Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

SRTTU

JECEI

Research paper

Clustering of Triangular Fuzzy Data Based on Heuristic Methods

N. Ghanbari [1]***, S. H. Zahiri*** [1,*]***, H. Shahraki*** [2]

[1]*Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.*

[2]*Department of Computer Engineering, Faculty of Industry and Mining, University of Sistan and Baluchestan, Khash, Iran.*

Article Info

Article History:
Received 18 March 2023
Reviewed 23 April 2023
Revised 16 May 2023
Accepted 14 June 2023

Keywords:
Heuristic clustering
Particle swarm optimization
Uncertain data
Fuzzy dataset
Ranking function

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: In this paper, a new version of the particle swarm optimization (PSO) algorithm using a linear ranking function is proposed for clustering uncertain data. In the proposed Uncertain Particle Swarm Clustering method, called UPSC method, triangular fuzzy numbers (TFNs) are used to represent uncertain data. Triangular fuzzy numbers are a good type of fuzzy numbers and have many applications in the real world.

Methods: In the UPSC method input data are fuzzy numbers. Therefore, to upgrade the standard version of PSO, calculating the distance between the fuzzy numbers is necessary. For this purpose, a linear ranking function is applied in the fitness function of the PSO algorithm to describe the distance between fuzzy vectors.

Results: The performance of the UPSC is tested on six artificial and nine benchmark datasets. The features of these datasets are represented by TFNs.

Conclusion: The experimental results on fuzzy artificial datasets show that the proposed clustering method (UPSC) can cluster fuzzy datasets like or superior to other standard uncertain data clustering methods such as Uncertain K-Means Clustering (UK-means) and Uncertain K-Medoids Clustering (UK-medoids) algorithms. Also, the experimental results on fuzzy benchmark datasets demonstrate that in all datasets except Libras, the UPSC method provides better results in accuracy when compared to other methods. For example, in iris data, the clustering accuracy has increased by 2.67% compared to the UK-means method. In the case of wine data, the accuracy increased with the UPSC method is 1.69%. As another example, it can be said that the increase in accuracy for abalone data was 4%. Comparing the results with the rand index (RI) also shows the superiority of the proposed clustering method.

Introduction

Data clustering plays an essential role in machine learning, data mining, statistical analysis, image segmentation, and pattern recognition [1]-[3]. Clustering groups data objects based on the information found in that qualifies the objects and their communications. The aim of clustering is that the objects within a group be similar or related to one another and different from the objects in the other groups.

The greater likeness within a group and the greater disagreement between the groups, the better the clustering [4]. Almost all standard heuristic clustering methods are developed for crisp datasets. At the same time, in real-world applications such as sensor measurements, biomedical, and microarray, information cannot often be expressed by crisp numbers. For simple example in the sentence "Ryan's intelligence is good." The goodness of Ryan's intelligence is a qualitative sense and cannot be measured by any quantity. These data are known as uncertain data.

Uncertain data can be used in many cases. There are some quantities that must be mentioned by interval data; For instance, the range of temperature during a day. Also, uncertainty may have resulted from the lack of knowledge, implicit randomness in data generation, data staling, and inability to perform adequate physical measurements or vagueness [5]. It is usually related to missing or incomplete information or the probability of occurrence of given information [6]. Therefore, it is a critical issue to consider uncertain data in clustering algorithms. The clustering of uncertain data has been considered in various papers [7]-[13]. In some studies, interval data is used to represent uncertain data [7]. A famous example of interval data is the temperature data set [7]. The temperature interval data set gives the minimum and the maximum monthly temperatures of 37 cities in degrees centigrade. This data set has 12 features and 4 cluster. In [8], an uncertain clustering algorithm using cloud model is proposed. In [9], one method of multiple kernel fuzzy clustering for uncertain data classification has been presented. A suitable cluster validity index for uncertain data clustering is presented in [10]. In [11], a new clustering algorithm is presented using the k-medoids for uncertain objects. In [12], uncertain data clustering in distributed peer-to-peer networks has been done. In [13], an upgrade of particle swarm optimization (PSO) algorithm for the fuzzy environment has been reported where particles have been defined as fuzzy numbers, and their motion has been reformulated by fuzzy equations.

In [14], interval data clustering based on the adaptive dynamic cluster method has been described. Carvalho et al. used adaptive Hausdorff distances and dynamic clustering to cluster the interval data [15]. In [16], a novel clustering method based on novel density and hierarchical density has been proposed for interval data. Likewise, using interval data in various methods and applications is described in [7]. Using fuzzy numbers is another common way to express uncertain data. Tayyebi proposed a fuzzy clustering method (FCM), which clusters trapezoidal fuzzy numbers [17]. In this method, a linear ranking function is applied to define a distance between trapezoidal fuzzy data.

On the other hand, the powerfulness of optimization algorithms has caused those the heuristic clustering methods such as the genetic algorithm-based clustering method [18], [19] particle swarm optimization algorithm-based clustering method [20], [21], learning automata-based clustering method [22], and clustering method based on gravitational search algorithm [23], [24] be created. In these clustering methods, the optimization algorithms are used to find the optimal centroids for clusters.

This type of clustering method (heuristic clustering method) is used in many types of research [18]-[24], and their results show that these clustering methods have good potential to apply in different applications. Despite all the advances of these clustering methods [18]-[24], clustering of uncertain data with these methods is one of the topics that have not yet been investigated. So, in this paper, we want to promote one of these clustering methods to cluster uncertain data. Thus, the particle swarm optimization algorithm-based clustering method, which is one of the best clustering methods of this family, is chosed. The particle swarm optimization (PSO) algorithm is a relatively novel heuristic algorithm introduced by Kennedy and Eberhart [25]. PSO algorithm efforts to find the optimal solution throughout the simulation of some concepts that obtained from bird flocking, fish schooling, and other social folks. Each particle can efficiently attain his objective using the information that is owned by itself and the information that is assigned between the folk. This means that the PSO algorithm is an optimization procedure that utilizes the laws of social behavior [26]. Data clustering using the PSO algorithm was first suggested by Engelbrecht and Merwe in [27]. The idea of this clustering algorithm was allocating all cluster centroids to each particle and update the particle following to the fitness function value computed for the particle. As mentioned, particle swarm optimization has used only for certain data until now. In this paper, an enhancement of the PSO algorithm is proposed to appropriate for clustering uncertain data (triangular fuzzy numbers).

In this paper, for the first time, the new version of the particle swarm optimization algorithm is proposed for clustering uncertain data. For presenting uncertain data, triangular fuzzy numbers are used. On the other hand, in the present paper, particle swarm optimization algorithm is promoted to proceed with clustering triangular fuzzy numbers. The proposed method is called the Uncertain Particle Swarm Clustering (UPSC) method. In the UPSC method, particles are triangular fuzzy numbers. Therefore, to upgrade a version of standard PSO, obtaining the distance between the fuzzy numbers is necessary. For this purpose, a linear ranking function is applied to describe the distance between fuzzy vectors.

Ranking function used in the proposed method (UPSC) is introduced for triangular fuzzy numbers; because triangular fuzzy numbers are a good type of fuzzy numbers and have many applications in the real world. Much of the data in the real world is expressed with specific precision, which can be well represented by a triangular fuzzy number. The proposed method can also be used for clustering datasets of the type of interval data, real data or trapezoidal fuzzy number.

The contributions of this paper are as follows:

1- The method proposed in this paper can cluster uncertain or fuzzy data. In other words, uncertain data can be given as input to the proposed algorithm and be clustered.
2- in order to capable cluster uncertain data by the proposed algorithm, it is necessary to make changes in the conventional PSO algorithm. An important change is the use of a linear ranking function in the fitness function of PSO algorithm. With this, the distance between fuzzy/uncertain numbers can be calculated and evaluated. The changes applied to the conventional PSO algorithm provide a new version of the conventional PSO algorithm that can cluster fuzzy/uncertain data.
3- The new proposed cost function (using linear ranking function) of the proposed method can find the optimal centers of the clusters for data whose features are expressed in triangular fuzzy numbers (TFNs).
4- The most important advantage of the proposed method is compatibility with real-life applications. For example, the proposed method is its applicability in missing value data because one way to express missing value data is to estimate it with triangular fuzzy numbers by using the Fuzzy Nearest Neighborhood Mean (FNNM) method [17].
5- The proposed method can also be implemented for crisp data because crisp data is a particular type of fuzzy data and can be represented as triangular fuzzy data easily. It is necessary to mention uncertain data is uncertain due to noise. The fact that we defined a fuzzy membership function for each input data indicates that it is noisy.
6- One of the strengths of the proposed method is its general and modular form. Most of the computational work of the proposed method, such as fitness function and fuzzification of features, are general, and other heuristic methods with minimum changes can be utilized instead of PSO.

The rest of this paper are constituted as follows: Initially some preliminary notations and concepts of fuzzy numbers and fuzzy theory is presented. Then the proposed particle swarm optimization algorithm for clustering fuzzy data and the experimental clustering results are reported. Finally, conclusion of the paper is reported.

Preliminary Concepts

This section is divided into two subsections. Subsection “A” describes several concepts of the fuzzy set theory. Due to the importance of ranking function in the method presented in the present paper, Subsection “B” will investigate this issue, and in it, a linear ranking function for comparing fuzzy numbers are introduced. While introducing ranking function, the necessary definitions and lemmas is provided.

A. A Brief Introduction to Fuzzy Set Theory

In this section, we review the main concepts of the fuzzy set theory, initialize by Bellman and Zadeh in [28], and is applied in this paper.

Let UNI be the universal set. A mapping $\widetilde{\mathrm{P}}: \mathrm{UNI} \rightarrow [0,1]$ is a fuzzy set. The value $\widetilde{P}(x)$ of $\widetilde{\mathrm{P}}$ at $x \in UNI$ stands for the grade of membership of x in $\widetilde{\mathrm{P}}$. A fuzzy set $\widetilde{\mathrm{P}}$ is normal if there exists $x_0 \in UNI$ such that $P(x_0) = 1$. An α-cut of fuzzy number $\widetilde{\mathrm{P}}$, $\alpha \in [0,1]$, is a crisp set as:

$$\alpha = \{\mathrm{x} \in \mathrm{UNI}: \widetilde{\mathrm{P}}\,(\mathrm{x}) \geq \alpha\} \tag{1}$$

If a fuzzy set $\widetilde{\mathrm{P}}$ satisfies that $\widetilde{\mathrm{P}}_{\alpha}$ is a closed interval for every $\alpha \in [0,1]$, then $\widetilde{\mathrm{P}}$ is called a fuzzy number. A particular type of fuzzy numbers is the triangular fuzzy number (TFN) to be determined as:

$$\widetilde{P}_{\alpha} = \begin{cases} 0 & \text{for} \quad x \leq P_1 \\ \frac{x-P_1}{P_2-P_1} & \text{for} \quad P_1 < x \leq P_2 \\ \frac{P_3-x}{P_3-P_2} & \text{for} \quad P_2 < x \leq P_3 \\ 0 & \text{for} \quad P_3 < x \end{cases} \tag{2}$$

For reduction, the TFN $\widetilde{\mathrm{P}}$ has been marked by $\widetilde{\mathrm{P}} = (\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3)$ (Fig. 1). Next, arithmetic operation on triangular fuzzy numbers is described. Suppose that $\tilde{\mathrm{p}} = (\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3)$ and $\tilde{\mathrm{q}} = (\mathrm{q}_1, \mathrm{q}_2, \mathrm{q}_3)$ be two triangular fuzzy numbers and c a real number. Scalar product and scalar addition operators are described as follows:

$$\mathrm{c}.\tilde{\mathrm{p}} = (c\mathrm{p}_1, c\mathrm{p}_2, c\mathrm{p}_3) \qquad \text{if } \mathrm{c} \geq 0, \mathrm{c} \in \mathrm{R} \tag{3}$$

$$\mathrm{c}.\tilde{\mathrm{p}} = (\mathrm{cp}_3, \mathrm{cp}_2, \mathrm{cp}_1) \qquad \text{if } \mathrm{c} \leq 0, \mathrm{c} \in \mathrm{R} \tag{4}$$

$$\tilde{\mathrm{p}} + \tilde{\mathrm{q}} = (\mathrm{p}_1 + \mathrm{q}_1, \mathrm{p}_2 + \mathrm{q}_2, \mathrm{p}_3 + \mathrm{q}_3) \tag{5}$$

$$\tilde{\mathrm{p}} - \tilde{\mathrm{q}} = (\mathrm{p}_1 - \mathrm{q}_3, \mathrm{p}_2 - \mathrm{q}_2, \mathrm{p}_3 - \mathrm{q}_1) \tag{6}$$

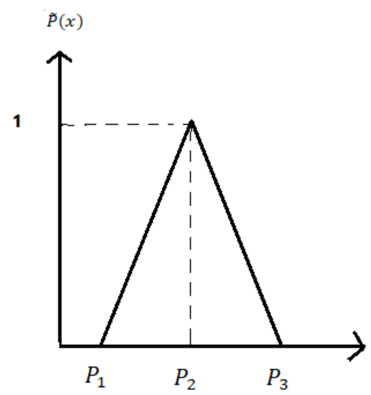

Fig. 1: Membership function of TFN $\widetilde{\mathrm{P}} = (\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3)$.

B. Ranking Function

There are several procedures for comparing fuzzy numbers, which can be viewed in Tanaka and Ichihashi [29], Lai and Hwang [30], Fang and Hu [31], and Shoacheng [32].

One of the most suitable of these procedures has relied on the notion of comparison of fuzzy numbers using ranking functions [33]-[41]. In this section, a ranking function have been investigated to determine a distance between fuzzy vectors.

An effective approach for arranging the point of F(R) is to determine a ranking function R: F(R)→R that maps each triangular fuzzy number into the actual line, where a natural order exists.

The ranking function is applied to specify a distance between triangular fuzzy vectors in the section 3. In this section, only linear ranking function are considered, i.e., a ranking function R such that

$$R(\tilde{p} + c\tilde{q}) = R(\tilde{p}) + cR(\tilde{q}) \tag{7}$$

for any $\tilde{p}$, $\tilde{q}$∈F(R) and any c ∈R, it is clear that R $(\tilde{0})$ = 0, where $\tilde{0}$ = (0,0,0). The ranking function is defined using Lemma.

Lemma 2.2.1 For constant nonnegative numbers α, β∈R, the function R: F(R)→R is determined as:

$$R(\tilde{p}) = \alpha p_1 + 2\beta p_2 + \alpha p_3 \tag{8}$$

where $\tilde{p} = (p_1, p_2, p_3)$∈F(R) is a linear ranking function. The proof was given in [17]. For example, if α=β=1/4, then $R(\tilde{p}) = p_1 + 2p_2 + p_3/4$, that has been suggested by Yager [42].

In the proposed method (UPSC), we set α=β=1/4 for the ranking function.

Then, for triangular fuzzy numbers $\tilde{p}$ and $\tilde{q}$, we have

$$\tilde{p} \geq \tilde{q} \quad \text{if and only if } \; p_2 + \frac{1}{2}(p_3 + p_1) \geq q_2 + \frac{1}{2}(q_3 + q_1) \tag{9}$$

Ranking function used in the proposed method (UPSC) are introduced for triangular fuzzy numbers; because triangular fuzzy numbers are a standard approximation for fuzzy numbers and have many applications in the real world. Much of the data in the real world is expressed with specific precision, which can be well represented by a triangular fuzzy number. For example, persons 'weight can be defined with a sure accuracy, which can be well represented by triangular fuzzy numbers. The proposed method (UPSC) can also be applied for clustering datasets of the type of interval data, real data or trapezoidal fuzzy number.

A ranking function to specify a distance between trapezoidal fuzzy vectors/numbers is explained in [17]. Obviously, any real number and any interval number can be rewritten as a trapezoidal fuzzy number.

Uncertain Particle Swarm Clustering (UPSC) Method

In this section, at first, an overview of PSO and then PSO clustering method for certain data and finally proposed method (UPSC) for uncertain data will be explained respectively.

C. An Overview of PSO

The particle swarm optimization (PSO) algorithm is a relatively novel heuristic algorithm introduced by Kennedy and Eberhart [25], [26]. In PSO method, the particles fly through the problem/search space by following the optimal particles. Each particle recollects the best position that it has searched (P_{best}) and also best position among all the particles in the Population/group (G_{best}). The position of each particle updates according to the P_{best} and G_{best} in the search space. A simple and standard implementation of PSO algorithm is as follows [43]:

1. Create a random population of particles in the search space
2. Calculate the fitness of each particle according to the fitness function defined in the problem
3. Update the velocity of each particle based on the velocity (10)

   $$vel_i^d(t+1) = Wvel_i^d(t) + rand \times c_1 \times [p_{best_i}^d - x_i^d(t)] + rand \times c_2 \times [g_{best_i}^d - x_i^d(t)] \tag{10}$$

   where, c_1 and c_2 are two positive constants, rand is random number uniformly distributed within the span [0,1], w is the inertia weight. In the proposed method (UPSC), we set c_1=c_2=2 and w=0.7. position of the ith particle shows with $X_i = (x_i^1, x_i^2, \dots, x_i^n)$ and velocity of the ith particle shows with $VEL_i = (vel_i^1, vel_i^2, \dots, vel_i^n)$. Also, the best previous position of the ith particle represent with $P_{besti} = (p_{best_i}^1, p_{best_i}^2, \dots, p_{best_i}^n)$ and the best previous position among all the particles in the group shows with $G_{besti} = (g_{best_i}^1, g_{best_i}^2, \dots, g_{best_i}^n)$.
4. Update the position of each particle using the position (11)

   $$x_i^d(t+1) = vel_i^d(t+1) + x_i^d(t) \tag{11}$$

5. Repeat loops 2 to 4 until the stop condition is met. The condition for stopping is usually to achieve the desired accuracy or maximum number of repetitions.

D. PSO Clustering Method for Certain Data

Among all the literature attempts to improve the particle swarm optimization algorithm to real/certain data clustering, [44] represents to be the one that is nearest to the main ideas of the PSO because each particle perceives an entire candidate solution to the problem. In other words, each particle represents the center of the cluster. Based on this method, a particle X_i is created as follows:

$$X_i = (m_{i1}, m_{i2}, \dots, m_{ic}) \quad \text{for } \; L=1,2,\dots,c \tag{12}$$

For clusters with n dimension m_{iL} is as follows.

$$m_{iL} = (m_{iL}^1, m_{iL}^2, \dots, m_{iL}^n) \quad \text{for } \; p=1, 2,\dots,n \tag{13}$$

where c is the number of clusters to be organized, and

m_{iL} corresponds to the Lth centroid of the ith particle, the centroid of the cluster C_{iL}. So, a lonely particle demonstrates a candidate solution to a given clustering problem.

The input data set D contains N samples with dimensions n.

$$D = (d_1, d_2, \dots, d_N) \quad \text{for } k = 1,2,3,\dots,N \tag{14}$$

where d_k defines as follows:

$$d_k = (d_k^1, d_k^2, \dots, d_k^n) \quad \text{for} \quad p = 1,2,3,\dots,n \tag{15}$$

The fitness of each particle is evaluated by the following fitness function.

$$\text{Fitness function} = \frac{\sum_{L=1}^{c}[\sum_{\forall d_k \in C_{iL}} ED(d_k, m_{iL})/|C_{iL}|]}{c} \tag{16}$$

where $d_k = (d_k^1, d_k^2, \dots, d_k^n)$ defines the kth data vector, $|C_{iL}|$ is the number of data vectors belonging to the cluster C_{iL} and ED is the Euclidian distance between d_k and m_{iL}.

E. The Proposed Uncertain Particle Swarm Clustering (UPSC) Method

In this section, a novel PSO algorithm is offered for clustering triangular fuzzy data. This algorithm expands the standard PSO algorithm introduced in [44] because the standard PSO algorithm just clusters certain/real data. Suppose we have a dataset such as:

$$\tilde{D} = [\tilde{d}_1, \tilde{d}_2, \dots, \tilde{d}_N] \quad \text{for } k = 1,2,3,\dots,N \tag{17}$$

where N is the number of data vector to be clustered.

In (17) we have:

$$\tilde{d}_k = (\tilde{d}_k^1, \tilde{d}_k^2 \dots, \tilde{d}_k^n)^T \quad \text{for} \quad p = 1,2,3,\dots,n \tag{18}$$

And

$$\tilde{d}_k^p = ((d_k^p)_1, (d_k^p)_2, (d_k^p)_3) \quad \text{for } k = 1,2,3,\dots,N \quad \text{and} \quad p = 1,2,3,\dots,n \tag{19}$$

Thus, $\tilde{D}$ is a part of $F^{n\times N}(R)$. The goal is to divide $\tilde{d}_k$'s into c clusters.

Allow $\tilde{X}_i$ be a triangular fuzzy matrix of the prototype. $\tilde{X}_i$ is defined as:

$$\tilde{X}_i = (\tilde{m}_{i1}, \tilde{m}_{i2}, \dots, \tilde{m}_{ic}) \quad \text{for} \quad L=1,2,\dots,c \tag{20}$$

$\tilde{X}_i$ is a part of $F^{n\times c}(R)$. In (20) we have:

$$\tilde{m}_{iL} = (\tilde{m}_{iL}^1, \tilde{m}_{iL}^2, \dots, \tilde{m}_{iL}^n) \quad \text{for} \quad p=1, 2,\dots,n \tag{21}$$

$$\tilde{m}_{iL}^p = ((m_{iL}^p)_1, (m_{iL}^p)_2, (m_{iL}^p)_3) \quad \text{for } L = 1,2,3,\dots,c \quad \text{and} \quad p = 1,2,3,\dots,n \tag{22}$$

Because the data are fuzzy, it is assumed that the prototype is also triangular fuzzy numbers. But in the implementation of the proposed method (UPSC), it was decided to obtain cluster centers as real numbers, so in (22) $(m_{iL}^p)_1 = (m_{iL}^p)_2 = (m_{iL}^p)_3$ is considered. In the following, the linear ranking function is applied to describe a distance between the fuzzy vector $\tilde{d}_k$'s and vector $\tilde{m}_{iL}$'s to be necessary to extend the standard PSO algorithm.

Definition 3.3.1 allow R be a linear ranking function. The mapping $d_R: F^n(R) \times F^n(R) \to R$ with

$$d_R(\tilde{f}, \tilde{g}) = \sqrt{\sum_{p=1}^{n} R^2(\tilde{f}_p - \tilde{g}_p)} = \sqrt{\sum_{p=1}^{n} (R(\tilde{f}_p) - R(\tilde{g}_p))^2} \tag{23}$$

is called a fuzzy distance with relation to R where $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_n), \tilde{g} = (\tilde{g}_1, \tilde{g}_2, \dots, g_n) \in F^n(R)$.

It is clear that the description of d_R is a direct expansion of the formal Euclidean distance. Based on Lemma 2.2.1, the ranking mapping R to be described by (24) is linear. So, for ranking function R, the fuzzy distance can rewrite as follows:

$$R(\tilde{p}) = \alpha p_1 + 2\beta p_2 + \alpha p_3 \tag{24}$$

$$d_R^2([\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_n], [\tilde{g}_1, \tilde{g}_2, \dots, \tilde{g}_n]) = \sum_{p=1}^{n}[\alpha((f_p)_1 + (f_p)_3 - (g_p)_1 - (g_p)_3) + 2\beta((f_p)_2 + -(g_p)_2)]^2 \tag{25}$$

where $\tilde{f}_p = ((f_p)_1, (f_p)_2, (f_p)_3)$ and $\tilde{g}_p = ((g_p)_1, (g_p)_2, (g_p)_3)$ for p=1, 2, …, n. The (25) specifies the value $d_R^2(\tilde{f}, \tilde{g})$ clearly; but the compact form of definition 3.3.1 is applied.

The proposed UPS clustering algorithm solves

$$\text{Fitness function(UPSC)} = \frac{\sum_{L=1}^{c}[\sum_{\forall \tilde{d}_k \in C_{iL}} d_R^2(\tilde{d}_k, \tilde{m}_{iL})/|C_{iL}|]}{c} \tag{26}$$

$$\text{Fitness function(UPSC)} = \frac{\sum_{L=1}^{c}[\sum_{\forall \tilde{d}_k \in C_{iL}} \sum_{p=1}^{n}[\alpha((d_k^p)_1 + (d_k^p)_3 - (m_{iL}^p)_1 - (m_{iL}^p)_3) + 2\beta((d_k^p)_2 - (m_{iL}^p)_2)]^2/|C_{iL}|]}{c} \tag{27}$$

where in $\tilde{d}_k \in F^{n\times N}(R)$.

Now, the proposed algorithm (UPSC) for clustering triangular fuzzy data is described.

It is clear that any real number can be written as triangular fuzzy number, so Algorithm 1 is converted to the standard PSO algorithm for real data.

Algorithm 1: Uncertain particle swarm clustering (UPSC) algorithm

Choose the number of clusters c and set iteration = 1
Initialize the cluster centroids of each particle randomly
For iteration =1 to iteration $_{max}$ do
Begin
For each particle i do
Begin
For each data vector $\tilde{d}_k$ do
Begin
Calculate the $d_R^2(\tilde{d}_k, \tilde{m}_{iL})$ to all cluster centroids $\tilde{m}_{iL}$ using fuzzy distance (28) obtained using ranking functions.

$$d_R^2(\tilde{d}_k, \tilde{m}_{iL}) = \sum_{p=1}^{n}[\alpha\left(\left(d_k^p\right)_1 + \left(d_k^p\right)_3 - \left(m_{iL}^p\right)_1 - \left(m_{iL}^p\right)_3\right) + 2\beta(\left(d_k^p\right)_2 - \left(m_{iL}^p\right)_2)]^2 \quad (28)$$

Assign $\tilde{d}_k$ to cluster C_{iL} such that:

$$d_R^2(\tilde{d}_k, \tilde{m}_{iL}) = \min\ \{d_R^2(\tilde{d}_k, \tilde{m}_{iV})\} \qquad \text{for } V = 1,2,\dots,c \quad (29)$$

End
End
Calculate the fitness function using (26)
Update the global best and local best positions
Update the cluster centroids using (10) and (11)
End

It is a preference for the proposed algorithm (UPSC) that if data is a specific type (such as real numbers, interval numbers, or TFNs), then prototypes are the same as type; prototypes are linear compounds of data.

Experimental Results

In this section, datasets, evaluation criteria, and numerical results are explained.

F. Datasets

To evaluate the proposed method (UPSC), several datasets, including six artificial datasets and nine UCI datasets, are used. These datasets have been converted to triangular fuzzy numbers and then used.

Nine Benchmark Datasets

These datasets are iris, wine, vehicle, yeast, image, abalone, libras, glass, ecoli. These datasets are existent at the UCI machine learning repository [45]. The specifications of these datasets are existent in Table 1.

Table 1: Characteristics of fuzzy benchmark datasets

Dataset	#Objects	#Classes	#Attributes
Iris	150	4	3
Wine	178	13	3
Vehicle	846	18	4
Yeast	1484	8	10
Image	2310	19	7
Abalone	4124	7	17
Libras	360	90	15
Ecoli	327	7	5
Glass	214	9	6

Six Artificial Datasets

For three artificial datasets using MATLAB functions, a uniformly distributed pseudorandom integer has been created. The first dataset has 20 points distributed among two clusters Fig. 2. The second dataset has 500 points distributed among five clusters Fig. 3, and the third dataset has 150 points distributed among three clusters Fig. 4.

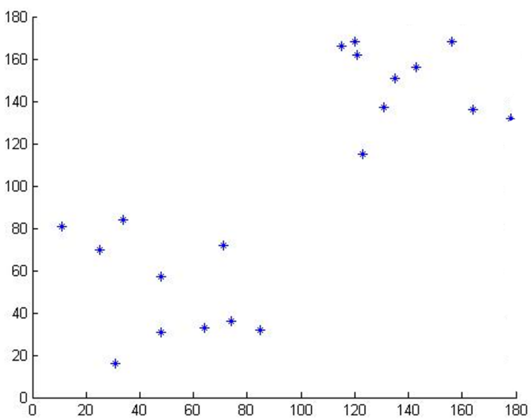

Fig. 2: Seed artificial dataset 1, which created using the uniform distribution of MATLAB software.

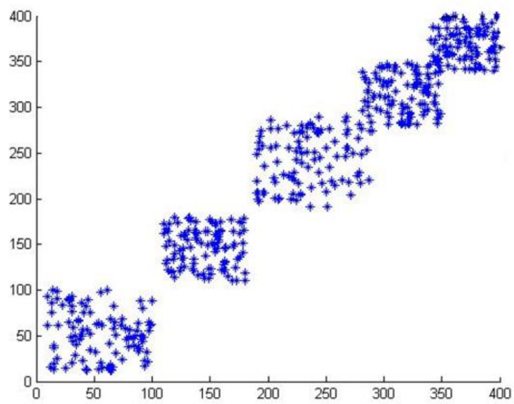

Fig. 3: Seed artificial dataset 2, which created using the uniform distribution of MATLAB software.

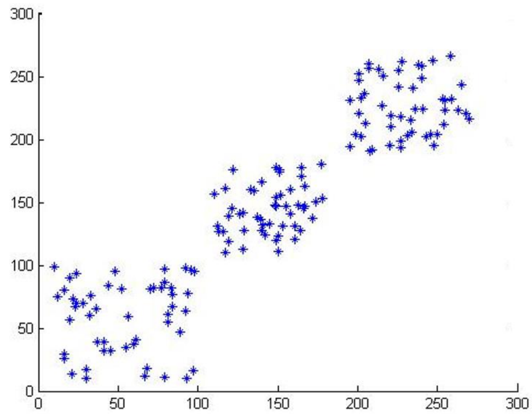

Fig. 4: Seed artificial dataset 3, which created using the uniform distribution of MATLAB software.

For the other two artificial datasets, the same data point formations introduced in [14] are considered. At first, two standard quantitative datasets in R^2 be regarded. Each dataset includes 350 points distributed among three clusters of unequal sizes and shapes: one cluster with size 50 and a spherical shape and two clusters with sizes 150 and ellipsis shapes. Data points of each cluster in datasets 4 and 5 were drawn following a bi-variate normal distribution of autonomous parts with th following mean vector and covariance matrix:

$$\mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \quad \text{and} \quad \Sigma_{11} = \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \tag{30}$$

Artificial dataset 4 exhibits well-separated clusters Fig. 5.

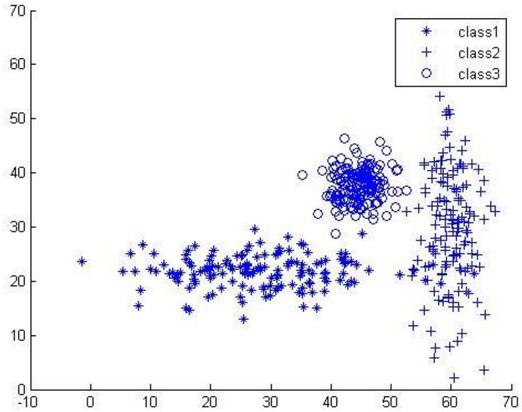

Fig. 5: Seed artificial dataset 4, which exhibit well-separated classes.

Data points of each cluster in dataset 4 were drawn following the subsequent parameters:

$$\text{Class 1}: \mu_1 = 28, \mu_2 = 22, \sigma_1^2 = 100, \sigma_2^2 = 9 \tag{31}$$

$$\text{Class 2}: \mu_1 = 60, \mu_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 144 \tag{32}$$

$$\text{Class 3}: \mu_1 = 45, \mu_2 = 38, \sigma_1^2 = 9, \sigma_2^2 = 9 \tag{33}$$

Artificial dataset 5 exhibits overlapping clusters Fig. 6.

Data points of each cluster in dataset 5 were drawn following the subsequent parameters:

$$\text{Class 1}: \mu_1 = 45, \mu_2 = 22, \sigma_1^2 = 100, \sigma_2^2 = 9 \tag{34}$$

$$\text{Class 2}: \mu_1 = 60, \mu_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 144 \tag{35}$$

$$\text{Class 3}: \mu_1 = 52, \mu_2 = 38, \sigma_1^2 = 9, \sigma_2^2 = 9 \tag{36}$$

The sixth dataset has non-linearly separable clusters. This dataset is very suitable for density-based clustering algorithms like DBSCAN. This dataset has 500 points and two clusters Fig. 7.

To create artificial data in this paper, an attempt has been made to use different types of data with different numbers, different distributions, different dispersions, different complexities, and different amounts of interference.

For example, dataset 1 is a straightforward artificial dataset. The purpose of creating it was merely to demonstrate the ability of the proposed method for solving the desired problem. In datasets 2 and 3, the number of samples and clusters has increased compared to dataset 1. In dataset 5, more interference was considered for data. In dataset 6, data distribution was considered more difficult than in other datasets.

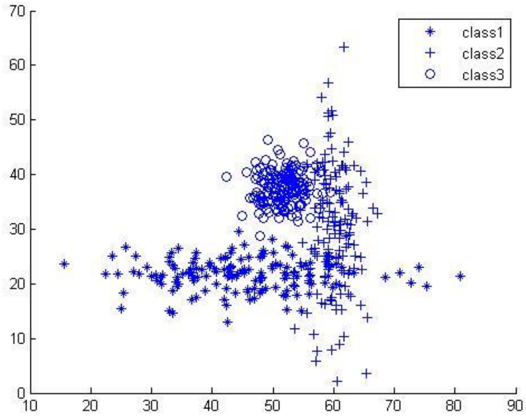

Fig. 6: Seed artificial dataset 5, which has overlapping classes.

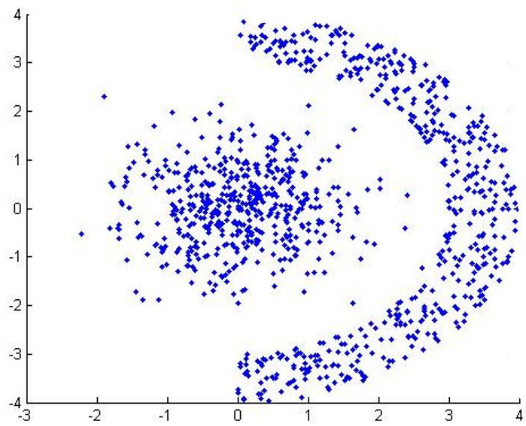

Fig. 7: Seed artificial dataset 6.

G. Converting Real Numbers to Triangular Fuzzy Number

A simple method is used to convert all datasets to triangular fuzzy datasets. Triangular fuzzy number in LR format is show in Fig. 8. In LR format we have:

$$P^* = P_3 - P_2, P_* = P_2 - P_1, P_0 = P_2 \tag{37}$$

If P_{0_i} be a crisp or certain data, firstly, the maximum and minimum values of each crisp dataset ($\min(P_0)$ and $\max(P_0)$) are obtained using equations (38, 39) (N is the number of data sets). The P_* and P^* parameters are set to a random value using (40) and uniform distribution (41). The "rand" function of MATLAB software is used in (41) to generate random real numbers with a uniform distribution.

$$\min(P_0) = \min_{i=1,2,\ldots,N} P_{0_i} \tag{38}$$

$$\max(P_0) = \max_{i=1,2,\ldots,N} P_{0i} \quad (39)$$

$$\text{TFN range} = \text{alpha} \times (\max(P_0) - \min(P_0)) \quad (40)$$

$$P_* = P^* = \text{rand} \times \text{TFN range} \quad (41)$$

Now crisp data is converted to triangular fuzzy data (TFN). Triangular fuzzy data ($\widetilde{P}$) exhibits in the following format (42).

$$\widetilde{P} = (P_*, P_0, P^*)_{LR} = (P_2 - P_1, P_2, P_3 - P_2) \quad (42)$$

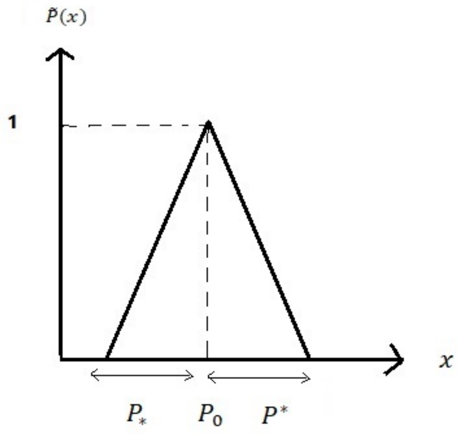

Fig. 8: Triangular fuzzy number in LR format ($\widetilde{P} = (P_0, P_*, P^*)_{LR}$).

Before using data to test the proposed method (UPSC), normalization preprocessing is performed on it.

H. Evaluation criteria

In this paper, the accuracy criterion mentioned in [16], which is one of the most popularly used criteria, is utilized to measure the accuracy of the clustering results. To calculate accuracy, each cluster is assigned to the class which is most repeated in the cluster. Then the accuracy of this assignment is evaluated by counting the number of correctly assigned objects and dividing them on all objects n. formally

$$\text{accuracy}(M, CL) = \frac{1}{n} \sum_c \max_k |m_c \cap cl_k| \quad (43)$$

where $M = \{m_1, m_2, \ldots, m_c\}$ is the set of clusters and CL=$\{cl_1, cl_2, \ldots, cl_k\}$ is the set of classes.

The Rand Index (RI) criterion is another criterion computed to exhibit the resemblance between the two procedures of labeling (44). The rand index for error-free clustering is equal to 1.

$$\text{Rand Index}(M, CL) = \frac{a+b}{n(n-1)/2} \quad (44)$$

In (44), a is the number of pairs that are together in both classes and clusters, and b is the number of pairs that are separated from each other both in classes and clusters.

I. Numerical Results

In this section, the proposed method and other methods are evaluated based on the mentioned criteria.

In all simulations, c_1=c_2=2 and w=0.7 for UPSC, and as discussed in section 2, α=β=1/4 for ranking function is set. The experiment is repeated ten times to decrease the variation from trial to trial. Table 2 shows the accuracy results on six artificial datasets for the proposed method (UPSC), UK-means, and UK-medoids. The Accuracy or purity is one of the most popular criteria for evaluating clustering methods. Two sample results for each dataset in Table 2 are reported.

These results are related to two different alpha parameters in (40). The first row is related to alpha = 0.01, and the second row is associated with alpha = 0.03.

The higher the alpha, the higher the margin of fuzzy triangular numbers (P_* and P^* in Fig. 8). The average, best and worst answers for ten times implementation of these algorithms are reported in Table 2. The mean is the average of the answers received from the ten times the implementation of these algorithms. Also, best and worst show the best and worst answers received from the ten times of implementing these algorithms, respectively. To evaluate the proposed method (UPSC), the Uk-Means and Uk-Medoids algorithms also implemented. In fuzzy artificial dataset 1, which is straightforward, all three methods performed the clustering without error. In artificial dataset 2, when the alpha coefficient increases to 0.03 (data uncertainty increases), firstly, our proposed method can find the answer without error in this case. Secondly, the mean, the best and the worst answer are slightly different together and can be said that the stability of our proposed method is higher. In the artificial dataset 3, as can be seen from Table 2, for alpha=0.01, answers of all three methods are without error. Still, by increasing the alpha coefficient to 0.03, the superiority of UPSC is quite apparent, because in this case, mean, best and worst answer is the same.

This means complete stability of the proposed method in the case of dataset 3, While in the UK-Means and UK-medoids methods, the best and worst answers are different. In artificial dataset 4, although the UK-Means and UK-medoids methods have better stability, the mean and the best answer obtained using UPSC are better. For the artificial dataset 5, which has high interference, and the artificial dataset 6, which has a more difficult distribution, according to Table 2, the UPSC method has the better answers and good stability.

After artificial datasets, to have a more realistic valuation, the UPSC method is applied to the UCI datasets as mentioned above. The optimal cluster prototypes are available for these datasets. Table 3 shows the accuracy results on nine fuzzy benchmark datasets.

The results are achieved by repeating the experiments 10 times.

Table 2: Clustering accuracy results on fuzzy artificial datasets using UPSC, UK-Means, and UK-medoids methods

Dataset		UPSC			UK-Means			UK-medoids		
		Mean	best	worst	Mean	best	worst	Mean	best	worst
Artificial dataset 1	alpha=0.01	1	1	1	1	1	1	1	1	1
	alpha=0.03	1	1	1	1	1	1	1	1	1
Artificial dataset 2	alpha=0.01	0.9712	0.9920	0.7840	0.8780	0.9960	0.7960	0.8688	0.9960	0.7680
	alpha=0.03	0.9976	1	0.9960	0.8748	1	0.7720	0.9228	0.9800	0.7960
Artificial dataset 3	alpha=0.01	1	1	1	1	1	1	1	1	1
	alpha=0.03	1	1	1	0.9547	0.9867	0.6667	0.9667	1	0.6667
Artificial dataset 4	alpha=0.01	0.8386	0.8786	0.8214	0.8086	0.8143	0.8000	0.8057	0.8143	0.8000
	alpha=0.03	0.8993	0.9071	0.8286	0.8621	0.8786	0.8571	0.8514	0.8571	0.8286
Artificial dataset 5	alpha=0.01	0.7843	0.7857	0.7714	0.7300	0.7500	0.7000	0.7357	0.7571	0.7143
	alpha=0.03	0.7814	0.7857	0.7786	0.7093	0.7143	0.7071	0.7071	0.7643	0.6929
Artificial dataset 6	alpha=0.01	0.8480	0.8640	0.8410	0.8377	0.8450	0.8290	0.7758	0.8590	0.5420
	alpha=0.03	0.8560	0.8680	0.8430	0.8442	0.8540	0.8400	0.8372	0.8600	0.8170

As can be seen in Table 3, UPSC has higher accuracy than the other two methods for iris data. Besides, the best and worst answers of UPSC for iris data are less different than the other two methods. In the case of wine data, UPSC method is more accurate and stable than the UK-medoids method and more accurate than the UK-Means Method. In the vehicle dataset, when the data uncertainty increases with increasing alpha, the ability of our proposed method increases, and we reach a higher accuracy than the other two methods. In the yeast dataset, the best answer with a slight difference is related to UPSC method, but the mean of our method is better with more differences than the other two methods. For the image dataset, UPSC method has the highest accuracy. In the abalone dataset, the UPSC method has a significantly better answer, and with increasing the alpha coefficient, the accuracy of the UPSC method increases. In contrast, for the UK-Means and the UK-medoids methods, the accuracy decreases with increasing alpha, and this demonstrates the ability of the UPSC method for clustering fuzzy data.

In the glass and ecoli datasets, according to Table 3, the proposed method has a better answer, and like the other the dataset, the clustering accuracy increases with increasing alpha. Only in the libras dataset the accuracy of UPSC method less than that of the other two methods. Although the UK-Means Method has the highest accuracy for the libras dataset, in this method and the UK-medoids method, accuracy decreased with increasing alpha. While with increasing alpha coefficient, UPSC's ability in clustering increased.

In the continuation of this section, Table 4 and Table 5 are presented to facilitate the comparison of UPSC method with other methods.

The best answer for the proposed method was for alpha=3. Therefore, Table 4 and Table 5 are reported for alpha=3. Table 4 is a summary of Table 2. The achieved results for UPSC in challenging with the other methods show the capability and robustness of UPSC method. As Table 4 shows, in all artificial datasets, UPSC method gives the best answer. Table 5 compares the results of UPSC method on the benchmark dataset with the UK-means, the UK-medoids, and uncertain ACO (Ant Colony Optimization) methods. In Table 5, the best accuracy, the mean accuracy and RI index are reported. The propinquity of the best accuracy and the mean accuracy in Table 5 illustrate the higher stability of UPSC method.

According to Table 5 in eight datasets, UPSC method has higher accuracy. Although in libras dataset the UK-means method has the highest accuracy, but with increasing the alpha coefficient, the ability of UPSC method for clustering libras data increase. In contrast, for the UK-means method, with increasing alpha, the clustering accuracy decrease. In Table 5, whatever the rand index (RI) is close to 1, The clustering error is less.

According to the calculations reported in Table 5, the proposed clustering method has a better Rand index in datasets. Due to the random and search-based nature of the proposed method, the volume of calculations and the execution time of the algorithm increase. Of course, since the calculations are done offline, this issue becomes less important. In other words, this time is related to clustering the existing data and finding the optimal centers of the clusters. After the clustering is done, the clustering of unknown and new samples is done in very little time.

Table 3: Clustering accuracy results on fuzzy benchmark datasets using UPSC, UK-Means, and UK-medoids methods

Dataset		UPSC			UK-Means			UK-medoids		
		Mean	best	worst	Mean	best	worst	Mean	best	worst
Iris	alpha=0.01	0.8973	0.9200	0.8867	0.8647	0.8867	0.6667	0.8300	0.9067	0.6667
	alpha=0.03	0.8960	0.9267	0.8800	0.8867	0.8867	0.8867	0.8493	0.9000	0.6667
Wine	alpha=0.01	0.9534	0.9663	0.9438	0.9449	0.9494	0.9326	0.8326	0.9270	0.6011
	alpha=0.03	0.9579	0.9663	0.9494	0.9478	0.9551	0.9326	0.8629	0.9663	0.6180
Vehicle	alpha=0.01	0.3827	0.4031	0.3700	0.3959	0.4078	0.3652	0.3858	0.4102	0.3641
	alpha=0.03	0.3987	0.4232	0.3712	0.3858	0.4043	0.3641	0.3892	0.4078	0.3629
Yeast	alpha=0.01	0.4739	0.5168	0.4387	0.4678	0.5004	0.4135	0.4725	0.5068	0.4422
	alpha=0.03	0.5054	0.5263	0.4650	0.4935	0.5218	0.4401	0.4980	0.5200	0.4161
Image	alpha=0.01	0.6541	0.7039	0.6117	0.6214	0.6814	0.5745	0.6403	0.6926	0.5667
	alpha=0.03	0.6518	0.7061	0.6160	0.6206	0.6771	0.5745	0.6292	0.6948	0.5502
Abalone	alpha=0.01	0.2602	0.2679	0.2512	0.2164	0.2204	0.1998	0.2225	0.2345	0.2072
	alpha=0.03	0.2611	0.2701	0.2500	0.2142	0.2200	0.202	0.2220	0.2330	0.2167
Libras	alpha=0.01	0.4158	0.4417	0.3917	0.4636	0.5083	0.4306	0.4414	0.4833	0.4028
	alpha=0.03	0.4156	0.4667	0.3611	0.4758	0.5028	0.4500	0.4342	0.4861	0.3861
Glass	alpha=0.01	0.5308	0.5561	0.5000	0.5108	0.5501	0.4800	0.5290	0.5561	0.5093
	alpha=0.03	0.5407	0.5794	0.4953	0.5192	0.5514	0.5000	0.5286	0.5415	0.5040
Ecoli	alpha=0.01	0.7908	0.8318	0.7523	0.7801	0.8043	0.7523	0.7839	0.8218	0.7454
	alpha=0.03	0.7920	0.8349	0.7584	0.7755	0.7890	0.7645	0.7801	0.8226	0.6453

Table 4: Comparing The best accuracy of UPSC method with the best accuracy of other methods for fuzzy artificial datasets

Dataset	UPSC (proposed method)	UK-Means	UK-medoids
Artificial dataset 1	**1**	**1**	**1**
Artificial dataset 2	**1**	**1**	0.98
Artificial dataset 3	**1**	0.9867	**1**
Artificial dataset 4	**0.9071**	0.8786	0.8571
Artificial dataset 5	**0.7857**	0.7143	0.7643
Artificial dataset 6	**0.8680**	0.8540	0.8600

Fig. 9 shows the implementation time of the proposed clustering method (UPSC) for all data. All tests and simulations in this paper (for proposed method and other conventional method) by the computer with Core (TM) 2 Duo 2.20GHz central processing unit and 4GB memory in MATLAB software environment performed. As shown in Fig. 9, the larger the dataset volume according to Table 1, the more time is required to implement the UPSC method.

For example, according to Table 1 abalone dataset contains 4124 objects.

This dataset has 7 features, and the goal is to separate the abalone dataset into 17 clusters. Therefore, more time is needed to run the UPSC method than (for example) the iris dataset; Because the iris dataset has 150 4-dimensional samples that should be divided into 3 clusters.

Artificial datasets are also small in volume, so little time is required to implement the UPSC method.

Due to the large number of datasets used in this paper, only clustering output diagrams for artificial datasets 1, 2 and, 3 are shown.

Fig. 10, Fig. 11, and Fig. 12 show the clustering diagrams for these datasets. The diagrams in Fig. 10, Fig. 11, and Fig. 12 associated with the UK-means and the UK-medoids methods are for alpha=0.03.

Table 5: Comparing the UPSC method with other methods for fuzzy benchmark datasets

	UPSC (proposed method)			UK-means			UK-medoids			Uncertain ACO		
	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI	Best accuracy	Mean accuracy	RI
Iris	**0.9267**	**0.8960**	**0.9400**	0.8867	0.8867	0.9273	0.9000	0.8493	0.9384	0.9211	0.8945	0.9388
Wine	**0.9663**	**0.9579**	**0.9478**	0.9494	0.9449	0.9344	0.9270	0.8326	0.9412	0.9479	0.9389	0.9320
Vehicle	**0.4232**	**0.3987**	0.3793	0.4043	0.3858	**0.3794**	0.4078	0.3892	0.3755	0.3992	0.3850	0.3782
yeast	**0.5263**	**0.5054**	**0.6548**	0.5218	0.4935	0.6117	0.5200	0.4980	0.6188	0.5243	0.5030	0.6517
image	**0.7061**	**0.6518**	**0.6784**	0.6771	0.6206	0.6617	0.6948	0.6292	0.6702	0.6954	0.6615	0.6720
Abalone	**0.2701**	**0.2611**	**0.7896**	0.2200	0.2142	0.7869	0.2330	0.2220	0.7725	0.2345	0.2301	0.7851
Libras	0.4667	0.4156	0.5873	**0.5028**	**0.4758**	**0.5990**	0.4861	0.4342	0.5897	0.4634	0.4020	0.5855
Glass	**0.5794**	**0.5407**	**0.5879**	0.5514	0.5192	0.5791	0.5415	0.5286	0.5788	0.5702	0.5326	0.5833
Ecoli	**0.8349**	**0.7920**	**0.8343**	0.7890	**0.7755**	0.8143	0.8226	0.7801	0.8275	0.8240	0.7899	0.8225

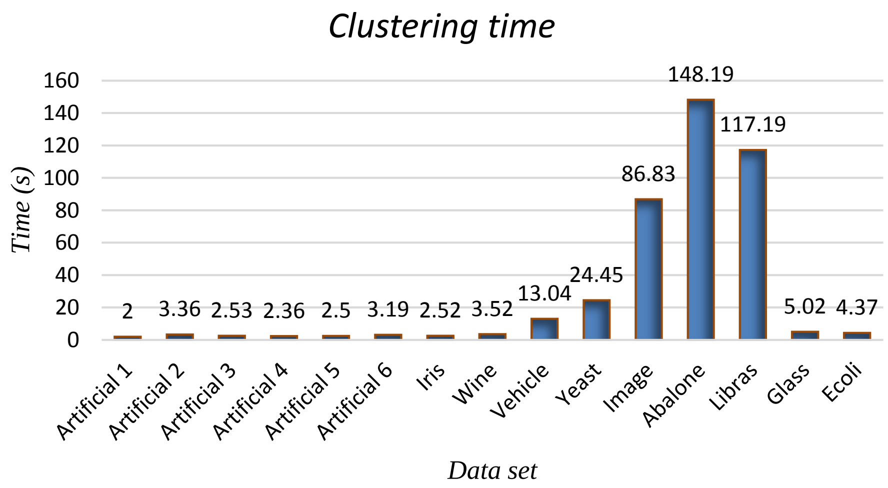

Fig. 9: Clustering time for all fuzzy datasets in UPSC method.

To see the results of the proposed method (UPSC) for alpha=0.01, we have also shown it in Fig. 10, Fig. 11, and Fig. 12. These clustering diagrams show that UPSC method can find centroids of clusters related to these datasets without error or minimum error.

Dataset 1 is a straightforward artificial dataset. The purpose of creating this dataset was mere to demonstrate the ability of UPSC method in fuzzy/uncertain data clustering. In fact, in clustering, the results are the centers of the clusters. The two obtained cluster centers for the artificial dataset 1 are shown in Fig. 10 with two squares, blue and red. As you can see in Fig. 10, UPSC method with two selected alpha parameters 0.01, and 0.03 (top row in Fig. 10), has been able to perform the clustering operation like the UK-means and UK-medoids methods (bottom row in Fig. 10) correctly and achieved the best centers of clusters.

All blue data is assigned to the cluster with the blue center and, all red data is transferred to the cluster with a red center.

Fig. 11 shows the clustering diagram related to artificial dataset 2. Fig. 11 shows the superiority of UPSC method clearly. In the above row of Fig. 11, which is related to UPSC method with alpha 0.01 and 0.03, all the color data are assigned to their respective clusters. The bottom row of Fig. 11, is corresponding to the UK-means and UK-medoids methods that have errors in this implementation and have not been able to obtain the centers of the clusters correctly. The five correct clusters created in the artificial dataset 2 are shown in Fig. 3. The last diagram that shown as an example is related to the artificial dataset 4. By comparing Fig. 12 with Fig. 5, the superiority of UPSC method is identified. The top row of Fig. 12 (UPSC method) has performed clustering with fewer errors than the bottom row methods (UK-means and UK-medoids).

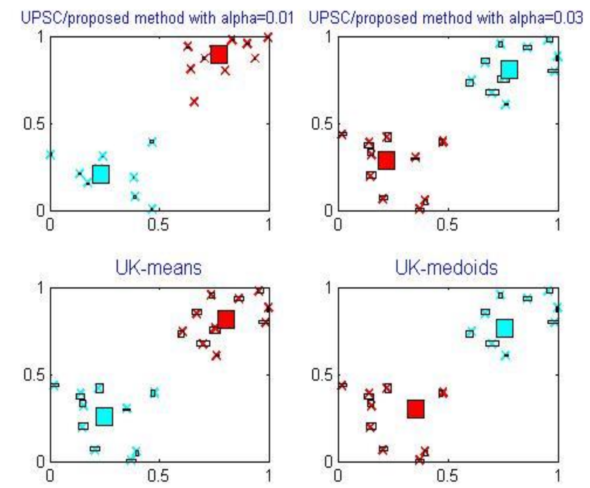

Fig. 10: Clustering diagrams for fuzzy artificial dataset 1.

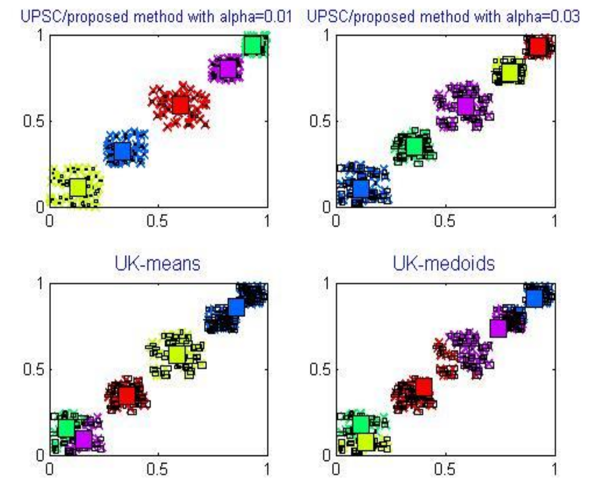

Fig. 11: Clustering diagrams for fuzzy artificial dataset 2.

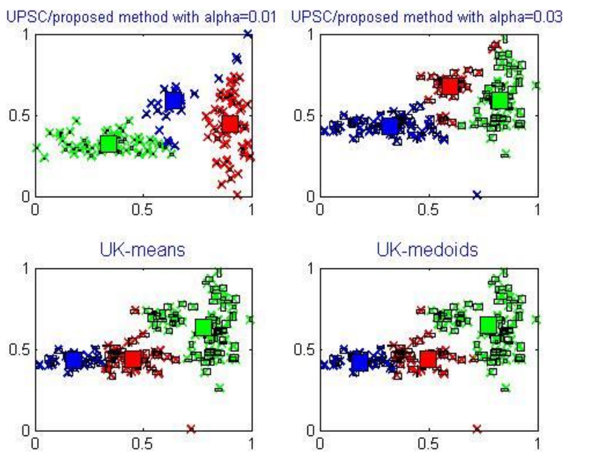

Fig. 12: Clustering diagrams for fuzzy artificial dataset 4.

Conclusion

In this paper, for the first time, the particle swarm optimization algorithm is upgraded for clustering of uncertain/fuzzy datasets. For this purpose, a new cost function is defined for the traditional particle swarm clustering method. The proposed cost function can find the optimal centers of the clusters for data whose features are expressed in triangular fuzzy numbers (TFNs). On the other hand, an uncertain particle swarm clustering (UPSC) method is proposed, which can cluster fuzzy datasets. The proposed method (UPSC) can be also used for clustering uncertain datasets with interval data, trapezoidal fuzzy data, or real data. The achieved results show that the proposed method (UPSC) is superior to the challenging and standard methods such as UK-means, UK-medoids, and Uncertain ACO algorithms. So, the UPSC is an efficient solution for the problem of clustering uncertain data. For example, in iris data, the clustering accuracy has increased by 2.67% compared to the UK-means method. In the case of wine data, the accuracy increase with the proposed method is 1.69%. As another example, it can be said that the increase in accuracy for abalone data was 4%. Comparing the results with the Rand index also shows the superiority of the proposed method. The most important limitations of heuristic algorithms, including the PSO algorithm, are capturing local optimum and matter of being offline. Of course, it should be mentioned that the most of the clustering algorithms are offline and can meet the needs of the problem. In this paper, the PSO algorithm is upgraded for clustering uncertain/fuzzy data. One of the works that can be done in the future is using other versions of the PSO algorithm that are more complex and may give better results. Using soft computing methods (e.g., fuzzified PSO / fuzzy and PSO) is another works that can be done in the future. Also, the proposed method is general and modular. Therefore, other heuristic methods such as inclined planes system optimization (IPO) or gravitational search algorithm (GSA) with minimum changes can be used instead of PSO to solve uncertain/fuzzy data clustering.

Author Contributions

Dr. Zahiri and Dr. Shahraki were the supervisor and adviser of the current research paper. They sketched the research framework and the roadmap. Also, they analyzed the results. N. Ghanbari searched in authentic journals to gather all relevant papers. Also, she collected the data and wrote the manuscript. Dr. Zahiri, Dr. Shahraki, and N. Ghanbari interpreted the results.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

PSO	Particle Swarm Optimization
UPSC	Uncertain Particle Swarm Clustering
TFNs	Triangular Fuzzy Numbers
UK- means	Uncertain K-Means
UK-Medoids	Uncertain K-Medoids
FCM	Fuzzy Clustering Method
IPO	Inclined Planes System Optimization
GSA	Gravitational Search Algorithm
RI	Rand Index
ACO	Ant Colony Optimization

References

[1] A. Dutt, M. A. Ismail, T. Herawan, "A systematic review on educational data mining," IEEE Access, 5: 15991-16005, 2017.

[2] Z. Wang "Determining the clustering centers by slope difference distribution," IEEE Access, 5: 10995-11002, 2017.

[3] T. T. Zhang, B. Yuan, "Density-based multiscale analysis for clustering in strong noise settings with varying densities," IEEE Access, 6: 25861-25873, 2018.

[4] P. N. Tan, M. Steinbach, V. Kumar, "Introduction to datamining," Addison-Wesley, 2005.

[5] H. Shahraki, S. H. Zahiri, "Classification of trapezoidal fuzzy data based on heuristic classifiers," Kasmera, 43(1): 128-144, 2015.

[6] F. Gullo, "An information-theoretic approach to hierarchical clustering of uncertain data," Inf. Sci., 402: 199-215, 2017.

[7] Y. Mao, Y. Liu, M.A. Khan, J. Wang, D. Mao, J. Hu, "Uncertain interval data EFCM-ID clustering algorithm based on machine learning," J. Rob. Mechatron., 31(2): 339-347, 2019.

[8] L. Yue, L. Zitu, L. Shuang, G. Yike, L. Qun, W. Guoyin, "Cloud-Cluster: An uncertainty clustering algorithm based on cloud model," Knowledge-Based Syst., 263, 2023.

[9] G. S. Nijaguna, K. Thippeswamy, "Multiple kernel fuzzy clustering for uncertain data classification," Int. J. Comput. Eng. Technol. (IJCET), 10(01): 253-261, 2019.

[10] C. Ko, J. Baek, B. Tavakkol, Y. S. Jeong, "Cluster Validity Index for Uncertain Data Based on a Probabilistic Distance Measure in Feature Space," Sensors, 23(7): 3708, 2023.

[11] B. Tavakkol, Y. Son, "Fuzzy kernel K-medoids clustering algorithm for uncertain data objects," Pattern Anal. Appl., 24(3): 1287-1302, 2021.

[12] J. Zhou, L. Chen, C. L. Philip Chen, Y. Wang, H. X. Li, "Uncertain data clustering in distributed peer-to-peer networks," IEEE Trans. Neural Networks Learn.Syst., 29(6): 2392-2406, 2018.

[13] H. Shahraki, S. H. Zahiri, "Fuzzy decision function estimation using fuzzified particle swarm optimization," Int. J. Mach. Learn. Cybern., 8: 1827-1838, 2017.

[14] R. M.C.R. de Souza, F. de A.T. de Carvalho, "Clustering of interval data based on city–block distances," Pattern Recognit. Lett., 25: 353–365, 2004.

[15] F. de A.T. de Carvalho, R. M.C.R. de Souza, M. Chavent, Y. Lechevallier, "Adaptive hausdorff distances and dynamic clustering of symbolic interval data," Pattern Recognit. Lett., 27: 167–179, 2006.

[16] X. Zhang, H. Liu, X. Zhang, "Novel density-based and hierarchical density-based clustering algorithms for uncertain data," Neural Networks, 93: 240–255, 2017.

[17] J. Tayyebi, E. Hosseinzadeh, "A fuzzy c-means algorithm for clustering fuzzy data and its application in clustering incomplete data," J. AI Data min., 8(4): 515-523, 2021.

[18] R. Adrian, S. Sulistyo, I.W. Mustika, S. Alam, "ABNC: Adaptive border node clustering using genes fusion based on genetic algorithm to support the stability of cluster in VANET," Int. J. Intell. Eng. Syst., 13(1): 354-363, 2020.

[19] T. P. Q. Nguyen, R. J. Kuo, "Partition-and-merge based fuzzy genetic clustering algorithm for categorical data," Appl. Soft Comput. J., 75: 254–264, 2019.

[20] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Finding roles of players in football using automatic particle swarm optimization-clustering algorithm," Big Data, 7(1): 35-56, 2019.

[21] Z. Liu, B. Xiang, Y. Song, H. Lu, Q. Liu, "An improved unsupervised image segmentation method based on multi-objective particle swarm optimization clustering algorithm," CMC-Comput. Mater. Continua, 58(2): 451-461, 2019.

[22] B. Anari, J. Akbari torkestani, A. M. Rahmani, "A learning automata-based clustering algorithm using ant swarm intelligence," Expert Syst., 35(6): e12310, 2018.

[23] M. S. Tomar, P. K. Shukla, "Energy efficient gravitational search algorithm and fuzzy based clustering with hop count-based routing for wireless sensor network," Multimedia Tools Appl., 78: 27849–27870, 2019.

[24] H. Mittal, M. Saraswat, "An automatic nuclei segmentation method using intelligent gravitational search algorithm based super pixel clustering," Swarm Evol. Comput., 18(9): S2210-6502, 2018.

[25] J. Kennedy, R. C. Eberhart, "Particle swarm optimization," in Proc. IEEE Internal Conference on Neural Networks, 4: 1942-1948, 1995.

[26] W. Xiong, "Initial clustering based on the swarm intelligence algorithm for computing a data density parameter," Comput. Intell. Neurosci.: 1-8, 2022.

[27] D. W. Van der Merwe, A. P. Engelbrecht, "Data clustering using particle swarm optimization in evolutionary computation," in Proc. The 2003 Congress on Evolutionary Computation, 2003. CEC '03, 2003.

[28] R. E. Bellman, L. A. Zadeh, "Decision making in a fuzzy environment," Manag. Sci., 17: 141-164, 1970.

[29] H. Tanaka, H. Ichihashi, "A formulation of fuzzy linear programming problem based on comparison of fuzzy numbers," Control Cyber., 13: 185-194, 1984.

[30] Y.J. Lai, C.L. Hwang, "Fuzzy mathematical programming methods and applications," Springer, Berlin, 1992.

[31] S. C. Fang, C. F. Hu, H. F. Wang, S. Y. Wu, "Linear programming with fuzzy coefficients in constraints," Comput. Math. Appl., 37(10): 63-76, 1999.

[32] T. Shaocheng, "Interval number and fuzzy number linear programming," Fuzzy Sets Syst., 66(3): 301-306, 1994.

[33] C. Garcia-Aguado, J. L. Verdegay, "On the sensitivity of membership functions for fuzzy linear programming problems," Fuzzy Sets Syst., 56(1): 47-49, 1993.

[34] H. R. Maleki, "Ranking functions and their applications to fuzzy linear programming," Far East J. Math. Sci, 4(3): 283-301, 2002.

[35] Y. L. P. Thorani, P. Phani Bushan Rao, N. Ravi Shankar, "Ordering generalized trapezoidal fuzzy numbers," Int. J. Contemp. Math. Scie., 7(12): 555 – 573, 2012.

[36] M. J. Ebadi, M. Suleiman, F. B. Ismail, A. Ahmadian, M. R. Baluch Shahryari, S. Salahshour, "A new distance measure for trapezoidal fuzzy numbers," Math. Probl. Eng., Article ID: 424186, 2013.

[37] T. Allahviranloo, M. A. Jahantigh, S. Hajighasemi, "A new distance measure and ranking method for generalized trapezoidal fuzzy numbers," Math. Probl. Eng., Article ID: 623757, 2013.

[38] N. Mahdavi-Amiri, S. H. Nasseri, "Duality in fuzzy number linear programming by use of a certain linear ranking function," Appl. Math. Comput., 180: 206-216, 2006.

[39] P. K. De. Debaroti Das, "Ranking of trapezoidal intuitionistic fuzzy numbers," in Proc. 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012.

[40] T. Hasuike, "Technical and cost efficiencies with ranking function in fuzzy data envelopment analysis," in Proc. Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[41] D. Ponnialagan, J. Selvaraj, L. G. N. Velu, "A complete ranking of trapezoidal fuzzy numbers and its applications to multi-criteria decision making," Neural Comput. Appl., 30: 3303-3315, 2018.

[42] R. R. Yager, "A procedure for ordering fuzzy sets of the unit interval," Inf. Sci., 24: 143-161, 1981.

[43] H. Shahraki, S. H. Zahiri, "Design and simulation of an RF MEMS switch for removing the self: actuation and latching phenomena using PSO method," Iran J. Electr. Comput. Eng., 12: 56-63, 2013.

[44] D. W. van der Merwe, A. P. Engelbrecht, "Data clustering using particle swarm optimization," in Proc. The 2003 Congress on Evolutionary Computation, 2003. CEC '03: 215-220, 2003.

[45] S. Hettich, C. L. Blake, C. J. Merz, "UCI repository of machine learning database," Department of Information and Computer Science. University of California, Irvine, CA.

Biographies

Najme Ghanbari was born in Iran. She is the coach with the Department of Electronics Engineering, University of Zabol, Iran. She received the B.Sc. and M. Sc. degrees in electronics engineering from the University of sistan and baluchestan, Zahedan, Iran and Birjand University, Birjand, Iran, in 2002 and 2008, respectively. She is currently a Ph.D. student at Birjand University to receive a Ph.D. degree in electronics engineering. Her research interests include Pattern Recognition, Evolutionary Algorithms, and Swarm Intelligence Algorithms.

- Email: najme.ghanbari@birjand.ac.ir
- ORCID: 0000-0002-2804-1790
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hadi Shahraki received the B.Sc. degree in Electronics Engineering from University of Sistan and Baluchestan, Zahedan, Iran, in 2009 and M.Sc. degree in Electronics Engineering from Ferdowsi University of Mashhad, Mashhad, Iran, in 2012, and Ph.D. degree in Electronics Engineering from University of Birjand, in 2016.He is currently is assistant professor at the University of Sistan and Baluchestan, Zahedan, Iran. His research interests include pattern recognition and swarm intelligence algorithms.

- Email: hadi_shahraki@eng.usb.ac.ir
- ORCID: 0000-0001-9234-3577
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
N. Ghanbari, S. H. Zahiri, H. Shahraki, "Clustering of Triangular Fuzzy Data Based on Heuristic Methods," J. Electr. Comput. Eng. Innovations, 12(1): 1-14, 2024.

DOI: 10.22061/jecei.2023.9641.645

URL: https://jecei.sru.ac.ir/article_1894.html