Iranian Journal of Fuzzy Systems

Volume 20, Number 6, (2023), pp. 1-20

Original paper

University of Sistan and Baluchestan

A robust fuzzy clustering model for fuzzy data based on an adaptive weighted L_1-norm

E. Eskandari[1] and A. Khastan[2]

[1,2] *Department of Mathematics, Institute for Advanced Studies in Basic Sciences, 444 Prof. Yousef Sobouti Blvd., Zanjan 45137-66731, Iran*

eskandari.e@iasbs.ac.ir, khastan@iasbs.ac.ir

Abstract

The imprecision related to measurements can be managed in terms of fuzzy features, which are characterized by two components: Center and spread. Outliers affect the outcome of the clustering models. In trying to overcome this problem, this paper proposes a fuzzy clustering model for L-R fuzzy data, which is based on a dissimilarity measure between each pair of fuzzy data defined as an adaptive weighted sum of the L_1-norms of the centers and the spreads. The proposed method is robust based on the metric and weighting approaches. It estimates the weight of a given fuzzy feature on a given fuzzy cluster by considering the relevance of that feature to the cluster; if outlier fuzzy features are present in the dataset, it tends to assign them weights close to 0.

To deeply investigate the capability of our model, i.e., alleviating undesirable effects of outlier fuzzy data, we provide a wide simulation study. We consider the ability to classify correctly and the ability to recover the true prototypes, both in the presence of outliers. The comparison made with other existing robust methods indicates that the proposed methodology is more robust to the presence of outliers than other methods. Moreover, the performance of our method decreases more slowly than others when the percentage of outliers increases. An application of the suggested method to a real-world categorical dataset is also provided.

Keywords: L-R fuzzy data, robust fuzzy clustering, L_1-norm, outliers.

1 Introduction

To overcome the uncertainty vagueness in real-life problems, Zadeh [40] introduced the fuzzy set theory [12]. Fuzzy rules are used for solving regression and classification problems [31]. *Clustering* is the task of grouping data into clusters of similar objects, and *fuzzy* clustering generalizes this notion by allowing an object belongs to more than one cluster. Fuzzy clustering models, such as Fuzzy C-Means (FCM) [3], have many applications (e.g., in image segmentation [2, 29], sentiment analysis [1], disease diagnosis [28], forecast [36], market segmentation [14], etc.).

In the real world, some measurements may be imprecise and some observations are vaguely defined. This kind of imprecise data can be represented using *fuzzy data*. In the literature, some fuzzy clustering models have been developed for dealing with fuzzy data. In particular, Sato and Sato [33] proposed a complete fuzzy model, i.e., a fuzzy clustering model for fuzzy data. D'Urso and Giordani [15] suggested a *weighted* fuzzy clustering model for fuzzy data, which is based on a dissimilarity measure between each pair of fuzzy data defined as a weighted sum of the squared Euclidean distances among the centers and the spreads. This model objectively estimates the weights and a larger weight is given to the distance among the centers than that among the spreads.

Outliers are completely arbitrary objects that just do not belong to the pattern or class being searched for [10]. A well-known drawback of the FCM is its lack of robustness to the presence of outliers. This occurs because $||x_k - g_i||_2^2 = \sum_{j=1}^{p}(x_{kj} - g_{ij})^2$, the datum-to-prototype dissimilarity term in FCM functional, can place considerable weight on outlying data, thus pulling prototypes away from the main distribution of the cluster [20].

Corresponding Author: E. Eskandari
Received: August 2022; Revised: February 2023; Accepted: September 2023.
https://doi.org/10.22111/IJFS.2023.43284.7606

An approach to overcome this shortcoming is where metrics with *robust* properties are incorporated into the FCM functional. A different approach is where each feature is given a weight and low weights are assigned to outliers. One approach is where outliers are assigned to a noise cluster. Another approach is where one tries to detect outliers and trims them away before the application of the clustering algorithms. A possibilistic approach is where outliers are included in all clusters with small membership degrees.

Our study of the literature over the last few years on fuzzy clustering of fuzzy data has found models which are robust in the presence of outliers:
- Butkiewicz [7] considered the defuzzification of fuzzy data and used a classical robust fuzzy clustering model for non-fuzzy data. As a consequence, the fuzzy nature of the data is not properly taken into account in the clustering process.
- Hung and Yang [23] proposed a robust fuzzy clustering model for univariate fuzzy data based on the metric approach. In this method, the fuzziness of the data is authentically taken into account in the clustering process. However, this method can only analyze univariate data.
- Hung et al. [24] adopted a strategy for clustering L-R fuzzy data based on a combination of three algorithms. In addition to the computational complexity, another major drawback is that the clustering methods utilized in this strategy are not based on fuzzy theory.
- Using the weighting approach, Zarandi and Razaee [41] proposed two robust fuzzy clustering models for triangular fuzzy data. This prevents these models from being used with other types of data belonging to the L-R family. The second model is based on a transformation that reduces the fuzzy clustering of fuzzy data to the fuzzy clustering of crisp data. As such, the second model is amenable to the same criticism of the model proposed by Butkiewicz [7].
- Coppi et al. [9] proposed a possibilistic k-means clustering model for multivariate L-R fuzzy data. However, it may suffer from the risk of obtaining coincident clusters, i.e., clusters characterized by the same prototypes.
- Considering Partitioning Around Medoids (PAM) approach [27], D'Urso and De Giovanni [13] proposed fuzzy clustering models for fuzzy data, called FkMedC-F, SFkMedC-F, FkMedC-NC-F, and TrFkMedC-F. Among the four mentioned models, in which estimated weights [15] are used, the last three are robust based on, respectively, the metric, noise cluster, and trimmed approaches. Nonetheless, the usability of the estimated weights is limited to cases where the centers of fuzzy data play a relevant role [17]. Moreover, the PAM approach provides only a timid robustification, i.e., the clustering outcome can still be disrupted by even a single outlier.
- Following the possibilistic approach, Ferraro and Giordani [18] proposed two robust clustering methods for fuzzy data. First, a fully possibilistic method is suggested, but an undesirable characteristic is a tendency to produce trivial solutions consisting of coincident clusters. Then a joint possibilistic and fuzzy method is introduced preventing the occurrence of coincident clusters, but the number of iterations is large and it is also required to set the parameters of the method at the beginning of the process.
- Following the PAM approach, D'Urso and Leski [16] proposed a robust fuzzy clustering model for fuzzy data, called FcOMdC-FD, based on the combination of loss functions (quadratic, Linear, Sigmoidal, Huber, and Logarithmic) and Ordered Weighted Averaging (OWA). This model can smooth the influence of the outlier fuzzy data through a parameter but it requires relatively high computational complexity owing to the implementation of operators such as Huber's M-estimators and Yager's OWA.

The results of theoretical studies indicate that L_1-*norm*-based methods are more robust than those based on the L_2-norm [25]. Jajuga [25] and Bobrowski and Bezdek [5] independently suggested replacing $||x_k - g_i||_2^2$ with $||x_k - g_i||_1^1 = \sum_{j=1}^{p} |x_{kj} - g_{ij}|$ in the FCM functional to increase robustness against outlying data. In the case of interval-valued data, De Carvalho and Simoes [11] proposed fuzzy c-means clustering model based on *adaptive* City-Block distance.

Furthermore, L_1-norm-based methods can be recommended for departure from the normality of the distributions in the classes [25]. Clusters of samples drawn from a mixture of uniform distributions have sharp or boxy edges. Many real datasets have clusters shaped like this. For data of this kind, inner product norms are inappropriate. The L_1 and L_∞ norms have open and closed sets that match these shapes and may yield better results. Figure 1, which shows the topological structure of several closed unit balls in $\mathbb{R}^2$, lends geometric plausibility to this supposition [5].

We are motivated to address the problem of the disruptive effect of the outlier fuzzy data by proposing an extension from the interval-valued data methodology developed by De Carvalho and Simoes [11], which in turn is related to Jajuga [25], to L-R fuzzy data. Experiments on synthetic and real-world datasets show that our proposed model outperforms other state-of-the-art models. To sum up, the main contributions of our work are highlighted as follows.

- We propose a robust fuzzy clustering method for L-R fuzzy data based on the metric and weighting approaches.
- We introduce a dissimilarity measure between each pair of fuzzy data defined as an adaptive weighted sum of the L_1-norms of the centers and the spreads.

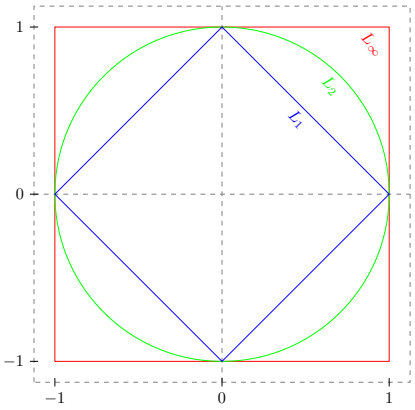


Figure 1: Some closed unit balls in $\mathbb{R}^2$.

- Due to the use of adaptive distance, we propose a fuzzy clustering algorithm with a step where a relevance weight for each fuzzy feature is learned. Such adaptive distance changes at each iteration of the algorithm and may be different from one cluster to another. If outlier fuzzy features are present in the dataset, our method tends to give them weights close to 0.
- For data drawn from the uniform distribution, our proposed method may yield better results.

The structure of the rest of the paper is as follows. In Section 2, some preliminary concepts about fuzzy data are briefly summarized. This section also includes the adaptation of the L_1-norm. Section 3 is devoted to describing the proposed algorithm. Section 4 contains a set of numerical experiments. Finally, Section 5 offers some concluding remarks.

2 Preliminaries

This section presents some preliminaries about the fuzzy framework.

Definition 2.1 (L-R fuzzy feature)**.** *Consider the fuzzy set $x_{kj} : \mathbb{R} \to [0,1]$, $k = 1, 2, \cdots, n$, $j = 1, 2, \cdots, p$, which denotes the j^{th} L-R fuzzy feature observed on the k^{th} L-R fuzzy datum $\mathbf{x}_k$. It is represented by the $\left(a_{1kj}^-, a_{1kj}^+, \underline{a}_{kj}, \overline{a}_{kj}\right)_{L,R}$, so that a_{1kj}^- and a_{1kj}^+ $(a_{1kj}^- \leq a_{1kj}^+)$ are called the left and the right centers of x_{kj}, and $\underline{a}_{kj}$ and $\overline{a}_{kj}$ are called the left and the right spreads of x_{kj}, respectively (see Figure 2). The membership degree of u to the fuzzy set x_{kj} is defined as*

$$x_{kj}(u) = \begin{cases} L\left(\dfrac{a_{1kj}^- - u}{\underline{a}_{kj}}\right), & u \leq a_{1kj}^-, \\ 1, & a_{1kj}^- \leq u \leq a_{1kj}^+, \\ R\left(\dfrac{u - a_{1kj}^+}{\overline{a}_{kj}}\right), & u \geq a_{1kj}^+, \end{cases}$$

where $L, R : [0,1] \to [0,1]$ are two continuous, decreasing functions with $L(0) = R(0) = 1$ and $L(1) = R(1) = 0$.

Particularly, when the functions L and R are linear, trapezoidal fuzzy features are obtained and can be represented by the quadruple $\left(a_{1kj}^-, a_{1kj}^+, \underline{a}_{kj}, \overline{a}_{kj}\right)$. Trapezoidal membership function is the most frequently used fuzzy set representation in fuzzy systems [26]. Moreover, if $a_{1kj}^- = a_{1kj}^+$ in the representation $\left(a_{1kj}^-, a_{1kj}^+, \underline{a}_{kj}, \overline{a}_{kj}\right)$, then the fuzzy feature is called a triangular fuzzy feature and the triple $(a_{1kj}, \underline{a}_{kj}, \overline{a}_{kj})$ is sufficient to represent it. We will denote by $\mathcal{F}_{L-R}(\mathbb{R})$ the space of L-R fuzzy numbers. The L-R fuzzy dataset, denoted by $\mathcal{D}$, can be stored in a matrix of dimension $n \times p$ (indeed, a tensor of dimension $n \times p \times 4$, see Figure 3).

The fuzzy nature of the data to be clustered leads to the fuzziness of the prototypes. The fuzzy set g_{ij}, $i = 1, 2, \cdots, c$, which denotes the j^{th} fuzzy feature observed on the i^{th} fuzzy prototype $\mathbf{g}_i$ is represented by the quadruple $\left(b_{1ij}^-, b_{1ij}^+, \underline{b}_{ij}, \overline{b}_{ij}\right)$. The fuzzy prototypes are stored in a matrix of dimension $c \times p$, denoted by $\mathbf{G}$.

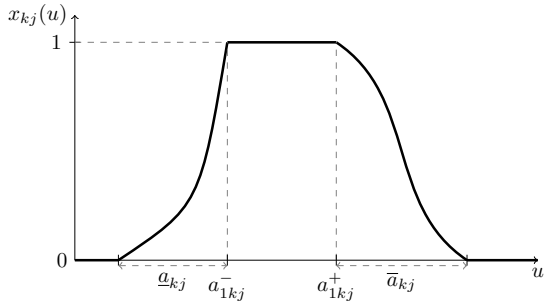


Figure 2: L-R membership function.

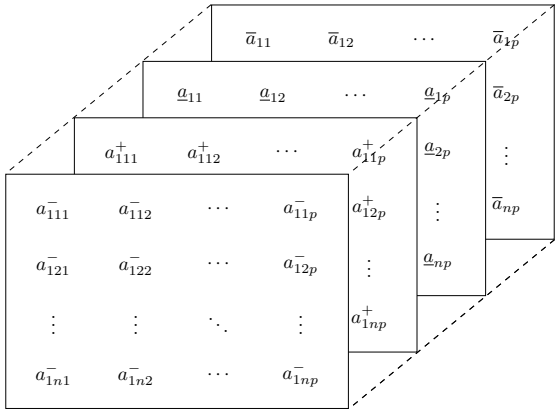


Figure 3: The L-R fuzzy data stored in a tensor of dimension $n \times p \times 4$.

Definition 2.2. *Let* $\mathbf{x}_k = (x_{k1}, \cdots, x_{kp})$ *and* $\mathbf{x}_{k'} = (x_{k'1}, \cdots, x_{k'p})$ *be a pair of fuzzy data in* $\mathcal{F}_{L-R}(\mathbb{R})$. *For* $i = 1, 2, \cdots, c$, *the adaptive dissimilarity measure between* $\mathbf{x}_k$ *and* $\mathbf{x}_{k'}$ *is defined by*

$$\Delta_{\boldsymbol{\lambda}_i, v}(\mathbf{x}_k, \mathbf{x}_{k'}) = \sum_{j=1}^{p} \lambda_{ij} \delta_v(x_{kj}, x_{k'j}),$$

where

$$\delta_v(x_{kj}, x_{k'j}) = (1-v)^2 \Big[\Big| a^{-}_{1kj} - a^{-}_{1k'j} \Big| + \Big| a^{+}_{1kj} - a^{+}_{1k'j} \Big| \Big] + v^2 \Big[\Big| \underline{a}_{kj} - \underline{a}_{k'j} \Big| + \Big| \overline{a}_{kj} - \overline{a}_{k'j} \Big| \Big], \tag{1}$$

is the weighted sum of the L_1*-norms of the centers and the spreads of* j^{th} *fuzzy feature of* $\mathbf{x}_k$ *and that of* $\mathbf{x}_{k'}$, $v \in [0, 1]$ *measures the importance of the distance among the spreads, and* λ_{ij}, *satisfying* $\lambda_{ij} > 0$ *and* $\prod_{j=1}^{p} \lambda_{ij} = 1$, *measures how much the* j^{th} *fuzzy feature is relevant to the* i^{th} *fuzzy cluster.*

Remark 2.3. *Since* $\Delta_{\boldsymbol{\lambda}_i, v}$ *is a linear combination of distances, it is a distance.*

3 Robust fuzzy clustering model for fuzzy data

The proposed fuzzy clustering model for L-R fuzzy data based on the dissimilarity measure introduced in Definition 2.2, namely FCM-F-L1, is

$$\min : \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ki}^{m} \Delta_{\boldsymbol{\lambda}_i, v}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ki}^{m} \sum_{j=1}^{p} \lambda_{ij} \Big\{ (1-v)^2 \Big[\Big| a^{-}_{1kj} - b^{-}_{1ij} \Big| + \Big| a^{+}_{1kj} - b^{+}_{1ij} \Big| \Big] + v^2 \Big[\Big| \underline{a}_{kj} - \underline{b}_{ij} \Big| + \Big| \overline{a}_{kj} - \overline{b}_{ij} \Big| \Big] \Big\}, \tag{2}$$

with the constraints

$$\begin{aligned}
&\sum_{i=1}^{c} u_{ki} = 1, u_{ki} \geq 0,\\
&\prod_{j=1}^{p} \lambda_{ij} = 1, \lambda_{ij} > 0,\\
&0 \leq v \leq 1,
\end{aligned} \tag{3}$$

where u_{ki} is the membership degree of k^{th} fuzzy datum in the i^{th} fuzzy cluster, $m \in (1, +\infty)$ is the fuzziness parameter, $\Delta_{\boldsymbol{\lambda}_i, v}(\mathbf{x}_k, \mathbf{g}_i)$ compares the k^{th} L-R fuzzy datum and the i^{th} fuzzy prototype, λ_{ij} is the relevance measure of the j^{th} fuzzy feature to the i^{th} fuzzy cluster, and v measures the importance of the distance among the spreads. The membership degrees are stored in a matrix of dimension $n \times c$, denoted by $\mathbf{U}$.

Remark 3.1. *The solution of v obtained by the method of Lagrange multipliers with the constraint $0 \leq v \leq 0.5$ [15] is*

$$v = \min \left\{ \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki}^{m} \sum_{j=1}^{p} \lambda_{ij} \left[\left| a_{1kj}^{-} - b_{1ij}^{-} \right| + \left| a_{1kj}^{+} - b_{1ij}^{+} \right| \right]}{\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki}^{m} \sum_{j=1}^{p} \lambda_{ij} \left[\left| a_{1kj}^{-} - b_{1ij}^{-} \right| + \left| a_{1kj}^{+} - b_{1ij}^{+} \right| + \left| \underline{a}_{kj} - \underline{b}_{ij} \right| + \left| \overline{a}_{kj} - \overline{b}_{ij} \right| \right]}, 0.5 \right\}.$$

However, in the present study, v is supposed to be fixed and equal to 0.5, which, for two reasons, is the best strategy [17]*:*

(i) *The estimated v is allowed to range between 0 and 0.5, while there is no external condition on v.*

(ii) *The method of Lagrange multipliers attempts to find a minimum of the cost function by giving less (or at most equal) importance to the distinguishing distance, i.e., either the distance among the centers or that among the spreads, whichever is longer.*

3.1 Algorithm

The FCM-F-L1 algorithm is performed in three steps, namely: update, weighting, and assignment (see Figure 4). The update step computes g_{ij}, which denotes the j^{th} fuzzy feature observed on the i^{th} fuzzy prototype. The weighting step computes λ_{ij}, which measures how much the j^{th} fuzzy feature is relevant to the i^{th} fuzzy cluster. Finally, the assignment step computes u_{ki}, which is the membership degree of the k^{th} fuzzy datum in the i^{th} fuzzy cluster.

Update step During the update step, λ_{ij}'s and $\mathbf{U}$ are kept fixed. For $i = 1, 2, \cdots, c$, we obtain the optimal prototype $\mathbf{g}_i$ that minimizes

$$\sum_{j=1}^{p} \sum_{k=1}^{n} u_{ki}^{m} \delta_v(x_{kj}, g_{ij}).$$

We extend the idea of De Carvalho and Simoes [11] and Jajuga [25] to the fuzzy framework. For $j = 1, 2, \cdots, p$, the left and the right centers and the left and the right spreads of the fuzzy feature g_{ij} observed on the fuzzy prototype $\mathbf{g}_i$ are computed from

$$\sum_{k=1}^{n} u_{ki}^{m} \delta_v(x_{kj}, g_{ij}) \rightarrow \min,$$

or

$$\sum_{k=1}^{n} u_{ki}^{m} \left[\left| a_{1kj}^{-} - b_{1ij}^{-} \right| + \left| a_{1kj}^{+} - b_{1ij}^{+} \right| + \left| \underline{a}_{kj} - \underline{b}_{ij} \right| + \left| \overline{a}_{kj} - \overline{b}_{ij} \right| \right] \rightarrow \min,$$

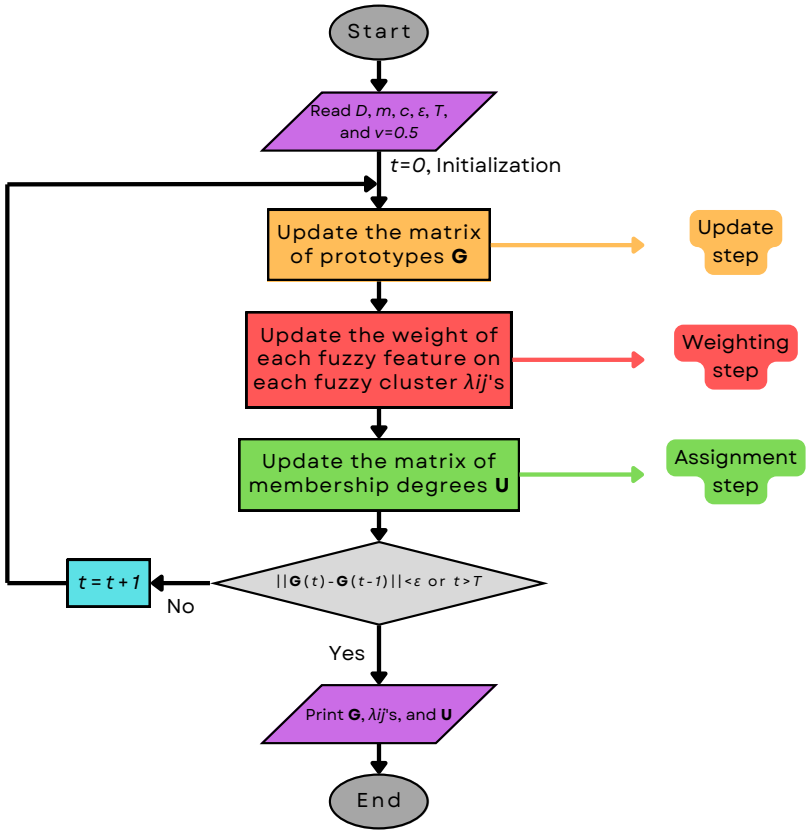


Figure 4: Overview of the proposed clustering model.

or, respectively, from

$$\sum_{k=1}^{n} u_{ki}^{m} \left| a_{1kj}^{-} - b_{1ij}^{-} \right| \to \min, \tag{4}$$

$$\sum_{k=1}^{n} u_{ki}^{m} \left| a_{1kj}^{+} - b_{1ij}^{+} \right| \to \min, \tag{5}$$

$$\sum_{k=1}^{n} u_{ki}^{m} \left| \underline{a}_{kj} - \underline{b}_{ij} \right| \to \min, \tag{6}$$

$$\sum_{k=1}^{n} u_{ki}^{m} \left| \overline{a}_{kj} - \overline{b}_{ij} \right| \to \min . \tag{7}$$

Each of the four minimization problems (4)–(7) brings to the minimization of

$$\sum_{k=1}^{n} |y_k - a z_k| , \tag{8}$$

where $y_k = u_{ki}^{m} a_{1kj}^{-}$ (respectively, $y_k = u_{ki}^{m} a_{1kj}^{+}$, $y_k = u_{ki}^{m} \underline{a}_{kj}$, and $y_k = u_{ki}^{m} \overline{a}_{kj}$), $a = b_{1ij}^{-}$ (respectively, $a = b_{1ij}^{+}$, $a = \underline{b}_{ij}$, and $a = \overline{b}_{ij}$), and $z_k = u_{ki}^{m}$. The problem of minimization of (8) is well-known and solved (the discussion is presented,

e.g., in [4]). To solve the optimization problem (8), the Algorithm 1 can be used.

Algorithm 1: Algorithmic solution of the optimization problem (8).

input : (y_k, z_k)
output: a
initialization;
$r \leftarrow 0$;
Rank (y_k, z_k) such that $\frac{y_{k_1}}{z_{k_1}} \leq \cdots \leq \frac{y_{k_n}}{z_{k_n}}$;
repeat
  $r \leftarrow r + 1$
until $-\sum_{l=1}^{n}|z_{k_l}| + 2\sum_{s=1}^{r+1}|z_{k_s}| > 0$;
$a \leftarrow \frac{y_{k_r}}{z_{k_r}}$;
if $-\sum_{l=1}^{n}|z_{k_l}| + 2\sum_{s=1}^{r}|z_{k_s}| = 0$ ***and*** $-\sum_{l=1}^{n}|z_{k_l}| + 2\sum_{s=1}^{r+1}|z_{k_s}| = 0$ **then**
  $a = \frac{\frac{y_{k_r}}{z_{k_r}} + \frac{y_{k_{r+1}}}{z_{k_{r+1}}}}{2}$
end

The optimum along the descent direction is a weighted median of ratios $\left\{\frac{y_k}{z_k}\right\}$. The Algorithm 1 finds the smallest ratio $a^{(1)} = \frac{y_{k_1}}{z_{k_1}}$ and tests whether it is the optimum, i.e., $\sum_{k=1}^{n}|y_k - az_k|$ increases for $a > a^{(1)}$. If not, the smallest remaining ratio is found and tested, etc. As a varies, $\sum_{k=1}^{n}|y_k - az_k|$ is minimized at the weighted median of the $\left\{\frac{y_k}{z_k}\right\}$ with weights $|z_k|$. Its graph looks like Figure 5 [4].

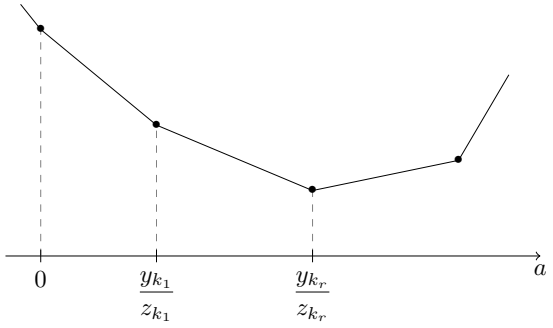


Figure 5: The line search of finding the minimum of $\sum_{k=1}^{n}|y_k - az_k|$.

Proposition 3.2 (**Weighting and Assignment steps**). *The iterative solution of the constrained optimization problem* (2)–(3) *is*

$$\lambda_{ij} = \frac{\left(\prod_{l=1}^{p}\sum_{k=1}^{n} u_{ki}^{m}\delta_v(x_{kl}, g_{il})\right)^{\frac{1}{p}}}{\sum_{k=1}^{n} u_{ki}^{m}\delta_v(x_{kj}, g_{ij})}, \tag{9}$$

$$u_{ki} = \left(\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\lambda_{ij}\delta_w(x_{kj}, g_{ij})}{\sum_{j=1}^{p}\lambda_{hj}\delta_w(x_{kj}, g_{hj})}\right)^{\frac{1}{m-1}}\right)^{-1}. \tag{10}$$

Proof of Proposition 3.2. The proof of (9) is analogous to [11]. Keeping fixed $\mathbf{G}$ and λ_{ij}'s, we obtain the optimal membership degrees. By considering the Lagrangian function

$$\mathcal{L}_m(u_{ki},\eta)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ki}^{m}\sum_{j=1}^{p}\lambda_{ij}\delta_w(x_{kj},g_{ij})-\eta\left(\sum_{i=1}^{c}u_{ki}-1\right),$$

we take the partial derivatives and set to 0:

$$\frac{\partial\mathcal{L}_m(u_{ki},\eta)}{\partial\eta}=\sum_{i=1}^{c}u_{ki}-1=0, \tag{11}$$

and

$$\frac{\partial\mathcal{L}_m(u_{ki},\eta)}{\partial u_{k'h}}=mu_{k'h}^{m-1}\sum_{j=1}^{p}\lambda_{hj}\delta_w(x_{k'j},g_{hj})-\eta=0. \tag{12}$$

From (12) we have

$$u_{ki}=\left(\frac{\eta}{m\left(\sum\limits_{j=1}^{p}\lambda_{ij}\delta_w(x_{kj},g_{ij})\right)}\right)^{\frac{1}{m-1}}, \tag{13}$$

and, upon substituting (13) into (11),

$$\left(\frac{\eta}{m}\right)^{\frac{1}{m-1}}\sum_{h=1}^{c}\frac{1}{\left(\sum\limits_{j=1}^{p}\lambda_{hj}\delta_w(x_{k'j},g_{hj})\right)^{\frac{1}{m-1}}}=1. \tag{14}$$

Finally, substituting (14) into (13) yields (10). □

The suggested algorithm can be summarized according to 2.

3.2 Complexity analysis

The time complexity of FCM-F-L1 algorithm can be evaluated by considering the unit cost of every individual step. It depends on the number of objects n, the number of clusters c, and the number of features p. The initialization costs $O(c\times n\times(c+p))$; the update step costs $O(c\times p\times n\times\log(n))$; the weighting step costs $O(c\times p\times n)$; and the assignment step costs $O(c\times n\times(c+p))$. Assuming that the Algorithm 2 needs T iterations to converge, the entire complexity is $O(c\times p\times n\times\max(c,\log(n))\times T)$ (see [11, Section 2.2.4.] for more details).

4 Numerical experiments

We implemented in MATLAB R2021a, the four suggested models in [13] and the FcOMdC-FD SQR, FcOMdC-FD LIN, FcOMdC-FD SIG ($\alpha=\beta=2$), FcOMdC-FD HUB ($\delta=5$), and FcOMdC-FD LOG models proposed in [16], based on the original papers, as well as our suggested clustering model. The effectiveness of the competing models is measured using the fuzzy Rand (frand) [8] and the Hullermeier (HUL) [22] indexes, which are measures of classifying correctly. The closer to 1 the values of the indexes frand and HUL, the better the performance of the model is. REC index introduced in [30] as

$$\mathrm{REC}=\sum_{i=1}^{c}\left[d^2\left(\mathbf{b}_{1i}^{-\,\mathrm{O}},\mathbf{b}_{1i}^{-\,\mathrm{E}}\right)+d^2\left(\mathbf{b}_{1i}^{+\,\mathrm{O}},\mathbf{b}_{1i}^{+\,\mathrm{E}}\right)+d^2\left(\underline{\mathbf{b}}_i{}^{\mathrm{O}},\underline{\mathbf{b}}_i{}^{\mathrm{E}}\right)+d^2\left(\overline{\mathbf{b}}_i{}^{\mathrm{O}},\overline{\mathbf{b}}_i{}^{\mathrm{E}}\right)\right],$$

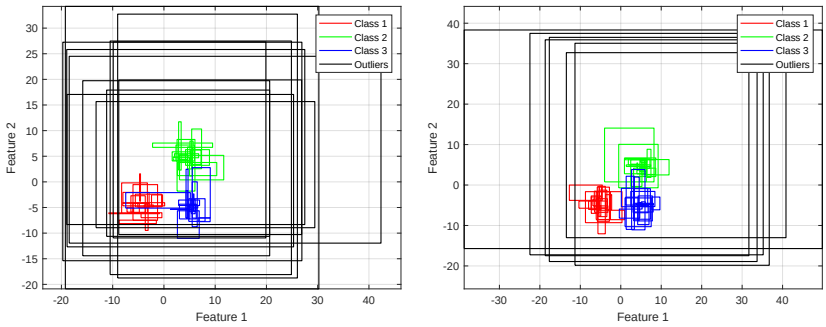
has also been used. The smaller the value of the REC index, the smaller the deviation of the computed (observed) prototypes from the true (expected) prototypes. The results are averaged over 100 runs and expressed as mean and 95% confidence interval (CI). All experiments were conducted on the same machine (OS: macOS, Memory: 8GB, Processor: 1.1GHz dual-core Intel Core i3).

Algorithm 2: FCM-F-L1 algorithm.

input : The L-R fuzzy dataset $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_k, \cdots, \mathbf{x}_n\}$, the fuzziness parameter $m \in (1, +\infty)$, the number of clusters $c \in \{2, 3, \cdots, n-1\}$, the minimum amount of improvement $\varepsilon = 1e-5$, the maximum number of iterations T, and the importance of the distance among the spreads $v = 0.5$.

output: The matrix of prototypes $\mathbf{G}$, the weight of each fuzzy feature on each fuzzy cluster λ_{ij}'s, and the matrix of membership degrees $\mathbf{U}$.

initialization;
$t \leftarrow 0$;
for $i = 1 : c$ **do**
  Randomly choose c distinct prototype $\mathbf{g}_i^{(0)} \in \mathcal{D}$ to obtain the $\mathbf{G}^{(0)}$
end
for $i = 1 : c$ **do**
  for $j = 1 : p$ **do**
    $\boldsymbol{\lambda}_{ij}^{(0)} \leftarrow 1$
  end
end
for $k = 1 : n$ **do**
  for $i = 1 : c$ **do**
    Calculate $u_{ki}^{(0)}$ to obtain $\mathbf{U}^{(0)}$, using Equation (10)
  end
end
repeat
  $t \leftarrow t + 1$;
  for $i = 1 : c$ **do**
    for $j = 1 : p$ **do**
      Calculate $g_{ij}^{(t)}$ to update $\mathbf{G}^{(t)}$, using Algorithm 1, keeping fixed $\lambda_{ij}^{(t-1)}$'s and $\mathbf{U}^{(t-1)}$; `/*Update step*/`
    end
  end
  for $i = 1 : c$ **do**
    for $j = 1 : p$ **do**
      Update $\lambda_{ij}^{(t)}$, using Equation (9), keeping fixed $\mathbf{G}^{(t)}$ and $\mathbf{U}^{(t-1)}$; `/*Weighting step*/`
    end
  end
  for $k = 1 : n$ **do**
    for $i = 1 : c$ **do**
      Calculate $u_{ki}^{(t)}$ to update $\mathbf{U}^{(t)}$, using Equation (10), keeping fixed $\mathbf{G}^{(t)}$ and $\lambda_{ij}^{(t)}$'s ; `/*Assignment step*/`
    end
  end
until $\left|\left|\mathbf{G}^{(t)} - \mathbf{G}^{(t-1)}\right|\right| < \varepsilon$ *or* $t > T$;



(a) $C_P = 0.2, C_D = 5$ (b) $C_P = 0.1, C_D = 10$

Figure 6: Examples of well-separated classes.

Table 1: Set-up of the simulation study, well-separated classes.

Subsample		Dim.	Left center	Right center	Spreads
Inliers	Class 1	1,2	$\mathcal{N}(-5,1)-\chi^2_1$	$\mathcal{N}(-5,1)+\chi^2_1$	χ^2_1
	Class 2	1,2	$\mathcal{N}(5,1)-\chi^2_1$	$\mathcal{N}(5,1)+\chi^2_1$	χ^2_1
	Class 3	1	$\mathcal{N}(5,1)-\chi^2_1$	$\mathcal{N}(5,1)+\chi^2_1$	χ^2_1
		2	$\mathcal{N}(-5,1)-\chi^2_1$	$\mathcal{N}(-5,1)+\chi^2_1$	χ^2_1
Outliers		1,2	$\mathcal{N}(0,3)-\chi^2_4$	$\mathcal{N}(0,3)+\chi^2_4+2C_D$	$\chi^2_4+C_D$

4.1 Two-dimensional datasets

Several mechanisms to generate outlier fuzzy data have been proposed in the literature. In this paper, we follow [34, 35, 13] to generate synthetic datasets, which are specially designed for the characteristics of the proposed method. The fuzziness parameter m is chosen to be equal to 1.5.

4.1.1 Well-separated classes

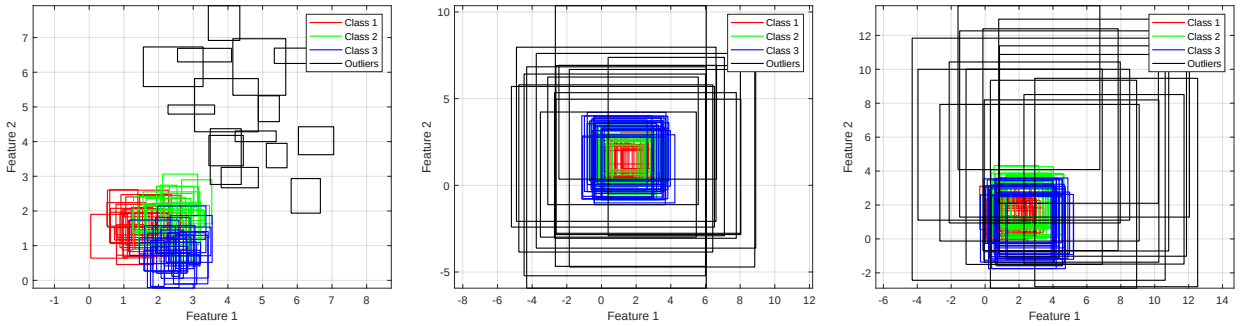
Each sample of size $n = 60$ of trapezoidal fuzzy data is assumed to be split into nC_P outliers and $n(1 - C_P)$ inliers. The inliers are supposed to be split into 3 well-separated groups of equal sizes. The proportion of anomality, C_P, is supposed to range in {0.1,0.2} and C_D, which measures how far the outliers are from the inliers, ranges in {5,10}. Each object of the datasets is a vector of intervals, i.e., a rectangle: $([a^-_{1k1} - \underline{a}_{k1}, a^+_{1k1} + \overline{a}_{k1}], [a^-_{1k2} - \underline{a}_{k2}, a^+_{1k2} + \overline{a}_{k2}])$, $k = 1, 2, \cdots, 60$ (see Figures 6a and 6b). Table 1 summarizes the data generation process.

4.1.2 Overlapped classes

The datasets are generated according to three schemes, i.e., center, spreads, and center and spreads. Each dataset consists of 3 overlapped classes (each of 25 triangular fuzzy data) and 15 outliers. Each object of the datasets is a vector of intervals, i.e., a rectangle: $([a_{1k1} - \underline{a}_{k1}, a_{1k1} + \overline{a}_{k1}], [a_{1k2} - \underline{a}_{k2}, a_{1k2} + \overline{a}_{k2}])$, $k = 1, 2, \cdots, 90$ (see Figures 7a, 7b, and 7c). Table 2 summarizes the data generation process.

The clustering results corresponding to the well-separated and overlapped classes are listed in Tables 3 and 4, respectively.

- FCM-F-L1 model presents the best performances for the $C_P = 0.1, C_D = 5$ and the spreads scheme.
- FCM-F-L1 model outperforms other comparative models in frand and HUL for the $C_P = 0.1, C_D = 10$ and $C_P = 0.2, C_D = 5$.
- FCM-F-L1 model is superior to other competitive models in REC for three schemes.



(a) Center (b) Spreads (c) Center and spreads

Figure 7: Examples of overlapped classes.

Table 2: Set-up of the simulation study, overlapped classes.

Scheme		Dim.	Center	Spreads
Center	Class 1	1,2	$\mathcal{U}[1,2]$	$\mathcal{U}[0,1]$
	Class 2	1	$\mathcal{U}[2,3]$	$\mathcal{U}[0,1]$
		2	$\mathcal{U}[1.5,2.5]$	$\mathcal{U}[0,1]$
	Class 3	1	$\mathcal{U}[2,3]$	$\mathcal{U}[0,1]$
		2	$\mathcal{U}[0.5,1.5]$	$\mathcal{U}[0,1]$
	Outliers	1,2	$\mathcal{N}(5,2)$	$\mathcal{U}[0,1]$
Spreads	Class 1	1,2	$\mathcal{U}[1,2]$	$\mathcal{U}[0.5,1.5]$
	Class 2	1,2	$\mathcal{U}[1,2]$	$\mathcal{U}[1,2]$
	Class 3	1,2	$\mathcal{U}[1,2]$	$\mathcal{U}[1.5,2.5]$
	Outliers	1,2	$\mathcal{U}[1,2]$	$\mathcal{N}(5,2)$
Center and spreads	Class 1	1,2	$\mathcal{U}[1,2]$	$\mathcal{U}[0.5,1.5]$
	Class 2	1	$\mathcal{U}[2,3]$	$\mathcal{U}[1,2]$
		2	$\mathcal{U}[1.5,2.5]$	$\mathcal{U}[1,2]$
	Class 3	1	$\mathcal{U}[2,3]$	$\mathcal{U}[1.5,2.5]$
		2	$\mathcal{U}[0.5,1.5]$	$\mathcal{U}[1.5,2.5]$
	Outliers	1,2	$\mathcal{N}(5,2)$	$\mathcal{N}(5,2)$

Table 3: The clustering results corresponding to the case of well-separated classes.

C_P	C_D	Method	frand		HUL		REC	
			Mean	95% CI	Mean	95% CI	Mean	95% CI
0.1	5	FkMedC-F 2014	0.7707	(0.7415,0.7999)	0.6780	(0.6356,0.7204)	278	(225,332)
		SFkMedC-F 2014	0.7724	(0.7433,0.8015)	0.6726	(0.6306,0.7146)	307	(248,366)
		FkMedC-NC-F 2014	0.7714	(0.7604,0.7823)	0.6467	(0.6308,0.6626)	431	(357,505)
		TrFkMedC-F 2014	0.7026	(0.6765,0.7287)	0.5657	(0.5241,0.6074)	434	(369,499)
		FcOMdC-FD, SQR 2020	0.7855	(0.7492,0.8218)	0.6701	(0.6084,0.7319)	299	(243,355)
		FcOMdC-FD, LIN 2020	0.7366	(0.7102,0.763)	0.5921	(0.5506,0.6337)	271	(221,321)
		FcOMdC-FD, SIG 2020	0.6972	(0.6803,0.714)	0.5052	(0.4825,0.5278)	162	(134,189)
		FcOMdC-FD, HUB 2020	0.8113	(0.7781,0.8445)	0.7034	(0.6482,0.7586)	248	(189,307)
		FcOMdC-FD, LOG 2020	0.7286	(0.7038,0.7535)	0.5567	(0.5201,0.5932)	220	(175,266)
		FCM-F-L1	**0.8457**	(0.8275,0.8638)	**0.7350**	(0.7143,0.7558)	**129**	(87,171)
	10	FkMedC-F 2014	0.7511	(0.7162,0.7861)	0.6364	(0.5852,0.6875)	769	(583,956)
		SFkMedC-F 2014	0.7937	(0.7651,0.8222)	0.6856	(0.6436,0.7276)	480	(353,607)
		FkMedC-NC-F 2014	0.7549	(0.7412,0.7685)	0.6154	(0.5963,0.6346)	606	(456,755)
		TrFkMedC-F 2014	0.7166	(0.6965,0.7367)	0.5928	(0.5629,0.6227)	601	(466,737)
		FcOMdC-FD, SQR 2020	0.7733	(0.7411,0.8055)	0.6636	(0.6115,0.7156)	683	(533,834)
		FcOMdC-FD, LIN 2020	0.7627	(0.7351,0.7903)	0.6172	(0.5754,0.659)	515	(384,645)
		FcOMdC-FD, SIG 2020	0.6945	(0.678,0.7109)	0.4781	(0.4552,0.5009)	**160**	(134,187)
		FcOMdC-FD, HUB 2020	0.7793	(0.7512,0.8074)	0.6692	(0.6259,0.7125)	517	(393,640)
		FcOMdC-FD, LOG 2020	0.7006	(0.6823,0.719)	0.5168	(0.4919,0.5416)	406	(306,506)
		FCM-F-L1	**0.8417**	(0.8257,0.8578)	**0.7228**	(0.7052,0.7404)	2nd 273	(172,373)
0.2	5	FkMedC-F 2014	0.6939	(0.6549,0.7328)	0.5404	(0.4773,0.6034)	604	(511,698)
		SFkMedC-F 2014	0.7697	(0.7439,0.7954)	0.6714	(0.6325,0.7103)	427	(362,493)
		FkMedC-NC-F 2014	0.7871	(0.7715,0.8027)	0.6548	(0.6326,0.677)	629	(519,738)
		TrFkMedC-F 2014	0.6822	(0.6561,0.7082)	0.5469	(0.5062,0.5877)	633	(532,734)
		FcOMdC-FD, SQR 2020	0.7490	(0.7115,0.7865)	0.6407	(0.5848,0.6965)	550	(463,638)
		FcOMdC-FD, LIN 2020	0.7164	(0.6895,0.7433)	0.5648	(0.5224,0.6072)	464	(397,532)
		FcOMdC-FD, SIG 2020	0.6859	(0.6659,0.7058)	0.4723	(0.4382,0.5063)	**175**	(144,205)
		FcOMdC-FD, HUB 2020	0.7485	(0.7193,0.7777)	0.6427	(0.6001,0.6854)	548	(471,625)
		FcOMdC-FD, LOG 2020	0.6932	(0.6713,0.7152)	0.5101	(0.4755,0.5446)	382	(321,442)
		FCM-F-L1	**0.7933**	(0.7719,0.8147)	**0.6786**	(0.6541,0.7031)	2nd 318	(259,378)
	10	FkMedC-F 2014	0.7046	(0.6698,0.7394)	0.5933	(0.5447,0.6418)	1075	(870,1280)
		SFkMedC-F 2014	0.7370	(0.701,0.773)	0.6246	(0.5755,0.6737)	979	(756,1201)
		FkMedC-NC-F 2014	0.7588	(0.7374,0.7802)	0.6101	(0.5804,0.6398)	1488	(1247,1729)
		TrFkMedC-F 2014	0.6655	(0.6359,0.6951)	0.5138	(0.4665,0.561)	1242	(1026,1457)
		FcOMdC-FD, SQR 2020	0.7472	(0.7115,0.7829)	**0.6464**	(0.596,0.6967)	982	(772,1192)
		FcOMdC-FD, LIN 2020	0.6899	(0.6606,0.7191)	0.5185	(0.4701,0.5668)	1056	(871,1241)
		FcOMdC-FD, SIG 2020	0.7120	(0.6939,0.7301)	0.5093	(0.4822,0.5364)	**375**	(267,483)
		FcOMdC-FD, HUB 2020	0.7364	(0.7032,0.7695)	0.6092	(0.5572,0.6612)	1002	(816,1189)
		FcOMdC-FD, LOG 2020	0.6988	(0.6777,0.7199)	0.5066	(0.4721,0.541)	639	(508,770)
		FCM-F-L1	**0.7598**	(0.7319,0.7876)	0.6300	(0.5961,0.6639)	3rd 818	(660,976)

The best result is marked in bold while the second and third best results are marked with 2nd and 3rd, respectively.

Table 4: The clustering results corresponding to the case of overlapped classes.

Scheme	Method	frand		HUL		REC	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
Center	FkMedC-F 2014	0.6279	(0.608,0.6478)	0.4506	(0.4257,0.4756)	14.71	(11.94,17.47)
	SFkMedC-F 2014	0.6694	(0.6542,0.6847)	0.4730	(0.4527,0.4934)	7.64	(5.59,9.7)
	FkMedC-NC-F 2014	0.5663	(0.5582,0.5744)	0.3453	(0.3332,0.3573)	11.84	(8.96,14.73)
	TrFkMedC-F 2014	0.6368	(0.621,0.6527)	0.4252	(0.4021,0.4482)	12.63	(8.83,16.43)
	FcOMdC-FD, SQR 2020	0.6656	(0.6425,0.6888)	0.5088	(0.4774,0.5403)	16.03	(13.24,18.83)
	FcOMdC-FD, LIN 2020	0.6619	(0.6423,0.6815)	0.4459	(0.4185,0.4732)	9.47	(7.09,11.84)
	FcOMdC-FD, SIG 2020	0.6486	(0.6286,0.6686)	0.4283	(0.4008,0.4558)	8.89	(6.26,11.53)
	FcOMdC-FD, HUB 2020	0.6785	(0.6557,0.7012)	0.5289	(0.4988,0.559)	13.37	(10.77,15.97)
	FcOMdC-FD, LOG 2020	**0.7193**	(0.6929,0.7456)	**0.5655**	(0.5281,0.6029)	10.07	(7.56,12.57)
	FCM-F-L1	3rd 0.6726	(0.6621,0.6831)	0.4197	(0.4069,0.4325)	**1.39**	(0.73,2.05)
Spreads	FkMedC-F 2014	0.5238	(0.5164,0.5311)	0.2872	(0.2739,0.3006)	13.36	(9.62,17.09)
	SFkMedC-F 2014	0.5426	(0.5335,0.5517)	0.2975	(0.2809,0.3141)	13.07	(10.01,16.12)
	FkMedC-NC-F 2014	0.5119	(0.506,0.5178)	0.2699	(0.262,0.2778)	14.20	(9.87,18.54)
	TrFkMedC-F 2014	0.5421	(0.5291,0.5552)	0.2942	(0.2783,0.3101)	21.95	(16.6,27.3)
	FcOMdC-FD, SQR 2020	0.5266	(0.517,0.5362)	0.2710	(0.2594,0.2826)	15.31	(11.52,19.1)
	FcOMdC-FD, LIN 2020	0.5167	(0.5119,0.5215)	0.2015	(0.1951,0.2078)	15.48	(12.14,18.81)
	FcOMdC-FD, SIG 2020	0.5003	(0.4921,0.5084)	0.1278	(0.1221,0.1336)	17.20	(13.65,20.75)
	FcOMdC-FD, HUB 2020	0.5255	(0.5155,0.5356)	0.2729	(0.2606,0.2852)	17.90	(14.36,21.44)
	FcOMdC-FD, LOG 2020	0.5275	(0.5196,0.5354)	0.2624	(0.2513,0.2736)	14.89	(11.58,18.21)
	FCM-F-L1	**0.6209**	(0.615,0.6267)	**0.3744**	(0.3665,0.3823)	**12.54**	(9.47,15.6)
Center and spreads	FkMedC-F 2014	0.6528	(0.628,0.6777)	0.4886	(0.4537,0.5235)	50.86	(39.1,62.61)
	SFkMedC-F 2014	0.7174	(0.6843,0.7504)	0.5647	(0.5188,0.6105)	41.67	(29.22,54.12)
	FkMedC-NC-F 2014	0.6904	(0.6788,0.702)	0.5018	(0.4829,0.5208)	54.80	(41.89,67.7)
	TrFkMedC-F 2014	0.6474	(0.6271,0.6677)	0.4583	(0.424,0.4927)	51.86	(38.67,65.06)
	FcOMdC-FD, SQR 2020	0.6248	(0.6014,0.6482)	0.4545	(0.4235,0.4854)	50.60	(40.89,60.32)
	FcOMdC-FD, LIN 2020	0.6360	(0.6137,0.6584)	0.4014	(0.3696,0.4333)	48.83	(37.88,59.78)
	FcOMdC-FD, SIG 2020	0.6497	(0.6329,0.6665)	0.4148	(0.3896,0.44)	39.61	(30,49.22)
	FcOMdC-FD, HUB 2020	0.6638	(0.6409,0.6868)	0.5106	(0.4795,0.5418)	44.50	(34.13,54.88)
	FcOMdC-FD, LOG 2020	0.7230	(0.6961,0.7499)	**0.5777**	(0.5419,0.6135)	45.51	(34.48,56.53)
	FCM-F-L1	**0.7260**	(0.713,0.7391)	3rd 0.5466	(0.5338,0.5595)	**27.53**	(20.46,34.6)

The best result is marked in bold while the second and third best results are marked with 2nd and 3rd, respectively.

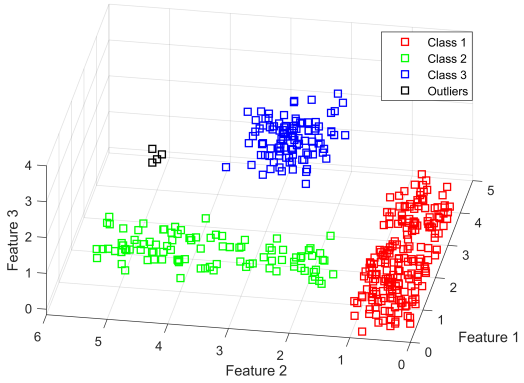


Figure 8: Lsun3D dataset.

- The leftmost column of Table 3 gives the proportion of outliers included in the three clusters in the sample. Note the effects of increasing the number of outliers as we move down the rows of the table. For any method, both the misclassification and deviation of the computed prototypes from the true prototypes steadily increase with the number of outliers. The values of the indexes indicate that resistance to outliers is slightly greater for FCM-F-L1 than for comparative models. We can observe that the performance of FCM-F-L1 decreases more slowly than comparative models when the percentage of outliers increases.
- It can be observed from Table 4 that, whatever the scheme (center, spreads, and center and spreads) and either considering the ability to classify correctly or considering the ability to recover the true prototypes, comparative models are less robust to the presence of outliers than the FCM-F-L1 model.

4.2 Benchmark datasets

For the experiments on datasets with complex structures, two well-known benchmark datasets named Lsun3D and Wine are used. The fuzziness parameter m is chosen to be equal to 1.1, 1.3, and 1.5. All crisp datasets are subjected to a fuzzification process as follows. Each object of the fuzzified dataset (a triangular fuzzy datum) is a vector of intervals, i.e., a rectangle (when $p = 2$) or a hyperrectangle (when $p \geq 3$): $([a_{1k1} - \underline{a}_{k1}, a_{1k1} + \overline{a}_{k1}], [a_{1k2} - \underline{a}_{k2}, a_{1k2} + \overline{a}_{k2}], \cdots, [a_{1kp} - \underline{a}_{kp}, a_{1kp} + \overline{a}_{kp}])$, $k = 1, 2, \cdots, n$. We assumed $a_{1kj} = z_{kj}$ and $\underline{a}_{kj}, \overline{a}_{kj} \sim \mathcal{N}(1, 0.09)$, $j = 1, 2, \cdots, p$, where z_{kj} is the j^{th} feature observed on the k^{th} object of crisp dataset.

4.2.1 Lsun3D

The Lsun3D dataset [37, section 9.1.8], shown in Figure 8, consists of one full sphere (100 samples), two bricks at a perpendicular angle to each other (200 and 100 samples), and four outliers in $\mathbb{R}^3$ [38].

4.2.2 Downsampled wine

The well-known Wine dataset, obtained from the UCI (University of California, Irvine) machine learning repository http://archive.ics.uci.edu/ml/, is a real-world benchmark dataset having 178 objects, 13 features, and 3 classes (59, 71, and 48 samples). Here, classes 2 and 3 are used as inliers and class 1 is downsampled to 10 instances to be used as outliers [32] (see, Figure 9).

The clustering results corresponding to the fuzzified Lsun3D and fuzzified downsampled Wine datasets are given in Tables 5 and 6, respectively. FCM-F-L1 model presents the best performance for all the cases, except one (the Lsun3D dataset, $m = 1.5$). For clusters of the Lsun3D dataset, which have sharp or boxy edges, FCM-F-L1 yields better results.

4.3 Application

A locus (plural loci) is a specific, fixed position on a chromosome where a particular gene or genetic marker is located [39]. Genes may possess multiple variants known as alleles. In this section, to experiment on a categorical dataset, we apply FCM-F-L1 to the Tetragonula dataset which was published in [19] and gives genetic information, pairs of alleles (codominant markers) for 13 loci, about 236 bees. Alleles have a three-digit code. Code 000 refers to missing values

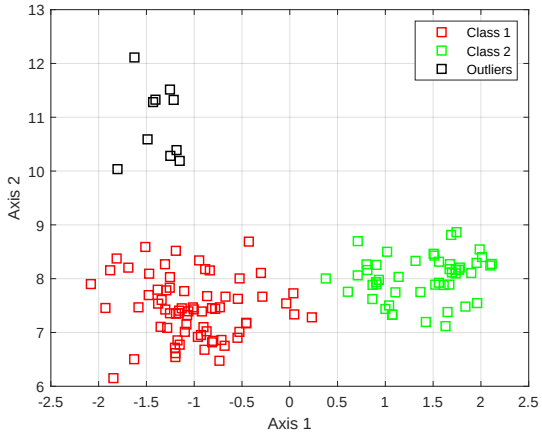


Figure 9: Visualization of the downsampled Wine dataset using Linear Discriminant Analysis.

Table 5: The clustering results corresponding to the fuzzified Lsun3D dataset.

m	Method	frand		HUL		REC	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
1.1	FkMedC-F 2014	0.8298	(0.8073,0.8522)	0.7586	(0.7269,0.7903)	8.98	(7.3,10.65)
	SFkMedC-F 2014	0.8072	(0.7837,0.8307)	0.7277	(0.695,0.7604)	9.74	(8.07,11.42)
	FkMedC-NC-F 2014	0.5578	(0.5453,0.5704)	0.3795	(0.3621,0.3969)	22.46	(20.25,24.67)
	TrFkMedC-F 2014	0.7325	(0.7102,0.7548)	0.6181	(0.5887,0.6475)	17.63	(15.68,19.59)
	FcOMdC-FD, SQR 2020	0.8283	(0.8059,0.8507)	0.7569	(0.7254,0.7883)	9.18	(7.51,10.86)
	FcOMdC-FD, LIN 2020	0.8441	(0.8228,0.8655)	0.7781	(0.7482,0.808)	8.44	(7,9.87)
	FcOMdC-FD, SIG 2020	0.8040	(0.783,0.8251)	0.7218	(0.692,0.7516)	10.08	(8.61,11.55)
	FcOMdC-FD, HUB 2020	0.8329	(0.8115,0.8543)	0.7629	(0.7327,0.7932)	8.83	(7.19,10.48)
	FcOMdC-FD, LOG 2020	0.8474	(0.8268,0.868)	0.7827	(0.7538,0.8117)	7.92	(6.57,9.26)
	FCM-F-L1	**0.9830**	(0.9829,0.9831)	**0.9736**	(0.9734,0.9737)	**0.33**	(0.32,0.34)
1.3	FkMedC-F 2014	0.7826	(0.7596,0.8057)	0.6899	(0.6578,0.7219)	9.73	(8.32,11.14)
	SFkMedC-F 2014	0.7204	(0.6967,0.7441)	0.5947	(0.563,0.6263)	13.00	(11.26,14.74)
	FkMedC-NC-F 2014	0.5507	(0.5415,0.56)	0.3877	(0.3756,0.3998)	21.74	(19.54,23.93)
	TrFkMedC-F 2014	0.6755	(0.6555,0.6956)	0.5250	(0.4989,0.5511)	18.80	(16.84,20.76)
	FcOMdC-FD, SQR 2020	0.7886	(0.7663,0.811)	0.6994	(0.6685,0.7303)	9.63	(8.27,11)
	FcOMdC-FD, LIN 2020	0.7812	(0.7584,0.8039)	0.6794	(0.649,0.7098)	9.72	(8.34,11.1)
	FcOMdC-FD, SIG 2020	0.7756	(0.7526,0.7985)	0.6820	(0.6506,0.7134)	10.23	(8.84,11.62)
	FcOMdC-FD, HUB 2020	0.8114	(0.789,0.8338)	0.7321	(0.7014,0.7628)	8.40	(7.01,9.79)
	FcOMdC-FD, LOG 2020	0.7744	(0.7533,0.7955)	0.6737	(0.645,0.7025)	11.18	(9.84,12.52)
	FCM-F-L1	**0.9143**	(0.9049,0.9237)	**0.8525**	(0.8406,0.8644)	**0.81**	(0.26,1.36)
1.5	FkMedC-F 2014	0.7610	(0.7395,0.7825)	0.6430	(0.6133,0.6728)	9.38	(8,10.77)
	SFkMedC-F 2014	0.6956	(0.676,0.7152)	0.5323	(0.5074,0.5573)	12.06	(10.36,13.77)
	FkMedC-NC-F 2014	0.5366	(0.5299,0.5433)	0.3602	(0.3515,0.3689)	21.33	(19.3,23.35)
	TrFkMedC-F 2014	0.6281	(0.6102,0.646)	0.4390	(0.4153,0.4627)	19.93	(17.82,22.03)
	FcOMdC-FD, SQR 2020	0.7544	(0.7328,0.776)	0.6435	(0.6144,0.6726)	10.38	(9.05,11.71)
	FcOMdC-FD, LIN 2020	0.7153	(0.6954,0.7352)	0.5586	(0.5329,0.5842)	10.97	(9.54,12.41)
	FcOMdC-FD, SIG 2020	0.7382	(0.7189,0.7574)	0.6109	(0.5853,0.6365)	10.58	(9.35,11.81)
	FcOMdC-FD, HUB 2020	0.7637	(0.7424,0.785)	**0.6566**	(0.6282,0.6851)	10.25	(8.85,11.65)
	FcOMdC-FD, LOG 2020	0.7458	(0.725,0.7667)	0.6144	(0.5862,0.6426)	10.84	(9.41,12.28)
	FCM-F-L1	**0.7669**	(0.7464,0.7874)	0.6107	(0.5848,0.6366)	**4.80**	(3.39,6.21)

The best result is marked in bold.

Table 6: The clustering results corresponding to the fuzzified Wine downsampled dataset.

m	Method	frand		HUL		REC	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
1.1	FkMedC-F 2014	0.5235	(0.523,0.5241)	0.3262	(0.3254,0.327)	184884	(166433,203336)
	SFkMedC-F 2014	0.5569	(0.5568,0.5571)	0.3734	(0.3732,0.3736)	16173	(15792,16555)
	FkMedC-NC-F 2014	0.5206	(0.517,0.5241)	0.3218	(0.3169,0.3268)	140954	(83374,198534)
	TrFkMedC-F 2014	0.5417	(0.5379,0.5455)	0.3518	(0.3464,0.3572)	105535	(60982,150089)
	FcOMdC-FD, SQR 2020	0.5234	(0.5227,0.5241)	0.3262	(0.3254,0.327)	179462	(161262,197663)
	FcOMdC-FD, LIN 2020	0.5536	(0.551,0.5563)	0.3688	(0.3652,0.3725)	41132	(25812,56452)
	FcOMdC-FD, SIG 2020	0.6758	(0.6506,0.7009)	0.5418	(0.5061,0.5775)	163339	(119357,207322)
	FcOMdC-FD, HUB 2020	0.5518	(0.5485,0.5551)	0.3661	(0.3616,0.3707)	58118	(40765,75470)
	FcOMdC-FD, LOG 2020	0.6676	(0.6486,0.6865)	0.5341	(0.5072,0.5611)	14080	(11428,16732)
	FCM-F-L1	**0.9150**	(0.9149,0.915)	**0.8830**	(0.8829,0.883)	**884.8**	(884.2,885.4)
1.3	FkMedC-F 2014	0.5247	(0.5237,0.5256)	0.3280	(0.327,0.329)	179620	(162325,196915)
	SFkMedC-F 2014	0.5581	(0.5581,0.5581)	0.3747	(0.3747,0.3747)	14513	(14513,14513)
	FkMedC-NC-F 2014	0.5161	(0.5134,0.5187)	0.3164	(0.3127,0.32)	137434	(78300,196568)
	TrFkMedC-F 2014	0.5382	(0.5345,0.542)	0.3473	(0.3424,0.3522)	119318	(73721,164916)
	FcOMdC-FD, SQR 2020	0.5256	(0.525,0.5262)	0.3289	(0.3281,0.3297)	171655	(156084,187226)
	FcOMdC-FD, LIN 2020	0.5540	(0.5525,0.5556)	0.3695	(0.3678,0.3711)	14126	(13809,14443)
	FcOMdC-FD, SIG 2020	0.5749	(0.5619,0.5879)	0.4042	(0.3856,0.4228)	204639	(155297,253981)
	FcOMdC-FD, HUB 2020	0.5559	(0.554,0.5579)	0.3722	(0.3702,0.3743)	14882	(14450,15314)
	FcOMdC-FD, LOG 2020	0.5603	(0.5515,0.5692)	0.3884	(0.3753,0.4015)	14491	(12833,16149)
	FCM-F-L1	**0.7469**	(0.7468,0.7471)	**0.6621**	(0.6619,0.6622)	**912.777**	(912.775,912.779)
1.5	FkMedC-F 2014	0.5264	(0.5259,0.527)	0.3300	(0.3292,0.3307)	157521	(144719,170323)
	SFkMedC-F 2014	0.5530	(0.5515,0.5545)	0.3678	(0.3663,0.3694)	14519	(14510,14527)
	FkMedC-NC-F 2014	0.5124	(0.5107,0.5142)	0.3137	(0.3114,0.3161)	144862	(85536,204188)
	TrFkMedC-F 2014	0.5323	(0.5285,0.536)	0.3410	(0.3364,0.3455)	105608	(60210,151006)
	FcOMdC-FD, SQR 2020	0.5250	(0.5241,0.5259)	0.3283	(0.3272,0.3293)	182437	(164956,199919)
	FcOMdC-FD, LIN 2020	0.5396	(0.5369,0.5423)	0.3517	(0.349,0.3544)	12490	(11638,13341)
	FcOMdC-FD, SIG 2020	0.5411	(0.5336,0.5487)	0.3616	(0.3508,0.3723)	170891	(127882,213899)
	FcOMdC-FD, HUB 2020	0.5420	(0.5393,0.5447)	0.3544	(0.3518,0.3571)	12576	(11551,13602)
	FcOMdC-FD, LOG 2020	0.5374	(0.5324,0.5423)	0.3585	(0.3511,0.3658)	13887	(11930,15844)
	FCM-F-L1	**0.6383**	(0.6382,0.6383)	**0.5099**	(0.5098,0.51)	**659**	(640,679)

The best result is marked in bold.

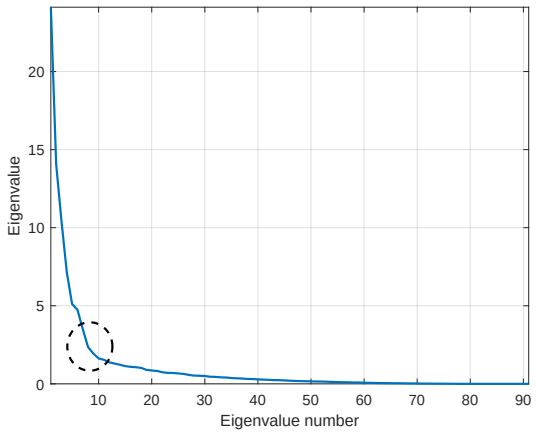


Figure 10: Plotting the eigenvalues of YY' as a function of the eigenvalue numbers.

[21]. The ground truth species labels (9 classes) are given by Franck et al. [19]. Here, class 4 is marked as outlier, while all other classes are inliers.

Bowcock et al. [6] defined the Shared Allele Distance (SAD) between pairs of individuals as one minus the proportion of alleles that they share averaged over loci. Loci with 000 values are not considered in the pairwise distance calculation.

Example 4.1. *Two individuals below share 9 of their 24 allele copies, so the SAD is* $1 - \frac{9}{2\times 13-2} = 0.625$. *Two alleles are missing in individual 167.*

Individual 98:	*Individual 167:*	*Matches:*
173176	191191	
143143	135147	
157159	151151	
111111	111109	111111
145145	145148	145145
170170	160154	
144144	144144	144144, 144144
116116	114114	
206208	190194	
256256	256256	256256, 256256
119119	119119	119119, 119119
108106	112108	108108
166166	000000	

The cmdscale function in MATLAB takes an $n \times n$ distance matrix D and returns an $n \times p$ matrix Y. Rows of Y are the coordinates of n objects in p-dimensional space. It also returns the eigenvalues of YY'. If the first k eigenvalues are much larger than the remaining $n - k$, then one can use the first k columns of Y as k-dimensional objects whose distance matrix approximate D.

We use SAD to compute the distance matrix D for the present dataset. It is 236×236 and symmetric, has values in $[0, 1]$ off the diagonal, and has zeros on the diagonal. The eight largest positive eigenvalues are much larger in magnitude than the remaining eigenvalues (see the elbow point in Figure 10). So, the first eight coordinates of Y are sufficient for a reasonable reproduction of D. We code the resulting data in a fuzzy manner by considering a symmetric triangular membership function with spreads equal to $(\mathcal{U}[0, 0.02], \mathcal{U}[0, 0.02], \cdots, \mathcal{U}[0, 0.02]) \in \mathbb{R}^8$. The fuzziness parameter m is chosen to be equal to 1.1. The clustering results corresponding to the Tetragonula dataset are listed in Table 7.

5 Conclusions

In this paper, we introduced a robust dissimilarity measure between each pair of fuzzy data defined as an adaptive weighted sum of the L_1-norms of the centers and the spreads. We also proposed a robust fuzzy clustering model for L-R fuzzy data, namely FCM-F-L1, which is based on the introduced dissimilarity measure.

Table 7: The clustering results corresponding to the Tetragonula dataset.

Method	frand		HUL		REC	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
FkMedC-F 2014	0.8957	(0.8883,0.9031)	0.8212	(0.8084,0.8339)	3.22	(3.04,3.39)
SFkMedC-F 2014	0.8747	(0.8646,0.8849)	0.7672	(0.7498,0.7847)	3.47	(3.27,3.67)
FkMedC-NC-F 2014	0.7626	(0.7478,0.7774)	0.6562	(0.6364,0.676)	4.17	(3.96,4.38)
TrFkMedC-F 2014	0.8357	(0.8233,0.8481)	0.6907	(0.6711,0.7102)	4.10	(3.89,4.3)
FcOMdC-FD, SQR 2020	0.8975	(0.8899,0.9051)	0.8267	(0.8143,0.8392)	3.20	(3.01,3.39)
FcOMdC-FD, LIN 2020	0.8583	(0.847,0.8695)	0.7320	(0.7125,0.7516)	3.47	(3.27,3.67)
FcOMdC-FD, SIG 2020	0.7749	(0.762,0.7878)	0.4881	(0.4676,0.5086)	3.60	(3.42,3.78)
FcOMdC-FD, HUB 2020	0.8955	(0.8874,0.9037)	0.8228	(0.8096,0.836)	3.14	(2.95,3.34)
FcOMdC-FD, LOG 2020	0.8959	(0.8876,0.9041)	0.8163	(0.8021,0.8305)	3.18	(3,3.36)
FCM-F-L1	**0.9074**	(0.9017,0.9132)	**0.8508**	(0.8416,0.8601)	**2.62**	(2.49,2.76)

The best result is marked in bold.

The FCM-F-L1 algorithm is performed in three steps. The update step computes fuzzy prototypes. The weighting step is where a relevance weight for each fuzzy feature is learned and if outlier fuzzy features are present in the dataset, weights close to 0 are given to neutralize their influence. Finally, the assignment step computes the membership degrees.

We have found that FCM-F-L1 worked in a satisfactory way also in comparison with its competitors. We observed that, for any method, both the misclassification and deviation of the computed prototypes from the true prototypes steadily increased with the number of outliers. Nevertheless, the performance of FCM-F-L1 decreased more slowly than comparative models when the percentage of outliers increased. Whatever the scheme (center, spreads, and center and spreads) and either considering the ability to classify correctly or considering the ability to recover the true prototypes, comparative models were less robust to the presence of outliers than the FCM-F-L1 model. For clusters of the Lsun3D dataset, which have sharp or boxy edges, FCM-F-L1 yielded better results. The complexity of our proposed method is $O(n \times \log(n))$, while that of other comparative methods is $O(n^2)$.

As a direction for a further research, we are working on generalizing the proposed method to the full fuzzy numbers. Another area of research is to find criteria to select v, which measures the importance of the distance among the spreads.

Acknowledgement

The authors would like to thank the editor in chief, associate editor, and reviewers for their valuable comments and suggestions for improving this manuscript.

References

[1] J. R. Alharbi, W. S. Alhalabi, *Hybrid approach for sentiment analysis of twitter posts using a dictionary-based approach and fuzzy logic methods: Study case on cloud service providers*, International Journal on Semantic Web and Information Systems (IJSWIS), **16**(1) (2020), 116-145.

[2] M. A. Alsmirat, Y. Jararweh, M. Al-Ayyoub, M. A. Shehab, B. B. Gupta, *Accelerating compute intensive medical imaging segmentation algorithms using hybrid CPU-GPU implementations*, Multimedia Tools and Applications, **76** (2017), 3537-3555.

[3] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithm*, Plenum Press, New York, 1981.

[4] P. Bloomfield, W. L. Steiger, *Least absolute deviations*, Birkh Iuser, Boston, MA, 1983.

[5] L. Bobrowski, J. C. Bezdek, *C-means clustering with the l_1 and l_∞ norms*, IEEE Transactions on Systems, Man, and Cybernetics, **21**(3) (1991), 545-554.

[6] A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, L. L. Cavalli-Sforza, *High resolution of human evolutionary trees with polymorphic microsatellites*, Nature, **368** (1994), 455-457.

[7] B. S. Butkiewicz, *Robust fuzzy clustering with fuzzy data*, in: P. S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Proceedings of Advances in Web Intelligence, Third International Atlantic Web Intelligence Conference, AWIC 2005, Lecture Notes in Computer Science, Vol. **3528**, Springer, Berlin, Heidelberg, (2005), 76-82.

[8] R. J. G. B. Campello, *A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment*, Pattern Recognition Letters, **28**(7) (2007), 833-841.

[9] R. Coppi, P. D'Urso, P. Giordani, *Fuzzy and possibilistic clustering for fuzzy data*, Computational Statistics and Data Analysis, **56**(4) (2012), 915-927.

[10] R. N. Dave, *Characterization and detection of noise in clustering*, Pattern Recognition Letters, **12**(11) (1991), 657-664.

[11] F. de A. T. De Carvalho, E. C. Simoes, *Fuzzy clustering of interval-valued data with City-Block and Hausdorff distances*, Neurocomputing, **266** (2017), 659-673.

[12] M. Deveci, D. Pamucar, I. Gokasar, M. Köppen, B. B. Gupta, *Personal mobility in metaverse with autonomous vehicles using Q-rung orthopair fuzzy sets based OPA-RAFSI model*, IEEE Transactions on Intelligent Transportation Systems, (2022), 1-10.

[13] P. D'Urso, L. De Giovanni, *Robust clustering of imprecise data*, Chemometrics and Intelligent Laboratory Systems, **136** (2014), 58-80.

[14] P. D'Urso, M. Disegna, R. Massari, *Fuzzy clustering in travel and tourism analytics*, in: Business and Consumer Analytics: New Ideas, Springer, (2019), 839-863.

[15] P. D'Urso, P. Giordani, *A weighted fuzzy c-means clustering model for fuzzy data*, Computational Statistics and Data Analysis, **50**(6) (2006), 1496-1523.

[16] P. D'Urso, J. M. Leski, *Fuzzy clustering of fuzzy data based on robust loss functions and ordered weighted averaging*, Fuzzy Sets and Systems, **389** (2020), 1-28.

[17] E. Eskandari, A. Khastan, S. Tomasiello, *Improved determination of the weights in a clustering approach based on a weighted dissimilarity measure between fuzzy data*, 2022 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), Padua, Italy, (2022), 1-6.

[18] M. B. Ferraro, P. Giordani, *Possibilistic and fuzzy clustering methods for robust analysis of non-precise data*, International Journal of Approximate Reasoning, **88** (2017), 23-38.

[19] P. Franck, E. Cameron, G. Good, J. Y. Rasplus, B. Oldroyd, *Nest architecture and genetic differentiation in a species complex of Australian stingless bees*, Molecular Ecology, **13**(8) (2004), 2317-2331.

[20] R. J. Hathaway, J. C. Bezdek, Y. K. Hu, *Generalized fuzzy c-means clustering strategies using Lp norm distances*, IEEE Transactions on Fuzzy Systems, **8**(5) (2000), 576-582.

[21] C. Hennig, *How many bee species? A case study in determining the number of clusters*, in: M. Spiliopoulou, L. Schmidt-Thieme, R. Janning (Eds.), Data Analysis, Machine Learning and Knowledge Discovery, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, (2014), 41-49.

[22] E. Hullermeier, M. Rifqi, S. Henzgen, R. Senge, *Comparing fuzzy partitions: A generalization of the Rand index and related measures*, IEEE Transactions on Fuzzy Systems, **20**(3) (2012), 546-556.

[23] W. Hung, M. Yang, *Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation*, Fuzzy Sets and Systems, **150**(3) (2005), 561-577.

[24] W. Hung, M. Yang, E. Lee, *A robust clustering procedure for fuzzy data*, Computers and Mathematics with Applications, **60** (2010), 151-165.

[25] K. Jajuga, *L_1-norm based fuzzy clustering*, Fuzzy Sets and Systems, **39** (1991), 43-50.

[26] S. Jin, *A bidirectional reasoning based on fuzzy interpolation*, International Journal of Software Science and Computational Intelligence (IJSSCI), **12**(1) (2020), 1-14.

[27] L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, Wiley, Hoboken, NJ, 2005.

[28] V. D. Minh, T. T. Ngan, T. M. Tuan, V. T. Duong, N. T. Cuong, *An improvement in integrating clustering method and neural network to extract rules and application in diagnosis support*, Iranian Journal of Fuzzy Systems, **19**(5) (2022), 147-165.

[29] N. M. Ralevic, M. Delic, Lj. Nedovic, *Aggregation of fuzzy metrics and its application in image segmentation*, Iranian Journal of Fuzzy Systems, **19**(3) (2022), 19-37.

[30] A. B. Ramos-Guajardo, M. B. Ferraro, *A fuzzy clustering approach for fuzzy data based on a generalized distance*, Fuzzy Sets and Systems, **389** (2020), 29-50.

[31] W. Rhmann, *An ensemble of hybrid search-based algorithms for software effort prediction*, International Journal of Software Science and Computational Intelligence (IJSSCI), **13**(3) (2021), 28-37.

[32] S. Sathe, C. C. Aggarwal, *LODES: Local density meets spectral outlier detection*, in: Proceedings of the 2016 SIAM International Conference on Data Mining (SDM), (2016), 171-179.

[33] M. Sato, Y. Sato, *Fuzzy clustering model for fuzzy data*, in: Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and the Second International Fuzzy Engineering Symposium, Proceedings of 1995 IEEE Int., Vol. **4**, IEEE, (1995), 2123-2128.

[34] B. Sinova, M. A. Gil, A. Colubi, S. Van Aelst, *The median of a random fuzzy number. The 1-norm distance approach*, Fuzzy Sets and Systems, **200** (2012), 99-115.

[35] B. Sinova, S. R. de Saa, M. A. Gil, *A generalized L1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data*, Information Sciences, **242** (2013), 22-34.

[36] V. V. Tai, L. D. Nghiep, *Interpolating time series based on fuzzy cluster analysis problem*, Iranian Journal of Fuzzy Systems, **17**(3) (2020), 151-161.

[37] M. C. Thrun, *Projection based clustering through self-organization and swarm intelligence*, Springer, Heidelberg, 2018.

[38] M. C. Thrun, A. Ultsch, *Clustering benchmark datasets exploiting the fundamental clustering problems*, Data in Brief, **30** (2020), 105501.

[39] E. J. Wood, *The encyclopedia of molecular biology*, Biochemical Education, **23**(2) (1995), 1165.

[40] L. A. Zadeh, *Fuzzy sets*, Information and Control, **8** (1965), 338-353.

[41] M. F. Zarandi, Z. S. Razaee, *A fuzzy clustering model for fuzzy data with outliers*, International Journal of Fuzzy System Applications (IJFSA), **1**(2) (2010), 29-42.