Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Fuzzy Sets and Systems ●●● (●●●●) ●●●–●●●
FUZZY sets and systems
www.elsevier.com/locate/fss

# A fuzzy clustering procedure for random fuzzy sets

Paolo Giordani [a], Ana Belén Ramos-Guajardo [b]

[a] *Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy*
[b] *Dpto de Estadística e I.O., Universidad de Oviedo, Spain*

Received 6 November 2014; received in revised form 5 February 2016; accepted 6 February 2016

**Abstract**

A fuzzy clustering method for random fuzzy sets is proposed. The starting point is a $p$-value matrix with elements obtained by comparing the expected values of random fuzzy sets by means of a bootstrap test. As such, the $p$-value matrix can be viewed as a relational data matrix since the $p$-values represent a kind of similarity between random fuzzy sets. For this reason, in order to cluster random fuzzy sets, fuzzy clustering techniques for relational data can be applied. In this context, the so-called NE-FRC algorithm is considered. One of the most important advantages of the NE-FRC is that the relational data could not be derived from Euclidean distances. Some simulations are presented to show the behavior of the proposed procedure and two applications to real-life situations are also included.

*Keywords:* Random fuzzy sets; Fuzzy clustering; Relational data

## 1. Introduction

The available information is usually affected by various sources of uncertainty. In the case of randomness, probability theory is the best tool for handling such a kind of uncertainty and the concept of random variable naturally arises. Fuzzy set theory [36] plays a relevant role in the presence of imprecision. If data are jointly random and imprecise, then the notion of Random Fuzzy Set (RFS for short) can be considered (see, e.g., Puri and Ralescu [23]). This is the case when fuzzy-valued observations of independent RFS's are registered on a sample of statistical units.

In the literature, several fuzzy clustering methods for fuzzy data have been proposed. Refer to, e.g., D'Urso [8], Coppi et al. [7], Hung and Yang [16] and the references therein. Nonetheless, as far as we know, no attention has been paid to the clustering problem of RFS's except for González-Rodríguez et al. [11] who developed a (non-fuzzy) hierarchical clustering procedure for RFS's. Their idea is to detect clusters of RFS's such that the RFS's in each group have the same population expected value. This goal is achieved by taking into account both imprecision and randomness in the clustering process by using the $p$-values of a multi-sample test for the expectations of RFS's (see,

*E-mail addresses:* paolo.giordani@uniroma1.it (P. Giordani), ramosana@uniovi.es (A.B. Ramos-Guajardo).

http://dx.doi.org/10.1016/j.fss.2016.02.006
0165-0114/© 2016 Published by Elsevier B.V.

for instance, Gil et al. [10] and González-Rodríguez et al. [12]). The higher the $p$-value, the more similar are the expectations between two RFS's. As highlighted by González-Rodríguez et al. [11], the $p$-values represent measures of similarities between RFS's. The so-obtained similarities consider both the fuzziness and the randomness. This is so because the adopted statistical test is able to handle data affected by fuzziness and the resulting $p$-value expresses a degree of similarity between two (random) expectations of RFS's.

In this work, we develop a non-hierarchical fuzzy clustering procedure for RFS's based on a $p$-value matrix. For this purpose, we apply fuzzy clustering algorithms for relational data, i.e., data coming from measures of similarity/dissimilarity either computed objectively according to any metric or based on subjective knowledge. Several proposals for clustering relational data can be found in the literature. See, e.g., Davé and Sen [9], Runkler [30], Khalilia et al. [17] and the references therein. As we shall see, we suggest a novel procedure for clustering RFS's using the algorithm introduced by Davé and Sen [9] applied to a $p$-value matrix rather than to a dissimilarity matrix. Differently from González-Rodríguez et al. [11], our proposal is much more efficient from a computational point of view and produces a fuzzy partition of the RFS's. Therefore, the RFS's are assigned to the clusters according to the so-called membership degrees ranging from 0 (complete non-membership) to 1 (complete membership). This is very welcome because it appears to be limited to get non-fuzzy clusters using fuzzy information.

The paper is organized as follows. In the next section some preliminaries on RFS's are given. In Section 3 the multi-sample test and the (non-fuzzy) hierarchical clustering algorithm for RFS's introduced by González-Rodríguez et al. [11] are recalled. Section 4 contains a review of fuzzy clustering algorithms for relational data. Section 5 is devoted to the proposed fuzzy non-hierarchical clustering algorithm for RFS's. The results of a simulation study are discussed in Section 6 where the empirical performance of the proposal also in comparison with González-Rodríguez et al. [11] and the ability of some cluster validity indices to detect the known number of clusters are investigated. Section 7 contains two real-life applications. Finally, concluding remarks are provided in Section 8.

## 2. Preliminaries

The space $\mathcal{F}_c(\mathbb{R})$ of fuzzy numbers to be considered includes the mappings $U : \mathbb{R} \to [0, 1]$ such that for each $\alpha \in (0, 1]$ the so-called $\alpha$-level set (or $\alpha$-cut) $U_\alpha = \{x \in \mathbb{R} | U(x) \geq \alpha\}$ belongs to the class of nonempty compact intervals in $\mathbb{R}$ (denoted by $\mathcal{K}_c(\mathbb{R})$). The 0-level, $U_0$, is the closure of the support of $U$.

The usual arithmetic between fuzzy values consists in the sum and the product by a scalar [36,22], which is an extension of the corresponding ones for intervals paying attention to the fuzzy meaning. Thus, if $U, V \in \mathcal{F}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, $U + \lambda V$ can be defined so that for each $\alpha \in [0, 1]$

$$(U + \lambda V)_\alpha = U_\alpha + \lambda V_\alpha = \{u + \lambda v : u \in U_\alpha, v \in V_\alpha\}. \tag{1}$$

The arithmetic is non-linear due to the lack of symmetric element w.r.t. the Minkowski addition although $(\mathcal{F}_c(\mathbb{R}), +, \cdot)$ has a semilinear-conical structure since the addition extends level-wise the Minkowski sum of sets.

The metric used in the formalizations is based on the concepts of midpoint and spread (or radius) of the interval $U_\alpha \in \mathcal{K}_c(\mathbb{R})$, denoted as $\text{mid}U_\alpha$ and $\text{spr}U_\alpha$, respectively. Then, the *distance between two fuzzy numbers* $U, V \in \mathcal{F}_c(\mathbb{R})$ (see Trutschnig et al. [34]) can be defined by

$$D_\theta^\varphi(U, V) = \sqrt{\int_{(0,1]} \left[(\text{mid}U_\alpha - \text{mid}V_\alpha)^2 + \theta\,(\text{spr}U_\alpha - \text{spr}V_\alpha)^2\right] d\varphi(\alpha)}. \tag{2}$$

In the latter definition, the value $\theta > 0$ determines the relative weight of the distance of the generalized spreads (or, equivalently, the difference in shape or imprecision) with respect to the distance of the generalized midpoints (or difference in location). The choice of $\varphi$ is based on the importance given to each $\alpha$-level and it is associated to a bounded density measure with positive mass and support in (0,1]. In most of practical situations the Lebesgue measure is considered, which assigns the same importance to every $\alpha$-level.

If $0 < \theta \leq 1$, the metric $D_\varphi^\theta$ is equivalent to the well-known Bertoluzza metric (see Bertoluzza et al. [1]) and hence the first one generalizes the second one. It should also be noted that a value of $\theta$ closer to 0 gives more importance to the midpoints, while a high value of $\theta$ gives more importance to the spreads of the $\alpha$-cuts. In practice, the most common choice is $\theta = 1/3$ which is equivalent to the Lebesgue measure for each $\alpha$-cut, unless more information about the data is available. Further discussions about this metric can be found in Trutschnig et al. [34].

Please cite this article in press as: P. Giordani, A.B. Ramos-Guajardo, A fuzzy clustering procedure for random fuzzy sets, Fuzzy Sets Syst. (2016), http://dx.doi.org/10.1016/j.fss.2016.02.006