ denotes the usual L_2-norm on the Hilbert space $\mathcal{H} = L^2(\mathbb{S}^{p-1} \times (0, 1])$ with respect to the product measure of the uniform surface measure ϑ_p on $\mathbb{S}^{p-1}$ and the weighting measure φ on $(0, 1]$, the metric can be equivalently expressed as

$$\left(D_\theta^\varphi(A, B)\right)^2 = \|\mathrm{mid}_A - \mathrm{mid}_B\|_{\tilde{2}}^2 + \theta \|\mathrm{spr}_A - \mathrm{spr}_B\|_{\tilde{2}}^2.$$

Let $(\Omega, \mathcal{A}, P)$ be a probability space. We define a *Random Fuzzy Set* (RFS) as a Borel measurable mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R}^p)$ (see [2,19]). The concepts of induced probability distribution and independence are the same as for general metric-space-valued random elements.

3. Density-based Classification Criteria for Fuzzy Data (DCCF)

Assume that we have a probability space $(\Omega, \mathcal{A}, P)$, and for each individual we observe a fuzzy datum. Each individual may belong to one of k different categories $g_1, \ldots, g_k$. As learning sample we have n independent individuals, the corresponding fuzzy data and categories. The goal is to find a rule allowing us to assign each new individual one of the k categories. For this purpose, our first criterion is a nonparametric density-based approach.

Formally, let $(\mathcal{X}, G) : \Omega \to \mathcal{F}_c(\mathbb{R}^p) \times \{1, \ldots, k\}$ be a random element such that $\mathcal{X}(\omega)$ is a fuzzy datum and $G(\omega)$ is the membership group (i.e., 1,...or k) of each individual $\omega \in \Omega$. Assume that we have n independent copies of $(\mathcal{X}, G)$ as training sample, that is, we have a simple random sample of size n, $\{(\mathcal{X}_i, G_i)\}_{i=1}^n$, from $(\mathcal{X}, G)$. As in a more general functional context (see [7]), we propose to nonparametrically estimate

$$P(G = g | \mathcal{X} = \widetilde{x})$$

for $g = 1, \ldots, k, \widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p)$, and then to assign the new data to the class of maximum estimated probability.

In order to find a reasonable estimator of the preceding probability, first consider $\delta > 0$ and a closed ball $B(\widetilde{x}; \delta)$ defined in terms of the metric introduced in Section 2. Assuming $P(X \in B(\widetilde{x}; \delta)) > 0$ Bayes Theorem implies

$$P(G = g | \mathcal{X} \in B(\widetilde{x}; \delta)) = \frac{P(D_\theta^\varphi(\mathcal{X}, \widetilde{x}) \leq \delta | G = g)) P(G = g)}{\sum_{l=1}^k P(D_\theta^\varphi(\mathcal{X}, \widetilde{x}) \leq \delta | G = l) P(G = l)}.$$

For each $\widetilde{x}$, $D_\theta^\varphi(\mathcal{X}, \widetilde{x})$ is a (real-valued) random variable. If $F_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}$ denotes the distribution function of this variable in group g, then,

$$P(G = g | \mathcal{X} \in B(\widetilde{x}; \delta)) = \frac{F_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}(\delta) P(G = g)}{\sum_{l=1}^k F_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=l}(\delta) P(G = l)}.$$

If we assume that $F_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}$ is absolutely continuous with a density (i.e., first derivative) continuous at 0 for all $g = 1, \ldots, k$ (as it is naturally supposed in the nonparametric setting), then there exists an underlying density function $f_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}$ with

$$\lim_{\delta \to 0} \frac{F_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}(\delta)}{\delta} = f_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}(0),$$

whence,

$$P(G = g | \mathcal{X} = \widetilde{x}) = \frac{f_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}(0) P(G = g)}{\sum_{l=1}^k f_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=l}(0) P(G = l)}. \tag{1}$$

It should be noted that in case $\mathcal{X}$ reduces to a real-valued random variable, $f_{D_\theta^\varphi(\mathcal{X}, \widetilde{x})|G=g}(0) = 2 f_{\mathcal{X}|G=g}(\widetilde{x})$ holds, where $f_{\mathcal{X}|G=g}$ is the density of $\mathcal{X}$ given the group g. Consequently, the distance-based classification criterion reduces in such a case to the well-known optimal criterion which assigns each new datum $\widetilde{x}$ to the class in g with highest density $f_{\mathcal{X}|G=g}(\widetilde{x}) P(G = g)$.

In case $\mathcal{X}$ is an $\mathbb{R}^p$-valued random vector ($p > 1$), due to the *curse of dimensionality* this second criterion cannot be applied in practice, even though it would be optimal. From a theoretical point of view, the situation is even more complex in infinite-dimensional situations like $\mathcal{F}_c(\mathbb{R}^p)$.

The denominator in (1) is just $f_{D_\theta^\varphi(\mathcal{X},\widetilde{x})}(0)$, for this reason, according to [7], we can estimate $P(G=g|\mathcal{X}=\widetilde{x})$ by means of

$$\widehat{P}(G=g|\mathcal{X}=\widetilde{x}) = \frac{\sum_{i=1}^n I_{G_i=g} K(h^{-1}D_\theta^\varphi(\mathcal{X}_i,\widetilde{x}))}{\sum_{i=1}^n K(h^{-1}D_\theta^\varphi(\mathcal{X}_i,\widetilde{x}))} = \frac{\left[(n_g h)^{-1}\sum_{i\in N_g} K(h^{-1}D_\theta^\varphi(\mathcal{X}_i,\widetilde{x}))\right]\cdot(n_g n^{-1})}{(nh)^{-1}\sum_{i=1}^n K(h^{-1}D_\theta^\varphi(\mathcal{X}_i,\widetilde{x}))},$$

where I denotes the indicator function, $N_g=\{i\in\{1,\ldots,n\}|G_i=g\}$, n_g is the cardinality of N_g, K is a kernel and h is the bandwidth.

The last expression shows that the estimator depends on an estimate of $f_{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=g}(0)$ (the densities within each group at point 0), the empirical estimate of $P(G=g)$ and an estimate of $f_{D_\theta^\varphi(\mathcal{X},\widetilde{x})}(0)$ (the overall density at point 0). In this approach the bandwidth is the same for all the estimates, and it is determined by using the full sample. Hence, in particular we have

$$\sum_{g=1}^k \widehat{P}(G=g|\mathcal{X}=\widetilde{x}) = 1.$$

However, since the real-valued random variable $D_\theta^\varphi(\mathcal{X},\widetilde{x})$ is a mixture of the variables $\{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=g\}_{g=1}^k$ whose distributions are expected to be different (recall that the aim is to discriminate among them), it seems more convenient to estimate each distribution separately (each one of them with its individual bandwidth), and then to combine them to estimate $f_{D_\theta^\varphi(\mathcal{X},\widetilde{x})}(0)$. Thus, we propose to estimate $f_{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=g}(0)$ by means of

$$\widehat{f}_{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=g}(0) = (n_g h_g)^{-1}\sum_{i\in N_g} K(h_g^{-1}D_\theta^\varphi(\mathcal{X},\widetilde{x}))$$

for all $g=1,\ldots,k$. In this way, we have

$$\widehat{P}(G=g|\mathcal{X}=\widetilde{x}) = \frac{n_g n^{-1}\widehat{f}_{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=g}(0)}{\sum_{l=1}^k (n_l n^{-1})\widehat{f}_{D_\theta^\varphi(\mathcal{X},\widetilde{x})|G=l}(0)}, \tag{2}$$

as well as

$$\sum_{g=1}^k \widehat{P}(G=g|\mathcal{X}=\widetilde{x}) = 1,$$

whenever (2) is well-defined, i.e. the denominator is not zero. Note that, as in the real case, this may happen in singular cases when the underlying densities are separated, and kernels with bounded support and small bandwidths are considered.

The considered estimators are statistically consistent under the unique assumption that the corresponding densities exist and are continuous for each $\widetilde{x}\in\mathcal{F}_c(\mathbb{R}^p)$ (plus the usual regularity conditions); however, there is no assumption about the changes in the distributions as $\widetilde{x}$ varies.

To summarize, consider a training sample $\{(\mathcal{X}_i, G_i)\}_{i\in\{1,\ldots,n\}}$ with $\mathcal{X}_i$ a fuzzy datum (in $\mathcal{F}_c(\mathbb{R}^p)$) and G_i the corresponding membership group (in $\{1,\ldots,k\}$). For each $g\in\{1,\ldots,k\}$ let n_g be the number of observations for which $G_i=g$ and denote by $\{\mathcal{Y}_{j,g}\}_{j=1}^{n_g}$ the corresponding collection of such fuzzy data (that is, the conditional samples). Given a fuzzy value $\widetilde{x}\in\mathcal{F}_c(\mathbb{R}^p)$, the proposed Density-based Classification Criteria for Fuzzy Data (DCCF) can be summarized by the following algorithm:

DCCF Algorithm

Step 1. Compute the distance between the datum $\widetilde{x}$ to be classified and the set of training fuzzy data, that is,

$$d_{j,g} = D_\theta^\varphi(\widetilde{x},\mathcal{Y}_{j,g}) \quad \text{for all } j\in 1,\ldots,n_g \text{ and all } g\in 1,\ldots,k.$$

Step 2. For each $g\in\{1,\ldots,k\}$ fix a bandwidth h_g associated with the real random sample $\{d_{j,g}\}_{j=1}^{n_g}$.

Step 3. Estimate the membership probabilities $p_g = P(G = g|\mathcal{X} = \widetilde{x})$ by means of

$$\widehat{p}_g = \frac{\sum_{j=1}^{n_g} K(h_g^{-1} d_{j,g})}{\sum_{g=1}^{k} \sum_{j=1}^{n_g} K(h_g^{-1} d_{j,g})}. \tag{3}$$

Step 4. Assign $\widetilde{x}$ to the group $g(\widetilde{x}) \in \{1, \ldots, k\}$ of maximum estimated probability, i.e. the group satisfying

$$\widehat{p}_{g(\widetilde{x})} = \max_{g \in \{1,\ldots,k\}} \hat{p_g}.$$

DCCF can be applied with classification purposes even if data were not collected by a random sampling method. In this case, it would not be possible to ensure that the computed values $\widehat{p_g}$ are good estimates of the corresponding conditional probabilities, but the method would provide a reasonable classification rule.

As it happens in the real-valued case, it is possible that in few situations either the maximum in Step 4 is attained at several groups or (3) is not well-defined (as it was above-mentioned). In these cases, alternative/complementary criteria could be applied.

The second step in the DCCF Algorithm is very important, because the accuracy in a nonparametric estimation of the density is highly related to the determination of the bandwidth. There are different methods in the literature, some of them to be analyzed in Section 5.

As it is well-known, the optimal bandwidth for the estimation of a density function is not appropriate in general for the estimation of its curvature. In the same way, the optimal bandwidth for the estimation of a density function is not necessarily appropriate for the classification problem. Consequently one could try to optimize the bandwidth selection for the classification problem by choosing, for instance, the values maximizing the proportion of right classifications. Nevertheless this approach is unfeasible in practice, because the bandwidths to be chosen (one per group) are specific for the new datum to be classified. Thus, it is not possible to measure the classification error of the new datum without using information about its membership group (which is going to be forecasted).

One of the well-known drawbacks of the nonparametric kernel density estimation is the need for moderate/large sample sizes. If data of a given group are scarce, then it is difficult to get accurate nonparametric density estimates. The problem of estimating the density in the boundary of its support is even harder. Furthermore, as pointed out in previous comments, the density-based criterion is not optimal in general, so there is still room for improvements.

4. Ball-based Classification Criteria for Fuzzy Data (BCCF)

In this section we will propose an alternative and simpler approach by essentially avoiding density estimation in the previous approach. Instead of focusing the classification technique on estimating the conditional probabilities

$$\{P(G = g|\mathcal{X} = \widetilde{x})\}_{g=1}^{k} \quad \text{with } \widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p),$$

we suggest a classification based on the quantities

$$\{P(G = g|\mathcal{X} \in B(\widetilde{x}; \delta))\}_{g=1}^{k}$$

for a value $\delta > 0$ to be chosen. On the one hand, this simplification could entail a loss of accuracy that will be analyzed in Section 5. On the other hand, it can be applied in a more general setting than DCCF, because no assumption about the existence of the conditional densities is made. Note that if all the conditional densities exist, then BCCF may be formally expressed as a special case of DCCF with a uniform kernel on $[0, 1]$ and $h = \delta$.

Consider $\delta > 0$ and $g \in \{1, \ldots, k\}$. A natural estimator for

$$P(D_\theta^\varphi(\mathcal{X}, \widetilde{x}) \leq \delta|G = g)$$

based on the sample information is

$$\widehat{P}(D_\theta^\varphi(\mathcal{X}, \widetilde{x}) \leq \delta|G = g) = \frac{n_{\delta,g}}{n_g},$$

where $n_{\delta,g}$ is the number of observations in the sample belonging to group g and for which the distance to $\widetilde{x}$ is lower than or equal to δ. Thus, by taking into account (1), $P(G = g|\mathcal{X} \in B(\widetilde{x}; \delta))$ can nonparametrically be estimated as follows:

$$\widehat{P}(G = g|\mathcal{X} \in B(\widetilde{x}; \delta)) = \frac{\frac{n_{\delta,g}}{n_g}\frac{n_g}{n}}{\sum_{l=1}^{k} \frac{n_{\delta,l}}{n_l}\frac{n_l}{n}} = \frac{n_{\delta,g}}{\sum_{l=1}^{k} n_{\delta,l}}.$$

Following the notation in the preceding section, given $\widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p)$, the suggested Ball-based Classification Criteria for Fuzzy Data (BCCF) can be summarized in the following algorithm:

BCCF Algorithm

Step 1. Compute the distance between the datum $\widetilde{x}$ to be classified and the set of training fuzzy data, that is,

$$d_{j,g} = D_\theta^\varphi(\widetilde{x}, \mathcal{Y}_{j,g}) \quad \text{for all } j \in 1, \ldots, n_g \text{ and all } g \in \{1, \ldots, k\}.$$

Step 2. Fix a value for $\delta > 0$ and for each $g \in \{1, \ldots, k\}$ compute

$$n_{\delta,g} = \sum_{j=1}^{n_g} I_{[0,\delta]}(d_{j,g}).$$

Step 3. Estimate the membership probabilities $p_g = P(G = g | \mathcal{X} \in B(\widetilde{x}; \delta))$ by means of

$$\widehat{P}(G = g | \mathcal{X} \in B(\widetilde{x}; \delta)) = \frac{n_{\delta,g}}{\sum_{l=1}^{k} n_{\delta,l}}.$$

Step 4. Assign $\widetilde{x}$ to the group $g(\widetilde{x}) \in \{1, \ldots, k\}$ of maximum estimated probability.

One of the essential issues of this approach is to select a suitable δ. In Section 5 this problem will be discussed, and two different estimates for δ will be proposed and compared.

5. Empirical results

In this section the performance of the classification methods for fuzzy data proposed in this manuscript is going to be analyzed on the basis of two different datasets. Several alternatives for estimating the densities involved in method DCCF, as well as for selecting an appropriate value of δ in method BCCF, will be considered.

The fuzzy data in both datasets reduce to trapezoidal fuzzy numbers which are characterized by 4-dimensional real vectors. Indeed, each trapezoidal fuzzy number $A \in \mathcal{F}_c(\mathbb{R})$ is characterized by $(\inf A_0, \inf A_1, \sup A_1, \sup A_0)$. Thus, any classification technique for multivariate data could also be applied. It should be underlined that the multivariate techniques do not take the order relationships inherited from the shape of the fuzzy sets into account. They are, nevertheless, feasible and reasonable tools available for comparison. Consequently we have considered the well-known linear discriminant analysis (LDA) as a benchmark.

5.1. Perceptions experiment

In an experiment about the visual perception of the length of a line segment relative to a maximum, we have shown images like those in Fig. 1 to different people. The aim was to express the length of the light line (below) relative to length of the dark line (above). Thereby people were allowed to express their lack of precision/their uncertainty by using a fuzzy scale. After a trial with the software designed for this purpose, people have been shown different line segments with random length and they have described their perception of the length in the fuzzy scale and the linguistic label (very small, small, medium, large, very large) they consider suitable.

The way of using the fuzzy scale was explained in the software as follows:

"*This experiment regards your perception about the relative length of different lines.*

On the top of the screen, we have plotted in dark color the longest line that we could show to you. This line will remain visible in the current position during all the experiment, so that you can always have a reference of the maximum length.

At each trial of the experiment we will show you a light line and you will be asked about its relative length (in comparison with the length of the reference dark line):

- *Firstly you will be asked for a linguistic descriptor of the relative length. We have considered five descriptors (very small; small; medium; large; very large). The aim is to select one of these descriptors at first sign (you can change it later if you want to).*
- *Secondly you will be asked for your own estimate or perception (without physically measuring it) of the relative length (in percentage) by means of a fuzzy set (the information about the design and interpretation of the fuzzy set will be shown to you at this time).*
- *Finally, in case your initial perception had been changed during the process you can readjust again the linguistic descriptor of its relative length."*

When the image like the one shown in Fig. 1 was shown the respondents were asked to draw a fuzzy set representing their perceptions. In accordance with the usual interpretation, the respondents had to choose the 0-level (set of all those points with a positive degree of membership) as the interval of all values that they considered to be compatible with the length

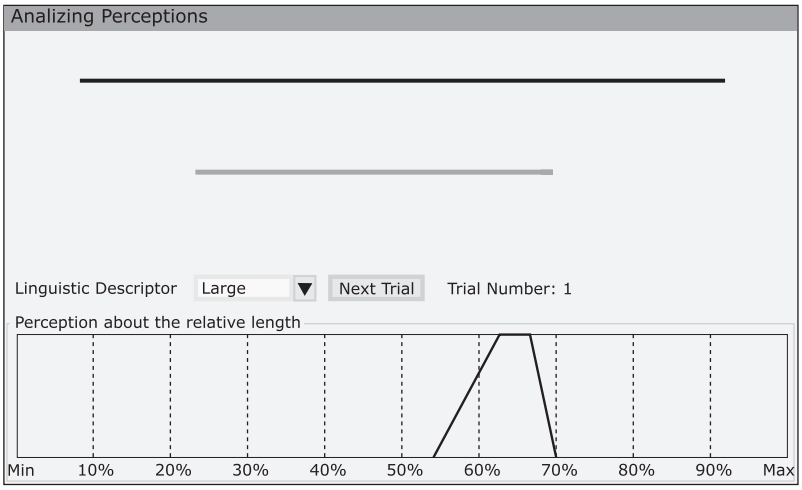

Fig. 1. Software to evaluate the visual perception of a line segment.

Table 1
Perceptions about the relative length of the light line segment.

Trial	inf P_0	inf P_1	sup P_1	sup P_0	Ling. descrip.
1	78.27	80.94	84.41	87.40	Large
2	54.93	58.00	62.20	65.67	Large
3	47.25	49.43	50.89	53.31	Medium
4	92.65	95.72	97.58	99.11	Very large
5	12.92	15.51	17.77	20.03	Very small
6	32.55	36.03	39.90	42.89	Small
7	2.50	4.44	6.22	9.21	Very small
8	24.80	28.19	30.45	33.28	Small
9	55.17	58.40	61.79	65.75	Large
10	2.26	3.63	5.57	8.08	Very small

to a greater or lesser extent (i.e., length not outside this interval). In the same way, the 1-level (set of all those points with maximum degree of membership) had to be an interval which they considered completely compatible with their perception about the measure of the dark line. Although it was possible to change the shape of the resulting fuzzy sets, by default the trapezoidal fuzzy set formed by the interpolation of both intervals was chosen (as shown in Fig. 1). Guidelines on collecting fuzzy data from random experiments as well as a detailed description of the Software and associated experiments can be found in [12].

Table 1 contains some of the data a person making 551 trials delivered. The complete dataset, as well as the software *Perceptions* providing it, can be found at http://bellman.ciencias.uniovi.es/SMIRE/perceptions.html (web page of the SMIRE research group).

The goal here is to predict the category (very small, small, medium, large or very large) that this person considers correct from the fuzzy perception that he/she has about the length of the light line. Note that in this case, there is an underlying (ill-known) precise quantity: the length of the light line. Nevertheless, the existence of that precise quantity is irrelevant for the stated goal, because the objective is to classify the fuzzy sets representing the perceptions in different groups, irrespectively of the real length that people observe. To classify the real length would require, of course, a different approach. It should be also remarked that the categories are treated here simply as different classes, which may be also labeled as 1, 2, 3, 4 and 5, irrespectively of the fuzzy representation that they might have. Considering fuzzy labels would lead to a different approach.

The line shown at each trial has been chosen at random – in order to obtain a good coverage of some interesting situations the precise generation procedure was the following one

- 479 lengths were generated from a uniform distribution on [0, 100].
- 9 lengths in the equally spaced discrete set $\{100/27 + (i/8)100(1 - 2/27)\}_{i=0,\ldots,8}$ were repeated 6 times. Thus, we had 54 lengths that are representative of quite different situations that may arise.
- All the random lengths were swapped and shown at random.

As explained in Section 3, the density of the random variable $D_\theta^\varphi(\mathcal{X}, \widetilde{x})$ at point 0 for any fuzzy response $\widetilde{x}$ should be estimated. The problem of estimating a density in the boundary of its support is not an easy task. Several approaches can be found in the literature (see, for instance, [18,15,13]), however, none of them has solved all the inconveniences that may arise. Consequently, we have considered the following three options to evaluate the performance

DCCF1: Firstly, kernel density estimation with the Quartic Kernel

$$K_1(u) = \frac{15}{16}(1-u^2)^2 I_{(|u|\leq 1)}$$

and bandwidth chosen by cross-validation was considered. It is well-known that this estimator is not consistent in the boundaries. In particular the estimation at 0 for this kernel is suboptimal in general (see, for instance, [18]). As we have mentioned before, several boundary corrections to overcome this problem are available. Many of these corrections (essentially based on jacknife) lead to consistent estimators at the boundaries, although taking negative values in many practical situations (which makes them useless in our case). There are some exceptions in the literature that yield both consistent and non-negative estimates.

DCCF2: As a second scenario we followed the proposal in [7] of using an asymmetric kernel. To be precise, we consider the Quartic kernel restricted to $u \geq 0$, K_1, that is,

$$K_2(u) = \frac{15}{8}(1-u^2)^2 I_{(u\in[0,1])}.$$

This approach leads to consistent estimates at the boundary (but with a low convergence speed). There are no studies about the optimal way of choosing the bandwidth for this asymmetric kernel, so we used the common one provided by cross-validation applied to this particular case.

DCCF3: Finally, as a third strategy we considered the estimator proposed by Jones and Foster [16] which consists in a kernel density estimation with boundary correction that provides non-negative estimates, that is,

$$f_P(0) = \bar{f}(0)\exp\left\{\frac{\hat{f}(0)}{\bar{f}(0)} - 1\right\},$$

where $\hat{f}(0)$ denotes the basic kernel density estimator divided by $a_0(0) = \int_{-1}^{0} K(u)du$ and $\bar{f}$ is the boundary corrected kernel density estimator based on generalized jacknifing by combining $\hat{f}$ with the analogous estimator based on the kernel $L(u) = 0.5K(2u)$. In this case an optimal (local) bandwidth is the solution of a complex variational problem (see [18] for a similar situation), which, in general, is not possible to solve. For this reason we have considered the global cross-validation bandwidth obtained in M1 for the kernel $K = K_1$.

As mentioned before, in the preceding cases the usage of a cross-validation technique for selecting the bandwidth is not theoretically justified. As simple alternative to the cross-validation method, we have also used the optimal AMISE bandwidth (that is, the bandwidth minimizing the Asymptotic Mean Integrated Squared Error assuming normality). Obviously the involved distributions cannot be Gaussian, and thus neither can this alternative be theoretically justified. Nevertheless, it is well-known that projections on high-dimensional spaces (although almost surely non-normal) are quite close to normality. Given that the considered distance is just an average of projections, it is expected that the distance-based variables are close to normality, and thus the proposed AMISE method could be accurate.

On the other hand, in order to apply the BCCF Algorithm we have to select an appropriate value for δ. As there are no studies concerning this selection, we have considered the following ones for comparative purposes:

BCCF1: In this first case, δ was chosen to be the maximum of the sample deviations in each group (trying to preserve the simplicity of this method). The reason for considering the maximum instead, for instance, the minimum is to try to ensure that the balls will be large enough for containing data points of at least one group.

BCCF2: In contrast to the above-mentioned simple approach, a more elaborate selection was considered. Namely δ was chosen as the value maximizing an accuracy measure, concretely the 10-random-3-fold (within each group) classification accuracy (see the details below).

In all situations φ was chosen to be the Lebesgue measure on $[0, 1]$ and $\theta = 1/3$ (i.e., the corresponding value for Bertoluzza et al. [1] distance assigning constant weight to all squared distances between the corresponding convex linear combinations within the α-levels).

In order to estimate the right classification percentage for each of the methods, we have considered a 100-random-10-fold cross validation within each group. That is, the sample corresponding to each class was split at random in 10 parts (subfolds) of approximately the same sample size. First, subfolds of all classes were grouped together in order to compose the first fold and so on. As a consequence, the whole data set was split in 10 folds of approximately the same sample size, keeping the proportion of observations of each class in each fold approximately equal to the original proportion in the whole sample. Each fold was selected as validation sample whereas the observations which were not included in this fold were used as training sample. Following this procedure, each data point in the sample was classified and the proportion of right classifications was computed. Finally, in order to avoid the dependency on the 10 particular selected folds the whole procedure was repeated 100 times by randomly selecting different fold composition. The classification accuracy employed in method BCCF2 is defined in the same way but with 10 replications (instead of 100) and 3 folds (instead of 100).

Table 2 shows a summary of the percentage of correct classifications in the 100 replications corresponding to each one of the considered methods. For comparative purposes the corresponding boxplots are depicted in Fig. 2.

Table 2
Summary of the percentage of correct classification with different methods in the different replicas.

	LDA	BCCF1–2	Cross-validation DCCF1–2–3	AMISE bandwidth DCCF1–2–3
Minimum	89.11	90.20–89.11	88.75–89.29–85.66	90.02–90.02–87.84
Median	90.20	90.74–90.56	90.20–90.38–87.57	90.93–90.93–89.29
Mean	90.14	90.74–90.50	90.19–90.32–87.53	90.82–90.82–89.27
Maximum	91.11	91.29–91.47	91.29–91.83–88.93	91.47–91.47–90.38
Deviation	0.36	0.21–0.44	0.46–0.46–0.61	0.34–0.34–0.55

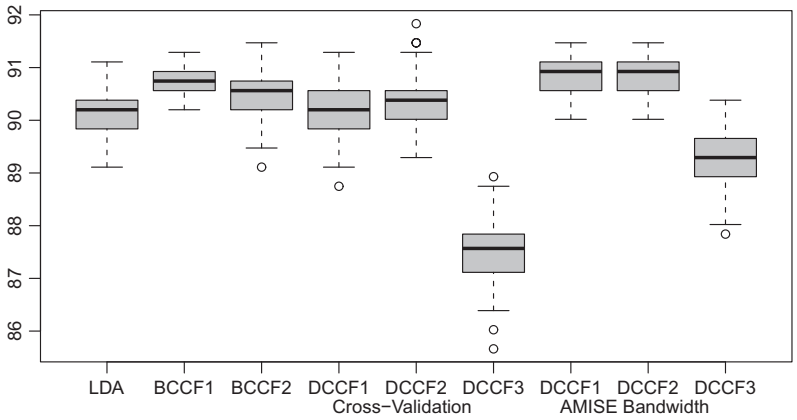

Fig. 2. Boxplots of the percentage of correct classifications in the 100 replicas of 10-fold cross validation for the different methods.

Fig. 2 shows an overall improvement of the percentage of correct classification (both in mean and deviation) for DCCF methods with AMISE bandwidth with respect to cross-validation bandwidth. The improvement is particularly remarkable in the case of method DCCF3. In general, a better behavior for DCCF1 and DCCF2 than for DCCF3 is shown. This fact points out the inappropriateness of the considered cross-validation bandwidth selection procedure in this case. It should be noted that, when AMISE bandwidth method is used, the behavior of DCCF1 and DCCF2 is identical, because exactly the same bandwidth is used in both cases. With a common bandwidth, both cases reduce to the same expression, although DCCF1 is theoretically inconsistent. In general, these facts indicate that the selection of a suitable bandwidth is more important than the theoretical accuracy of the considered estimators (DCCF1 and DCCF2 are, from a theoretical point of view, worse than DCCF3, provided that the optimal bandwidth is considered). To summarize, from the accuracy viewpoint, DCCF1 and DCCF2 with the AMISE bandwidth are the best ones among all the DCCF methods.

Concerning the BCCF methods, the behavior is better when δ is chosen according to the maximum of in-group deviations method. In this case, the mean percentage of correct classifications is quite similar, but the variability for the optimization criterion is twice the variability for the maximum of the deviations method. This fact is quite surprising, as no improvement is obtained by applying the optimization criterion. The main problem seems to be connected with the flatness of the optimization function, which introduces additional variability in the method, but with no improvement.

We can also observe a smaller variability in the 100 replicas for BCCF1 than for the DCCF methods. This means that BCCF1 has a more stable behavior than DCCF, that is, the classification obtained with the BCCF1 method is less affected by the concrete data set than in case of DCCF methods. This is probably due to the dependence of DCCF methods on the chosen bandwidth, and the high sensibility of bandwidth selection methods. Moreover, BCCF1 has a very similar mean accuracy as the DCCF methods. This situation is also reflected in Table 2.

Finally, the Linear Discriminant Analysis method has a quite good behavior in this particular case. Concerning the mean, its accuracy is slightly worse than BCCF1 and DCCF1–2 with AMISE bandwidth, and concerning variability it is comparable with that of the last two methods.

5.2. Flood analysis

A pilot study was carried out by the Institute of Natural Resources and Zoning (INDUROT) of the University of Oviedo in order to assess the flood return period of different flood plains along an Asturian river. The flood frequency of each sampled flood plain was classified by expert criteria as high, medium or low according to the geomorphological and historical information collected.

One of the main variables related to the flood frequency is the height of the flood plain, which is usually quantified by means of Digital Elevation Models (DEMs) with precisions up to a millimeter. Nevertheless, the accuracy of the classification based on DEM heights is partially lost when high and medium classes are considered probably because the usual abundance of vegetation in these areas involves a lower reliability of DEM measurements.

Table 3
Summary of the percentage of correct classification for the flood analysis.

	LDA	BCCF1–2	Cross-validation DCCF1–2–3	AMISE bandwidth DCCF1–2–3
Minimum	61.54	71.79–66.67	69.23–71.79–66.67	69.23–69.23–71.79
Median	79.49	84.62–84.62	84.62–87.18–82.05	87.18–87.18–87.18
Mean	79.56	83.28–83.44	84.41–85.62–81.61	85.59–85.59–87.36
Maximum	92.31	89.74–97.44	94.87–94.87–92.31	94.87–94.87–94.87
Deviation	5.01	2.62–4.58	4.81–4.02–4.56	4.49–4.49–3.80

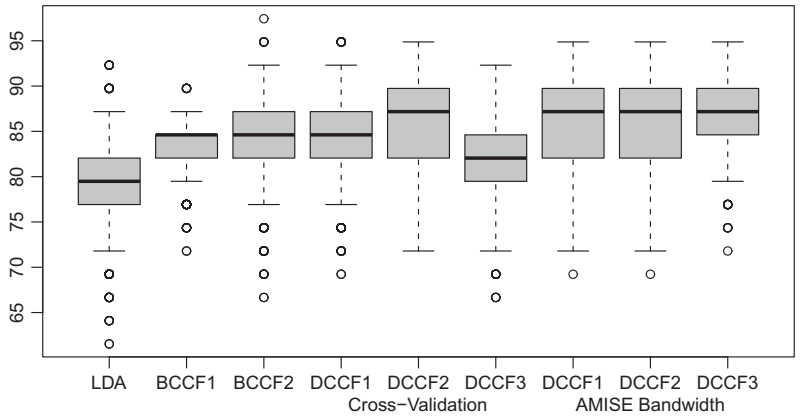

Fig. 3. Boxplots of the percentage of correct classification in the 1000 replicas of 10-fold cross validation for the Flood analysis.

An exact measurement of the height during the field work is very expensive. On the contrary a simple visual inspection is cheap and can be easily done by using a similar approach to the one proposed in the perceptions experiment. In order to analyze the quality of this approach, trapezoidal fuzzy numbers representing the perception about the height of each flood area where collected during the mentioned pilot study. The complete dataset can be found at http://bellman.ciencias.uniovi.es/SMIRE/floods.html.

The goal is to predict the flood frequency category (high, medium or low) from the fuzzy perception about height of the flood area. As in the previous case, the height is an underlying (ill-known) precise quantity, but the precise value is not related to the pursued aim.

Being a pilot study the sample sizes are small (9 flooding plains with low flooding frequency, 10 with medium and 20 with height). We decided to estimate the percentage of right classification by means of a 1000-random-3-fold cross validation (instead of 10 folds as in the previous experiment, that would be unfeasible, and using 1000 replications to compensate the smaller size of the folds). In this way each fold consists of at least three observations. In Table 3 we show a summary of the percentage of right classifications in the 100 replications corresponding to each one of the considered methods. The corresponding boxplots are also depicted in Fig. 3.

First of all, for the DCCF methods the AMISE Bandwidth criterion outperforms the cross-validation and shows a better mean classification accuracy and smaller variability. In the case of DCCF3, the improvement is really important. Consequently, the AMISE Bandwidth criterion seems recommendable. Among the three proposed density estimation techniques, the boundary corrected estimator proposed in [16] (DCCF3) is preferable.

With respect to BCCF methods, the mean classification accuracy for both techniques for selecting δ are similar. Nevertheless there is again an important difference regarding the variability, which is almost twice as large when using the optimization criterion as compared with the the maximum of deviation method. Thus, the recommendation is again to use the simpler method (BCCF1). In this scenario, the mean accuracy of DCCF3 with AMISE Bandwidth is better than the one provided by BCCF1, the last one being the best classification criteria for this data set.

Finally, it has to be remarked that in this case, the Linear Discriminant Analysis technique performs worse than the proposed methods. Nevertheless, the situation might be the contrary for other datasets, because in high-dimensional spaces the existence of a uniformly best classifier cannot be assured. Summing up, for this dataset DCCF3 with AMISE Bandwidth estimation is the best technique.

6. Concluding remarks

This work is an initial study concerning the problem of supervised classification of random fuzzy sets. We propose to use two nonparametric approaches to better capture nonlinear boundaries. However, other interesting viewpoints may be used

(either by extending/adopting those from functional data analysis, as the penalized or flexible discriminant analyses, or by developing *ad-hoc* procedures). Further theoretical and empirical comparative studies should be developed.

As it was mentioned in Section 5, the method inspired directly by functional data analysis entails a nonparametric estimation problem which is not easy to solve, and for which there is no general optimal method. In this respect, it seems that the AMISE selection criteria under normality assumption is a promising technique for this particular context.

Finally, it would be very interesting to consider the case in which the group membership of the training data is imprecise, as it was done in [4] for real-valued learning data (see also [5]).

Acknowledgments

This research has been partially supported by the Grants from the Spanish Ministry of Education and Science MTM2009-09440-C02-01 and MTM2009-09440-C02-02, the Grants from the Principality of Asturias IB09-042C1 and IB09-042C2 as well as by the European COST Action IC0702. This financial support is gratefully acknowledged. The authors also want to thank the anonymous reviewers and the guest editor for their helpful comments.

References

[1] C. Bertoluzza, N. Corral, A. Salas, On a new class of distances between fuzzy numbers, Mathware & Soft Computing 2 (1995) 71–84.
[2] A. Colubi, J.S. Domínguez-Menchero, M. López-Díaz, D.A. Ralescu, On the formalization of random fuzzy sets, Information Sciences 133 (2001) 3–6.
[3] A. Colubi, Statistical inference about the means of random fuzzy sets: applications to the analysis of fuzzy- and real-valued data, Fuzzy Sets and Systems 160 (3) (2009) 344–356.
[4] E. Côme, L. Oukhellou, T. Denœux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognition 42 (3) (2009) 334–348.
[5] T. Denœux, L.M. Zouhal, Handling possibilistic labels in pattern classification using evidential reasoning, Fuzzy Sets and Systems 122 (3) (2001) 47–62.
[6] T. Denœux, M.-H. Masson, P.-A. Hébert, Nonparametric rank-based statistics and significance tests for fuzzy data, Fuzzy Sets and Systems 153 (1) (2005) 1–28.
[7] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric functional approach, Computational Statistics & Data Analysis 44 (2003) 161–173.
[8] J.L. Flórez, A. Colubi, G. González-Rodríguez, M.A. Gil, Nonparametric regression with random fuzzy sets through the support functions, in: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty (IPMU'2006, Paris), 2006, pp. 724–730.
[9] W.G. González-Manteiga, P. Vieu, Statistics for functional data, Computational Statistics & Data Analysis 51 (10) (2007) 4788–4792.
[10] G. González-Rodríguez, A. Colubi, W. Trutschnig, Simulation of random fuzzy sets , Information Sciences 179 (5) (2009) 642–653.
[11] G. González-Rodríguez, A. Colubi, P. D'Urso, M. Montenegro, Multi-sample test-based clustering for random fuzzy sets, International Journal of Approximate Reasoning 50 (2009) 721–731.
[12] G. González-Rodríguez, A. Colubi, M.A. Gil, Fuzzy data treated as functional data: a one-way ANOVA test approach, Computational Statistics & Data Analysis, in press, doi:10.1016/j.csda.2010.06.013.
[13] P. Hall, B.U. Park, New methods for bias correction at endpoints and boundaries, Annals of Statistics 30 (5) (2002) 1460–1479.
[14] P. Hall, H.-G. Müller, J.-L. Wang, Properties of principal component methods for functional and longitudinal data analysis, Annals of Statistics 34 (3) (2006) 1493–1517.
[15] M.C. Jones, P.J. Foster, Generalized jackknifing and higher order kernels, Journal of Nonparametric Statistics 3 (1993) 81–94.
[16] M.C. Jones, P.J. Foster, A simple nonnegative boundary correction method for kernel density estimation, Statistica Sinica 6 (1996) 1005–1013.
[17] R. Kruse, K.D. Meyer, Statistics with Vague Data, D. Reidel Publishing Company, 1987.
[18] H.-G. Müller, Smooth optimum kernel estimators near endpoints, Biometrika 78 (1991) 521–530.
[19] M.L. Puri, D.A. Ralescu, Fuzzy random variables, Journal of Mathematical Analysis and Applications 114 (1986) 409–422.
[20] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis. Methods and Case Studies, in: Springer Series in Statistics, Springer-Verlag, New York, 2002.
[21] W. Trutschnig, G. González-Rodríguez, A. Colubi, M.A. Gil, A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread, Information Sciences 179 (2009) 3964–3972.
[22] R. Yang, Z. Wang, PA. Heng, L. Kwong-Sak, Classification of heterogeneous fuzzy data by Choquet integral with fuzzy-valued integrand, IEEE Transactions on Fuzzy Systems 15 (5) (2007) 931–942.
[23] J.-T. Zhang, J. Chen, Statistical inferences for functional data, Annals of Statistics 35 (3) (2007) 1052–1079.