International Journal of Industrial Engineering & Production Research June 2022 Vol. 33, No. 2: 1-12
DOI: 10.22068/ijiepr.33.2.13

RESEARCH PAPER

Clustering of Fuzzy Data Sets Based on Particle Swarm Optimization with Fuzzy Cluster Centers

Najme Ghanbari[1], Hadi Shahraki[2] & Seyed-Hamid Zahiri*[3]

Received 14 September 2021; Revised 10 January 2022; Accepted 26 February 2022;

ABSTRACT
In the current study, a particle swarm clustering method is suggested for clustering of triangular fuzzy data. This clustering method can find the centers of the fuzzy cluster. Whatever the centers of the fuzzy cluster have more points from the corresponding cluster, clustering accuracy increases. Triangular fuzzy numbers are utilized to demonstrate uncertain data. To compare triangular fuzzy numbers, a similarity criterion based on the intersection region of the fuzzy numbers is used. The performance of the suggested clustering method has experimented on both benchmark and artificial datasets. These datasets are used in the fuzzy form. The experiential results represent that the suggested clustering method with fuzzy cluster centers can cluster triangular fuzzy datasets like or superior to other standard uncertain data clustering methods. Experimental results demonstrate that, in almost all datasets, the proposed clustering method provides better results in accuracy when compared to Uncertain K-Means and Uncertain K-medoids algorithms.

KEYWORDS: *Clustering; Particle swarm clustering method; Uncertain data; Triangular fuzzy data; Similarity value.*

1. Introduction

The particle swarm optimization (PSO) algorithm is a swarm intelligence algorithm [1,2]. PSO algorithm is an effort to discover the optimal solution via the simulation of some concepts obtained from bird flocking, bees, fish, and other collective folks. Each particle can efficiently attain his objective using the data that is owned by him and the information that is assigned by the folk. This confirms that particle swarm optimization is an optimization procedure that follows the laws of collective treatment. PSO algorithm has been used in several engineering and optimization problems [3-6]. Also, some PSO algorithms have been proposed to solve pattern recognition problems [7,8].

Data clustering is one of the most essential data mining [9]. Data clustering using PSO was first suggested by Engelbrecht and Merwe in [10]. The idea of the PSO clustering method was to allocate all the centers of the cluster to each particle and update the particle by the fitness value computed by each particle. This clustering method is utilized on many occasions, and its conclusions demonstrate that this clustering method is possible to utilize in various utilizations [3,4].

The particle swarm optimization has been utilized only for crisp data until now. At the same time, in real-world applications such as biomedical, microarray, and sensor measurements, data is mainly demonstrated by uncertain data. Uncertain data cannot express by crisp numbers.

Uncertain data can utilize in various cases. Some amounts should express by interval data, for instance, the domain of the temperature during a specified time.

Also, uncertainty may conclude from lack of information, implicit randomness in data generation, data staling, and weakness to do good physical measurements or ambiguity [11]. It is commonly relevant to lacking or incomplete knowledge or the possibility of incidence of given data and expertise [12]. Hence, it is a serious subject to regard uncertain data in a clustering algorithm.

* **Corresponding author:** ***Seyed-Hamid Zahiri***
hzahiri@birjand.ac.ir

1. *Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.*
2. *Department of Computer Engineering, Faculty of Industry and Mining, University of Sistan and Baluchestan, Khash, Iran.*
3. *Department of Electrical Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran.*

Clustering of uncertain data is regarded in different articles [13-19]. In several papers, interval data is utilized to demonstrate uncertain data [13]. In [20], interval data clustering based on the adaptive dynamic cluster method is described. In [21], adaptive Hausdorff intervals and dynamic clustering cluster interval data. In [22], a novel clustering method based on novel density and hierarchical density is proposed for interval data. Likewise, using interval data in various methods and applications is explained in [13].

Utilizing fuzzy numbers is a popular alternative method to illustrate uncertain data. Tayyebi et al. suggested a fuzzy clustering method (FCM), which clusters trapezoidal fuzzy numbers [23]. In the mentioned paper, a linear ranking function is applied to specify a distance for a fuzzy number. This method can generalize to real data.

Clustering of uncertain data with heuristic methods is one of the topics that have not yet been investigated. So, in this paper, for the first time, the particle swarm clustering procedure is applied to find the centers of the optimal fuzzy cluster in the clustering of uncertain data. In other words, the aim of this work is to upgrade the particle swarm clustering method to cluster triangular fuzzy data. In the proposed method, whatever the centers of the fuzzy cluster have more points from the corresponding cluster, clustering accuracy increases.

In the proposed method, the cluster centers and data are fuzzy. Therefore, it is necessary to determine the similarity degree of two fuzzy numbers.

Calculations related to determining the similarity of two fuzzy numbers are done in three phases [24]. The intersection region between two fuzzy numbers is getting, firstly. In the second phase, the value of the shape of the intersection region is computed. Eventually, a similarity value of two fuzzy numbers is acquired using the similarity equation. How to compute the similarity between two triangular fuzzy numbers in one dimension is briefly explained in section 3.

Since any real number or interval number is a particular kind of triangular fuzzy number, the proposed clustering procedure can cluster uncertain data sets of the type interval, fuzzy or real data.

The remainder of the current article is constituted as follows: Section 2 presents some preliminary concepts of fuzzy theory and fuzzy numbers. The proposed PSO algorithm for clustering triangular fuzzy data with fuzzy cluster centers is expressed in section 3. Section 4 reported the experiential clustering outcomes. At last, Section 5 explains the conclusion.

2. Preliminary Concepts

In this part, essential concepts of fuzzy sets and theory, initialized by Bellman and Zadeh in [23], are reviewed to apply throughout this paper. The following impression and definitions are obtained from [25].

Suppose G is a universal set. $\tilde{a}$: G→ [0,1] is a mapping and shows a fuzzy set. The value $\tilde{a}$ (g) of $\tilde{a}$ at g∈G stands for the degree of membership of g in $\tilde{a}$. A fuzzy set $\tilde{a}$ is normal if there be g_0 ∈G such that $\tilde{a}$ (g_0) = 1. An alpha-cut of fuzzy number $\tilde{a}$, alpha ∈ [0,1], is a crisp set as

$$\text{alpha} = \{g \in G: \tilde{a}(g) \geq \text{alpha}\} \quad (1)$$

If a fuzzy set $\tilde{a}$ satisfies that $\tilde{a}_{\text{alpha}}$ is a closed interval for every alpha ∈ [0,1], then $\tilde{a}$ is called a fuzzy number.

Trapezoidal fuzzy number (TFN), $\tilde{a}$= (a^1, a^2, a^3, a^4) is determined as:

$$\tilde{a}_{\text{alpha}} = \begin{cases} \frac{g-a^1}{a^2-a^1} & a^1 < g \leq a^2 \\ 1 & a^2 < g \leq a^3 \\ \frac{a^4-g}{a^4-a^3} & a^3 < g \leq a^4 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

To reduce the TFN $\tilde{a}$ is illustrated by (a^1, a^2, a^3, a^4) (view Fig. 1.a). If $a^2 = a^3$, then $\tilde{a}$ is named a triangular fuzzy number. For example, (1.5, 2, 2, 2.5) and (1.7, 2, 3, 3.4) are trapezoidal fuzzy numbers, which may exert to qualify the fuzzy significance of around number 2 and around interval [2,3], respectively. The set of all trapezoidal fuzzy numbers is specified by F(R). Because any real number c and any interval [a, b] can be written as (c, c, c, c) and (a, a, b, b), respectively, it is evident that TFNs are an expansion of real numbers and intervals. A trapezoidal fuzzy vector $\tilde{a}$ is a member of the cartesian product F^n(R). Fig. 1.b shows an exhibition of vector ($\tilde{a}$, $\tilde{b}$) ∈F^2(R). The black areas offer full membership and the gray areas partial membership. The representations of different numbers are shown in Fig. 2.

Next, arithmetic operations on trapezoidal fuzzy numbers is described. Suppose $\tilde{p}$ = (p^1, p^2, p^3, p^4) and $\tilde{q}$ = (q^1, q^2, q^3, q^4) be two TFNs and w be a real value. The addition and scalar production operators are described as follows:

$$-w.\tilde{p} = (wp^1, wp^2, wp^3, wp^4) \quad \text{if} \quad w \geq 0, w \in R \qquad (3)$$

$$-w.\tilde{p} = (wp^4, wp^3, wp^2, wp^1) \quad \text{if} \quad w \leq 0, w \in R \qquad (4)$$

$$-\tilde{p} + \tilde{q} = (p^1 + q^1, p^2 + q^2, p^3 + q^3, p^4 + q^4) \qquad (5)$$

$$-\tilde{p} - \tilde{q} = (p^1 - q^4, p^2 - q^3, p^3 - q^2, p^4 - q^1) \qquad (6)$$

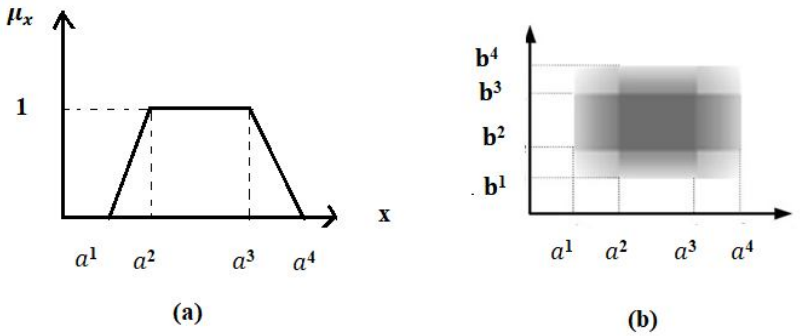

Fig. 1. a) Membership function of TFN $\tilde{a}$= (a^1, a^2, a^3, a^4), b) exhibition of ($\tilde{a}$, $\tilde{b}$) ∈F^2(R).

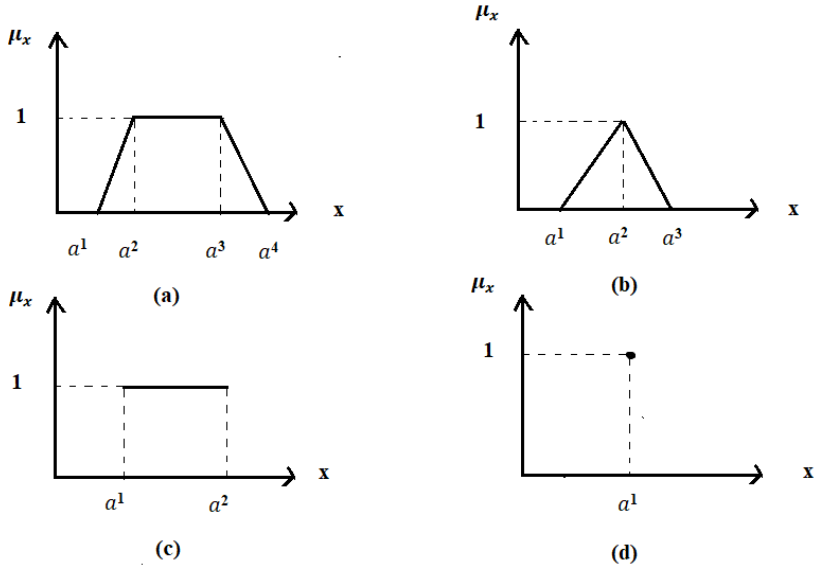

Fig. 2. Types of fuzzy numbers a) Trapezoidal, b) Triangular, c) Interval, and d) Crisp.

3. Proposed Clustering Method

Clustering of uncertain data with heuristic methods is one of the topics that have not yet been investigated. In the current section, a new particle swarm optimization algorithm is offered to cluster patterns whose features are triangle fuzzy numbers. This clustering method can find fuzzy cluster centers. In the proposed method, each particle provides a complete solution to solve the intended problem.

Assumed that $\tilde{Z}_p = [\tilde{Z}_{p1}, \tilde{Z}_{p2}, \dots, \tilde{Z}_{pn}]$ is a dataset where $\tilde{Z}_{pk} = [\tilde{Z}_{p1k}, \tilde{Z}_{p2k}, \dots, \tilde{Z}_{ptk}]^T$, k=1, 2, 3, ..., n and $\tilde{Z}_{lk} = [\tilde{Z}^1_{lk}, \tilde{Z}^2_{lk}, \tilde{Z}^3_{lk}] \in F(R)$ for all k=1, 2, ….,n and l=1, 2, …, t. Thus, $\tilde{Z}_p$is a part of $F^{t\times n}(R)$ .The goal is to division $\tilde{x}_k{}'s$ into c clusters. Also, $\tilde{m}_{ij} = [\tilde{m}_{ij1}, \tilde{m}_{ij2}, \dots, \tilde{m}_{ijN_c}]$ is a fuzzy matrix of the prototype where $\widetilde{mI}_{J_i} = [\widetilde{mI}_{J_{1i}}, \widetilde{mI}_{J_{2i}}, \dots, \widetilde{mI}_{J_{ti}}]^T$, i=1, 2, …, N_c and $\tilde{m}_{li} = [\tilde{m}^1_{li}, \tilde{m}^2_{li}, \tilde{m}^3_{li}] \in F(R)$ for all i=1, 2, …, N_c and l=1, 2, …, t.

In the following of the suggested clustering procedure, the degree of the similarity of all samples with each of the random cluster centers offered by the particle swarm algorithm is obtained. Each sample is devoted to the cluster whose center of the cluster is most similar to that sample. If the similarity of a sample with all clusters is zero, that sample is considered noise.

In the proposed method, clusters centers and data are fuzzy. Therefore, it is necessary to determine the similarity degree of two fuzzy numbers.

Calculations related to determining the similarity of two fuzzy numbers are done in three phases. The intersection region between two fuzzy numbers is getting, firstly. In the second phase,

the value of the shape of the intersection region is computed, and eventually, a similarity value of two fuzzy numbers is acquired using the similarity equation.

In the first step, the intersection points of the common area of two triangular fuzzy numbers are obtained. To find the intersection points, equations of the triangular fuzzy number line are obtained. With the help of line equations, all intersection points, if any, are obtained. Fig. 3 shows several different intersection area cases for one-dimensional triangular fuzzy numbers.

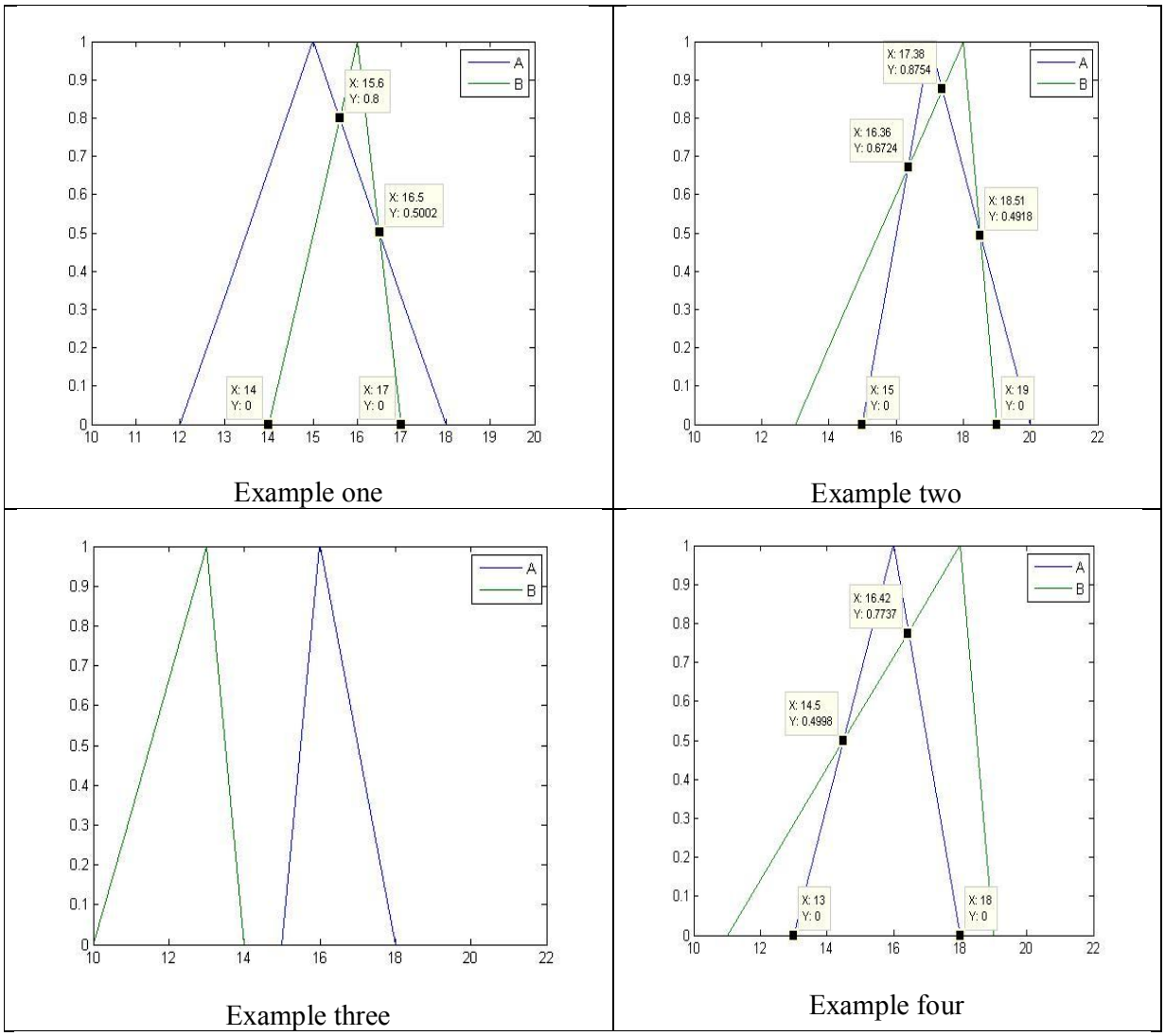

Fig. 3. Intersection area for some examples of triangular fuzzy numbers A and B.

In the second step, the area of the intersection of two TFNs is computed using the intersection points obtained from the previous step and equation (7).

$$\text{Area} = \left|\frac{(x_1y_2 - x_2y_1) + (x_2y_3 - x_3y_2) + \cdots + (x_ny_1 - x_1y_n)}{2}\right| \quad (7)$$

In equation (7), xi and yi are the coordinates of the intersection points for i = 1, 2,… n. In a triangular fuzzy number, yi is between 0 and 1. Also, n is the number of intersection points between two TFNs.

In the third step, the similarity value of two TFNs is obtained according to equation (8).

$$S = \frac{P(H \cap I)}{P(H)} \quad (8)$$

$P(H \cap I)$ and $P(H)$ is the intersection between two fuzzy numbers H and I and the area of the source fuzzy number H, respectively.

$$P(H \cap I) = \text{Area} \quad (9)$$

The area of triangular fuzzy number can compute by:

$$P(H) = \left|\frac{width \times height}{2}\right| \quad (10)$$

S is calculated in equation (8) for two one-dimensional fuzzy numbers. Since the features of artificial and benchmark datasets are multidimensional, to obtain S for two multidimensional fuzzy numbers, these features are considered independently. The degree of the similarity of the features in each dimension is obtained separately. The final similarity value is the product of the values obtained.

$$S_{total} = S_1 \times S_2 \times S_3 \times \ldots \times S_k \quad (11)$$

In equation (11), k is the number of features .For example, the iris set has 150 samples related to three types of iris flowers. Each iris sample has four features. So, the total similarity value is:

$$S_{total(Iris)} = S_1 \times S_2 \times S_3 \times S_4 \quad (12)$$

A similarity value is a number that alters from zero to one. 0 means dissimilar, and 1 means entirely similar.

So, a lonely particle demonstrates a candidate solution to a specified clustering problem. The fitness of every particle is obtained by the next formula.

$$J_e = \frac{\sum_{j=1}^{N_c}[\sum_{\forall \tilde{Z}_p \in C_{ij}} S_{total}(\tilde{Z}_p, \tilde{m}_{ij})/|C_{ij}|]}{N_C} \quad (13)$$

where in $\tilde{Z}_p \in F^{t\times n}(R)$.

In other words, the proposed algorithm solves equation (13). In the following, the proposed algorithm for clustering triangular fuzzy data is described.

Proposed algorithm. The particle swarm clustering algorithm for triangular fuzzy data with triangular fuzzy cluster centers:

Choose the number of clusters N_c, and set iteration = 1

Initialize the fuzzy cluster centroids of every particle accidentally

For iteration =1 to iteration $_{max}$ do

Begin

For each particle i do

Begin

For every data vector $\tilde{Z}_p$ do

Begin

Compute the $S_{total}(\tilde{Z}_p, \tilde{m}_{ij})$to all centers of the clusters $\tilde{m}_{ij}$

Allocate $\tilde{Z}_p$ to cluster $\tilde{C}_{ij}$ so that:

$$S_{total}(\tilde{Z}_p, \tilde{m}_{ij}) = \max\{S_{total}(\tilde{Z}_p, \tilde{m}_{ij})\} \quad \forall k = 1, \ldots, N_c \quad (14)$$

End

End

Compute the fitness function via equation (13)

Update the global best and local best locations using equations (15) and (16)

Update the centers of the clusters via formulas (17) and (18)

End

Equations (15) to (18) correspond to the PSO algorithm. Each particle j in the PSO has three attributes: a present location in the search region, P_i, a current velocity, V_i, and a personal best location in the search region, X_i. X_j is the location in the search region where particle j introduced the lowest error as specified by the target function g, supposing a minimalization work.

The global best (gb) location marked by $\tilde{X}$ demonstrates the location that produces the smallest error among all of the Xj. Formulas (15) and (16) describe how the personal and gb amounts are updated, respectively. It is supposed hereunder that the group includes of n particles, therefore j ∈ 1…n

$$X_j(time+1) = \begin{cases} X_j(time) & \text{if } g(X_j(time)) \le g(P_j(time+1)) \\ P_j(time+1) & \text{if } g(X_j(time)) \ge g(P_j(time+1)) \end{cases} \quad (15)$$

$$\tilde{X}(time) = \min\left\{g(X), g\left(\tilde{X}(time)\right)\right\} \quad X \in \{X_0(time), X_1(time), \ldots, X_n(time)\} \quad (16)$$

Within each iteration of the PSO, each particle is updated by formulas (17) and (18). Equations (17) and (18) are the velocity update formula and position update formula in PSO.

$$V_{j,k}(time+1) = wV_{j,k}(time) + c_1 r_{1,k}(time)[X_{j,k}(time) - P_{j,k}(time)] + c_2 r_{2,k}(time)[\tilde{X}_k(time) - P_{j,k}(time)] \quad (17)$$

$$P_j(time+1) = P_j(time) + V_j(time+1) \quad (18)$$

where w is the inertia weight, c_1 and c_2 are two positive fixed values, r_1 and r_2 are two random values in the span [0,1], and k=1, 2, …, N_d. N_d denotes the input dimensions or the number of features. Equation (17) includes three sections. The first section shows the current velocity of the

particle, which exhibits its current condition; The second section is the cognition phrase, which; The third section is called social phrase, which reflects the data subscription amongst the swarm. These three sections together specify the space searching ability.
Since real numbers, interval numbers, and TFNs are the particular types of triangular fuzzy numbers (TFNs), the proposed algorithm can also use for clustering of the real, interval, or trapezoidal data.

4. Simulation and Results

To exhibit the accuracy of the suggested procedure, experiments on different artificial and real data sets are discussed in this section. Also, the evaluation criterion of the suggested procedure will explain.

A. Evaluation criteria

In this paper, the accuracy criterion mentioned in [22] is used to measure the accuracy of the clustering outcomes (equation (19)). To calculate accuracy, each cluster is devoted to the class which is most repeated in the cluster. The accuracy of this allocation is evaluated by enumerating the number of samples allocated correctly and dividing them by N. N is the total number of samples.

$$\text{accuracy}\ (W, CC) = \frac{1}{N} \sum_j \max_i \left|\Omega_j \cap CC_i\right| \qquad (19)$$

In equation (19) the set of clusters is shown with $W = \{\Omega_1, \Omega_2, \ldots, \Omega_j\}$ and the set of classes is shown with $CC = \{cc_1, cc_2, \ldots, cc_i\}$.
Also, the Rand Index (RI) criterion is used to show the similarity between the two methods of labeling ((equation (20)).
If the clusters are created according to the classes, the rand index will be equal to 1.

$$\text{Rand Index}\ (W, CC) = \frac{A+B}{N(N-1)/2} \qquad (20)$$

In equation (20), A is the number of pairs that are together in both clusters and classes, and B is the number of pairs that are separated from each other both in clusters and in classes.

B. Artificial datasets

At first, we considered an artificial fuzzy dataset in R_2. This dataset is a straightforward artificial dataset. The purpose of creating this dataset was mere to represent the ability of the suggested procedure in fuzzy/uncertain data clustering. This dataset includes two clusters with ten dots. Each dot is a TFN with two dimensions. Fig. 4 exhibits the consequence acquired from the clustering of the first artificial dataset by utilizing the suggested clustering method. It is evident from Fig. 4 that the suggested clustering method can cluster this dataset correctly.

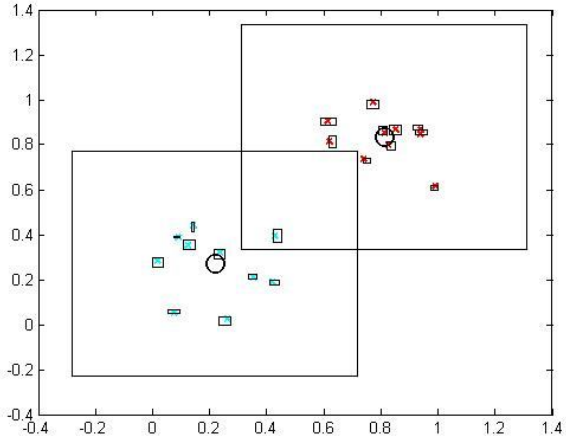

Fig. 4. The results of the suggested procedure for artificial dataset 1.

In fact, in clustering, the results are the centers of the clusters. The proposed method can find the fuzzy cluster centers. Each of the fuzzy cluster centers has two parameters. The central points of the clusters are shown with circles, and the margins of the centers are exhibited with rectangles. Our proposed method is performed the clustering operation like other standard uncertain clustering methods such as the Uncertain K-means and Uncertain K-medoids methods correctly. The proposed method achieved the best fuzzy centers of clusters. All

data is assigned to the correct corresponding cluster. As mentioned in the proposed method, whatever the centers of the fuzzy cluster have more points from the related cluster, clustering accuracy increases. As you see in Fig. 4, the result of clustering is such that the margins of the fuzzy clusters cover the entire samples of each cluster.

Two other artificial datasets with multi clusters are considered for appraising the proposed clustering method to solve the clustering problem. Fig. 5 and Fig. 6 exhibit the outcome acquired from the clustering of two multi-cluster datasets by the suggested clustering method. The second dataset has three clusters with fifty points. The third dataset has five clusters with hundred points. It can conclude that the suggested clustering method can apply to cluster multi-cluster fuzzy datasets. As you see in Fig. 4 and Fig 5, the clustering outcomes are such that the margins of the fuzzy clusters cover the entire samples of each cluster. All data is assigned to the correct corresponding cluster without error.

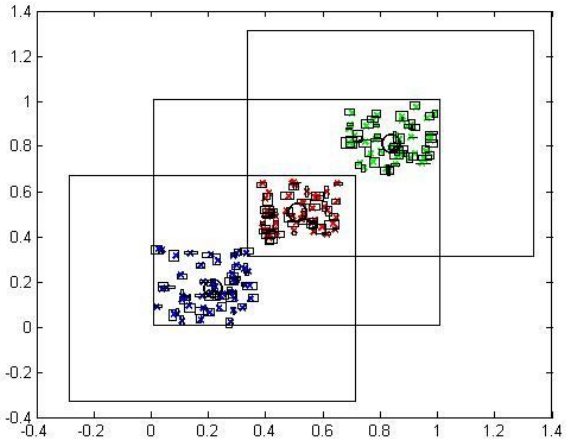

Fig. 5. The results of the suggested procedure for artificial dataset 2.

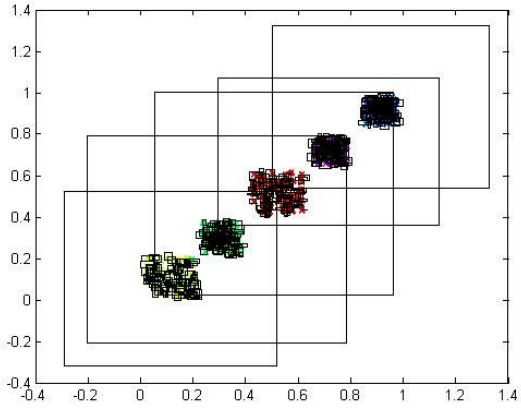

Fig. 6. The outcomes of the proposed method for artificial dataset 3.

Also, in Fig. 7, we consider the fourth artificial data set. This dataset has non-linearly separable clusters. This dataset consists of two clusters with five hundred dots. Fig. 7 shows the suggested clustering procedure can discover an optimum fuzzy cluster center with a minimum error like other standard uncertain clustering methods such as the UK-means and UK-medoids methods. The suggested clustering procedure is evaluated based on the criterion mentioned in the "A" section (equation (19)).

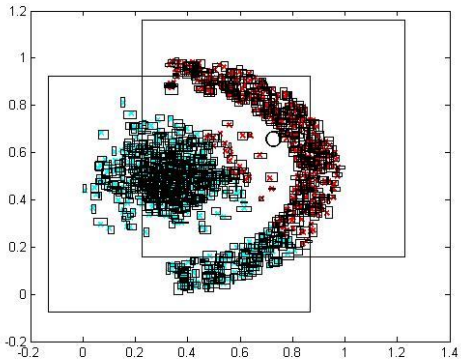

Fig. 7. The results of the proposed method for artificial dataset 4.

C. Triangular fuzzy benchmark datasets

To validate the proposed clustering method, seven real datasets exhibited by triangular fuzzy variables are considered. These triangular fuzzy datasets are iris, wine, vehicle, yeast, abalone, glass, ecoli. The specifications of these datasets are existent in Table 1. The accuracy of the suggested clustering procedure is compared with two standard uncertain data clustering methods, including uncertain-k means and uncertain-k medoids methods.

Tab. 1. Characteristics of benchmark datasets

Dataset	Features	Samples	Classes
Iris	4	150	3
Vehicle	18	846	4
Wine	13	178	3
Yeast	8	1484	10
Abalone	7	4124	17
Ecoli	7	327	5
Glass	9	214	6

Iris data set

The iris triangular fuzzy data set contains 150 measurements, including petal length, petal width, sepal length, and sepal width. These features are described by 4 triangular fuzzy variables. In this paper, only three features are considered. These features represent Versicolor, setosa, and Virginica of iris flowers. The experimental results of the iris dataset clustering are provided in Table 2. It can achieve from Table 2 that the suggested clustering procedure gives the best accuracy by 4% mis clustering error.

Tab. 2. Mis clustering percent of the different uncertain clustering methods on the iris dataset.

	Proposed clustering method	uncertain-k means	uncertain-k medoids
Mis clustering percent	4%	11.33%	10%

Wine dataset

The wine triangular fuzzy data set consists of 178 chemic to analyze of wines grown in the identic zone an Italy but derivative from various kinds described by 13 triangular fuzzy variables. The number of clusters is 3. In this paper, the four features are considered. Table 3 provides the experimental results derived from the clustering of the wine dataset. Table 3 confirms that the proposed clustering procedure is the most accurate clustering procedure for the wine dataset by 4.49% mis clustering error.

Tab. 3. Mis clustering percent of the different uncertain clustering methods on wine dataset.

	Proposed clustering method	uncertain-k means	uncertain-k medoids
Mis clustering percent	4.49%	5.06%	7.3%

Vehicle dataset

The vehicle triangular fuzzy data set consists of 846 vehicles described by 18 triangular fuzzy variables. 4 "Corgie" model vehicles consist of Chevrolet van, double-decker bus, Opel Manta 400, and Saab 9000 be used. In this paper, the four features are considered.

Ecoli dataset

The ecoli triangular fuzzy data set consists of 327 protein localization sites described by 7 triangular fuzzy numerical attributes. The number of clusters is 5. In this paper, three features are considered.

Abalone dataset

Abalone is a mollusk with a peculiar ear-shaped shell lined with a mother of pearl. Its age can be approximate by counting the number of rings in its shell with a microscope, but it is a time-consuming method. The abalone triangular fuzzy dataset contains 4124 physical measurements of abalones, large, edible sea snails. This dataset has 7 features consisting of Length, Height, Diameter, Shell weight, Shucked weight, Viscera weight, Whole weight. In this dataset, the number of clusters is 17.

Glass dataset

The triangular fuzzy glass dataset is achieved from USA Forensic Science Service and contains 214 pieces of glass. Each piece has a measured reflectivity index and chemical composition. In this dataset, 6 types of glass exist in terms of their oxide content.

Yeast dataset

The triangular fuzzy yeast dataset consists of 1484 data about a set of yeast cells. The task is to cluster the localization site of each cell amongst 10 conceivable alternatives. The number of features in the yeast dataset is 8.

Table 4 exhibits the results for iris, wine, vehicle, ecoli, abalone, glass, and yeast, respectively. Table 4 confirms that the accuracy of the suggested clustering method is quite comparable to uncertain-k means and uncertain-k medoids methods.

Also, Table 4 demonstrates that in all datasets except yeast, the proposed clustering method provides better results in accuracy when compared to UK-means and UK-medoids algorithms. Also, in Table 4, the average accuracy of the answers for ten times the execution of the algorithms is reported. The proximity of the mean and the best accuracy reported answer indicates the higher stability of the proposed method. Also, In Table 4, whatever the rand index (RI) is close to 1, the clustering is better.

Tab. 4. Comparing the proposed clustering method with uncertain-k means, and uncertain-k medoids methods on benchmark datasets

	Proposed clustering method			uncertain-k means			uncertain-k medoids		
	Accuracy	Mean	RI	Accuracy	Mean	RI	Accuracy	Mean	RI
Iris	96%	94.40%	0.9495	88.67%	84.27%	0.9273	90%	84.93%	0.9384
Wine	95.51%	95.04%	0.9488	94.94%	94.66%	0.9344	92.70%	83.26%	0.9412
Vehicle	41.96%	39.50%	0.3801	40.43%	37.71%	0.3792	40.78%	38.92%	0.3755
Ecoli	83.25%	80.15%	0.8312	78.90%	74.40%	0.8123	82.26%	78.37%	0.8266
Abalone	26.65%	23.25%	0.7872	22%	19.80%	0.7892	23.30%	20.21%	0.7723
Glass	56.74%	52.21%	0.5852	55.14%	52.90%	0.5801	54.15%	52.20%	0.5791
Yeast	46.02%	44.54%	0.5687	52.18%	49.35%	0.6117	52%	49.80%	0.6188

The clustering diagrams of the proposed method for some benchmark datasets are shown in Fig. 8.

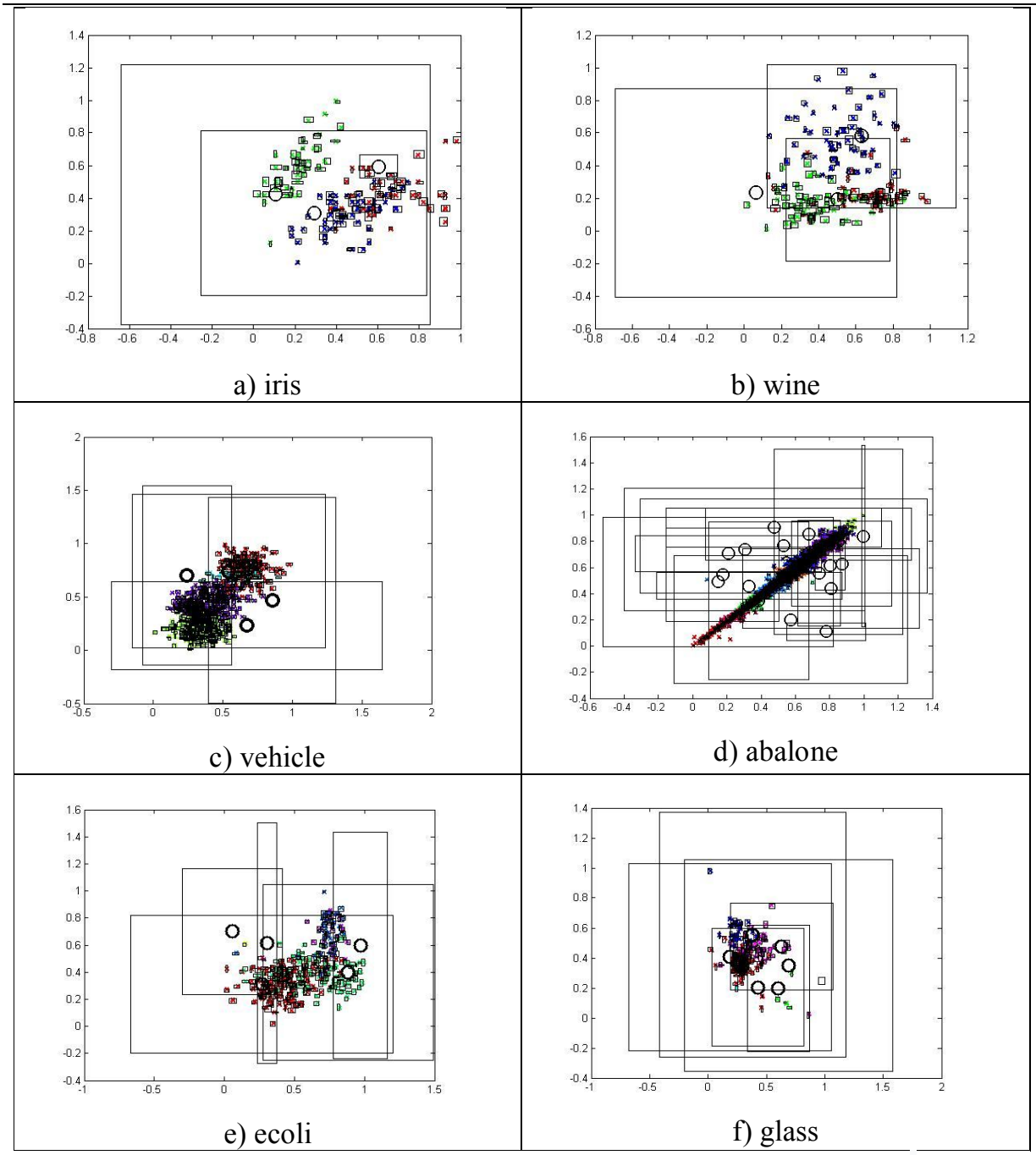

Fig. 8. The clustering diagrams for triangular fuzzy benchmark datasets.

These graphs are drawn for feature values in dimensions 1 and 2. Because the dimensions of the features of the benchmark dataset are more than 2, and it is not possible to show these graphs for all dimensions, it is not possible to analyze and compare the results of the suggested clustering procedure from the charts. Here, too, the centers of the clusters obtained by the proposed method have a central point, which is shown as a circle, and the margins of the centers of the cluster obtained, which are represented by large rectangles.

5. Conclusion

The goal of the present perusal was to suggest a particle swarm clustering method to cluster triangular fuzzy datasets. For this purpose, we defined a new cost function for the traditional particle swarm clustering method. The proposed cost function can find optimal fuzzy cluster centers for data whose features are expressed in

triangular fuzzy numbers. In the proposed method, Whatever the centers of the fuzzy cluster have more points from the corresponding cluster, clustering accuracy increases. To compare triangular fuzzy numbers, a similarity criterion based on the intersection region of the fuzzy numbers is used.

The accuracy of the proposed clustering method is evaluated with different artificial datasets. The empirical outcomes of the clustering of these datasets exhibit that our suggested procedure can apply to cluster fuzzy datasets. Also, seven triangular fuzzy benchmark datasets are selected to appraise our proposed clustering method to deal with real datasets. The results affirm that our proposed procedure is comparable to standard uncertain data clustering methods, including uncertain-k means and uncertain-k medoids methods. The experimental results on benchmark datasets (Table 4) demonstrate that in all datasets except Yeast, the proposed clustering method provides better results in accuracy when compared to UK-means and UK-medoids algorithms.

Besides, since any real or interval data is a particular kind of trapezoidal fuzzy data, the proposed clustering method can apply to cluster uncertain data sets of the type fuzzy, real, or interval data.

References

[1] J. Kennedy, and R. c. Eberhart, "Particle Swarm Optimization. In proceedings of IEEE internal conference on neural networks", Perth, Australia, Vol. 4, (1995), pp. 1942-1948.

[2] R. C. Eberhart, and J. Kennedy, "A new optimizer using particle swarm theory", In proceeding of the sixth international symposium on micro machine and human science, Nagoya, Japan, (1995), pp. 39-43.

[3] I. Behravan, et al., "Finding roles of players in football using automatic particle swarm optimization-clustering algorithm", Big data, Vol. 7, No. 1, (2019), pp. 35-56.

[4] Z. Liu, et al., "An improved unsupervised image segmentation method based on multi-objective particle swarm optimization clustering algorithm", CMC, Vol. 58, No. 2, (2019), pp. 451-461.

[5] M. Soltani, et al., "A New Model for Blood Supply Chain Network Design in Disasters Based on Hub Location Approach Considering Intercity Transportation", International Journal of Industrial Engineering & Production Research, Vol. 32, No. 2, (2021), pp. 1-28.

[6] S. Jafarian-Namin, et al., "Economic-Statistical Design of an Integrated Triple-Component Model Under Various Autocorrelated Processes", International Journal of Industrial Engineering & Production Research, Vol. 32, No. 4, (2021), pp. 1-18.

[7] S-H. Zahiri, and S-A. Seyedin, "Swarm intelligence-based classifiers", Journal of the franklin institute, Vol. 344, (2007), pp. 362-376.

[8] S-H. Zahiri, and S-A. Seyedin, "Using multi-objective particle swarm optimization for designing novel classifiers", in swarm intelligence for multi-objective problems in data mining, Vol. 242, (2009), pp. 65-92.

[9] E. Bakhshizadeh, et al., "Customer Clustering Based on Factors of Customer Lifetime Value with Data Mining Technique (Case Study: Software Industry)", International Journal of Industrial Engineering & Production Research, Vol. 33, No. 1, (2022), pp. 1-16.

[10] van der Merwe, D.W., and A.P. Engelbrecht, "Data clustering using particle swarm optimization in evolutionary computation", CEC '03. The 2003 congress on. (2003).

[11] H. Shahraki, and S-H. Zahiri, "Particle swarm classifier for fuzzy data sets", The artificial intelligence and signal processing (AISP), (2015).

[12] F. Gullo, "An information-theoretic approach to hierarchical clustering of uncertain data", Information sciences, Vol. 402, (2017), pp. 199-215.

[13] Y. Mao, et al., "Uncertain interval data EFCM-ID clustering algorithm based on

machine learning", Journal of robotics and mechatronics, Vol. 31, No. 2, (2019), pp. 339-347.

[14] A. Makhmutova, I. Anikin, "Online clustering on uncertain data stream", Journal of physics: conference series 1189, 012025, (2019).

[15] GS. Nijaguna, K. Thippeswamy, "Multiple kernel fuzzy clustering for uncertain data classification", International journal of computer engineering and technology (IJCET), Vol. 10, No. 01, (2019), pp. 253-261.

[16] Xu. Weixiang, Li. Jiaojiao, "An improved algorithm for clustering uncertain traffic data streams based on hadoop platform", International journal of modern physics B, Vol. 33, No. 19, (2019), pp. 1950203-1-19.

[17] K. K. Sharma, A. Seal, "Modeling uncertain data using monte carlo integration method for clustering", Expert systems with applications, Vol. 6, (2019), pp. S0957-4174,.

[18] J. Zhou, et al., "Uncertain data clustering in distributed peer-to-peer networks", IEEE transaction on neural networks and learning systems, Vol. 29, No. 6, (2018), pp. 2392- 2406.

[19] H. Shahraki, and S-H. Zahiri, "Fuzzy decision function estimation using fuzzified particle swarm optimization", International journal of machine learning and cybernetics, Springer, (2016).

[20] Renata M.C.R. de Souza, Francisco de A.T. de Carvalho, "Clustering of interval data based on fa–block distances", Pattern recognition letters, Vol. 25, (2004), pp. 353-365.

[21] F.A.T. de Carvalho, et al., "Adaptive hausdorff distances and dynamic clustering of symbolic interval data", Pattern recognition letters, Vol. 27, (2006), pp. 167-179.

[22] Xianchao Zhang, Han Liu, Xiaotong Zhang, "Novel density-based and hierarchical density-based clustering algorithms for uncertain data", Neural networks, Vol. 93, (2017), pp. 240-255.

[23] R. E. Bellman, L. A. Zadeh, "Decision making in a fuzzy environment", Manag. Sci, Vol. 17, (1970), pp. 141-164.

[24] S. Sukpisit, et al., "Polygon intersection-based algorithm for fuzzy set compatibility calculations", International journal of machine learning and computing, Vol. 6, No. 1, (2016), pp. 32-35.

[25] X. Wang, D. Ruan, and E. E. Kerre, "Mathematics of fuzziness basic issues", Springer-verlag berlin heidelberg, (2009).

Follow This Article at The Following Site:

zahiri S H, ghanbari N, Shahraki H. Clustering of fuzzy data sets based on particle swarm optimization with fuzzy cluster centers. IJIEPR. 2022; 33 (2) :1-12
URL: http://ijiepr.iust.ac.ir/article-1-1326-en.html