An adaptive FCM model for interval type-2 fuzzy data and its corresponding validity index

Elahe Hajigol Yazdi
Department of Industrial Engineering, Yazd University, Iran
Elahehajigol@gmail.com

Abstract

In this paper a modified fuzzy c-means clustering method is presented which is applicable for interval type-2 fuzzy data sets. A quantifying similarity measure based on Euclidean distances function which measured the similarity between the information contained in each interval type-2 fuzzy data point is implemented for data clustering. Moreover, a modified Xie-Beni index is presented which is used for interval type-2 fuzzy sets. Some numerical examples used to show the usefulness of IT2 fuzzy c-mean clustering method. The results show IT2FCM is the superior and efficient method for interval type-2 fuzzy sets clustering.

Keywords: Interval type-2 fuzzy set, Fuzzy c-means clustering, Validity index.

Introduction

Clustering [1, 13, 17, 23] is a major branch of data mining used for discovering groups and identifying patterns and distributions of a given set of data. The goal of every clustering algorithm as an unsupervised classification [3, 14] is grouping data elements according to some (dis)similarity measures so that unobvious relations and structures in the data can be revealed. The use of clustering is supported by the cluster hypothesis which assumes that data relevant to a given cluster tend to be more similar to each other than to irrelevant data and hence are likely to be clustered together. Clustering problems arise in many fields of (computer) science, in particular in computer vision, pattern recognition, data mining and machine learning, etc.

On the other hand, in the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process [Berry and Linoff, 1996]. In classical clustering each data must be assigned to exactly one cluster. This can be a source of vagueness in case of the cluster have overlaps so afford loss of information. So this kind of ambiguity can be taken into account by means of fuzzy theory.

Depending on way of measuring similarity, information of size and shape that contains in cluster prototypes and restriction on fuzziness degree, several fuzzy clustering algorithms could be known [19, 20]. The most common fuzzy clustering techniques is fuzzy c-mean clustering algorithm which uses only cluster centers and a Euclidean distance function in comparison to the Gaustafson-Kessel algorithm which may detect clusters of different geometrical shapes in a data set by introducing an adaptive distance norm for each clusters. Here we focus on fuzzy c-means (FCM) algorithm based on that one proposed by Bezdek [6]. In this algorithm the number of cluster should be defined at first by the users. Depending on the choice of C fuzzy partitions obtained, so, it's not always possible to know c in advance. So, it is necessary to validate each of the fuzzy C-partitions once they are found. This validation is performed by a cluster validity index. By means of a validity index, we can evaluate each of fuzzy c-partitions and find the optimum partition and optimal number of clusters [24, 21,22]. There are many criteria for evaluating clustering data. In particular, Bezdek's partition coefficient (PC) and partition entropy (PE) and Xie-Beni's index have been frequently used in recent research. Major characteristic for evaluating clusters that is used in many indices is compactness and separation

In the present work here, a modified fuzzy c-mean algorithm for interval type-2 fuzzy set (IT-2) is purposed. In the purposed Type-2 FCM, we define the distance between two interval type-2 data. Then, a modified Xie-Beni index that is applicable for interval type-2 data are developed. The purposed index is consisting of two properties: compactness measure and separation measure. The compactness measure quantifies the deviation of data from center of clusters. Since data are interval type-2 fuzzy sets, we define an index for each data based on its mean and variance and then approximate the distance between two IT2 set by means of the distance between two indices. Thus, a good partition is expected to have a low degree of overlap and a large separation distance.

The remainder of this paper is organized as follows: in section 2, we present some major concepts of fuzzy c-means clustering method and some corresponding validity indices that are applied for designing optimal clustering. In section 3 we introduce a new fuzzy C-means clustering method that is useful for clustering interval type-2 fuzzy data and its corresponding validity index based on estimation of interval type-2 fuzzy data with type-1 one and measuring similarity by means of it. We implemented the purposed model for clustering some questions that is applied for calculating investor risk propensity who invests in stock market in section 4.

Fuzzy c-means algorithm

FCM clustering algorithm was first proposed by Bezdek [3]. We focus on the criterion-based methods, more specifically on the class of methods based on the fuzzy C-means (FCM) algorithm [Bezdek, 1981]. FCM algorithm which is a very powerful tool, deals with nontrivial and uncertain data. FCM as an iterative algorithm following the same steps as the K-means algorithm but C-means algorithm is much more flexible because it present objects have some inference with more than one cluster. So in This way, overlapping clusters may be realized.

In FCM clustering method, after selecting initial centers for the clusters randomly, Euclidian distance is employed as the similarity measure. The difference between each object and center of the cluster (i.e., the mean object in cluster) is weighted by a function of that object’s membership value in that cluster. The membership functions are then iteratively altered to minimize the sum of weighted differences.

Assume the vector xk , k=1,2, …, N, contained in the columns of data matrix X, will be partitioned into c clusters, represented by their prototypical vectors vi=[vi1, vi2,…,vin]T∈ Rn , i=1,2,…,c. Denote V∈Rn×c the matrix having vi in its i-th column. This matrix is called the prototype matrix. The fuzzy partitioning of data among the c clusters is represented as the fuzzy partition matrix U∈Rn×c whose element μik∈[0,1] are the membership degree of the data vector xk in the i-th cluster. A class of clustering algorithms searches for the partition matrix and the cluster prototypes such that the following objective function is minimized:

$$J(X;V,U)=\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{i,k})^{m}d^{2}(x_{k},v_{i}),$$

subjet to:

$$\sum_{i=1}^{c}(\mu_{i,k})=1, k=1,2,\dots,N$$

$$0<\sum_{k=1}^{N}(\mu_{i,k})<N, i=1,2,\dots,c \qquad (1)$$

where, m>1 is a parameter that controls the fuzziness of the clusters. The function d(xk, vi) is the distance of the data vector xk from the cluster prototype vi. So, FCM in brief is as follows:

1- Fix c (2 ≤c<n) and select a value for parameter m. Initialize the partition matrix, U(0). Each step in this algorithm will be labeled r, where r = 0, 1, 2, … .

2- Calculate the c centers {vi(r)} for each step.

3- Update the partition matrix for the rth step, U(0) as follows:

$$\mu_{ik}^{(r+1)}=\left[\sum_{j=1}^{c}\left(\frac{d_{ik}^{r}}{d_{jk}^{r}}\right)^{\frac{2}{(m-1)}}\right]^{-1} \qquad for\ I=\phi$$

$$\mu_{ik}^{(r+1)}=0 \qquad for\,all\ classes\ i\ where, \quad i\in\tilde{I}_{k}$$

$$\tilde{I}_{k}=\{1,2,\dots,c\}-I_{k}\ ,\ I_{k}=\{i\,|\,2\le c<n; d_{ik}^{(r)}=0\}$$

$$,\sum_{i\in I_{k}}\mu_{ik}^{(r+1)}=1 \qquad (2)$$

4- If $\left\|U^{(l)}-U^{(l-1)}\right\|<\varepsilon$, stop; otherwise set r = r+1 and return.

The optimal fuzzy set will be determined by using this iterative method where J is successively minimized with respect to U, V.

Traditional clustering validity indices

To determine an appropriate number of clusters C for a given data set, a cluster validity function needs to be selected. Cluster validation refers to the problem whether a given fuzzy partition fits to the data all. So it is used for evaluating the clustering results by means of a quantities objective function. Validity index is applicable to measure fitness of number of clusters and the parameterized cluster shapes. Some kinds of validity indices are usually adopted to measure the adequacy of structure recovered through cluster analysis.

In the case of fuzzy clustering algorithm, some validity indices such as partition coefficient, partition entropy, backer-Jane index, etc. for evaluating results are used only the information of membership functions. But, there are some validity index that not use only he information of membership grads but also the structure of data.

A fuzzy clustering algorithm is run over a range of c values, 2,...,cmax. Bezdek proposed the Partition Coefficient (VPC) to measure the amount of overlap between clusters as follows:

$$V_{PC}=\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{2}}{n} \qquad (3)$$

He also purposed another validity index based on Shannon's information theory named partition entropy (PE):

$$V_{PE}=-\frac{1}{n}\left(\sum_{k=1}^{n}\sum_{i=1}^{c}[\mu_{ik}\ \log_a(\mu_{ik})]\right) \qquad (4)$$

The optimal fuzzy partition is obtained by maximizing V_{PC} (or minimizing V_{PE}) with respect to $c = 2; . . . ; c_{max.}$ Xie and Beni proposed a validity index (V_{XB}) that focuses on two properties: compactness and separation. V_{XB} is defined as:

$$V_{XB}=\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}^{2}\left\|x_i-v_k\right\|^{2}}{n(\underbrace{\min}_{i\neq j}\left\{\left\|v_i-v_j\right\|^{2}\right\})} \qquad (5)$$

In this equation, the numerator is the sum of the compactness of each fuzzy cluster and the denominator is the minimal separation between fuzzy clusters. The optimal fuzzy partition is obtained by minimizing V_{XB} with respect to $c = 2,...,c_{max}$, V_{XB} decreases monotonically as $c\to n$. Partition Index V_{SC} which is presented by Bensaid and Hall [19] is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster:

$$V_{SC}=\sum_{i=1}^{c}\frac{\sum_{j=1}^{n}(\mu_{ij})^{m}\left\|x_j-v_i\right\|^{2}}{n_i\sum_{k=1}^{c}\left\|v_k-v_i\right\|^{2}} \qquad (7)$$

V_{SC} is useful when comparing different partitions having equal number of clusters. A lower value of V_{SC} indicates a better partition

All of the cluster validity indices are applicable for type-1 fuzzy sets. So they can't cover all of the ambiguity of type-2 fuzzy sets because similarity measure based on distances between crisp vectors. In the case inputs are IT-2 fuzzy sets distance between vectors defined at the different manner.

Fuzzy C-Mean clustering method for interval type-2 fuzzy data

This section introduces a new fuzzy c-mean clustering algorithm for interval type-2 fuzzy system. This method is a suitable extension of the standard fuzzy clustering that will be discussed later.

Assume both X and V_0 are fuzzy denoted by $\tilde{X}$ and $\tilde{V}_0$. It should be mentioned that the distance between x_k and v_i or x_k and v_j are calculated crisply .Moreover, m is crisp number thus u_{ik}'s are crisp.

By extending X to $\tilde{X}$ and V to $\tilde{V}$, all u_{ik}'s have membership function, too. Therefore, the elements in U_t matrix are fuzzy numbers. In other words, by obtaining fuzzy membership, we have type-2 fuzzy clustering. For this purpose, we should define a suitable fuzzy distance to generate suitable matrix of MFs.

Modified Fuzzy c-mean algorithm

Assume $\tilde{x}_i \in \tilde{X}, i=1,2,...,n$ is unlabeled fuzzy data and $\tilde{v}_{0j} \in \tilde{V}_0, j=1,2,...,c$ is initial fuzzy cluster centers and $1 \prec c \prec n$, m>1, T=iteration limit, $\varepsilon > 0$ is termination criterion and m is a parameter which determines the fuzziness of the resulting clusters.

Since our data and centers of clusters are interval type-2 fuzzy sets we have to define a new distance measure that is applicable for IT-2 sets. For achieving this mean we introduce a determination index that can approximates each IT-2 fuzzy datum with type-1 one. Many researches apply on constructing indices that is useful for fuzzy ranking. By means of this instrument, we design an index that approximates each Interval type-2 fuzzy set by a type-1 fuzzy set. As mentioned two fuzzy set with same mean may be not equal and this non equality is because of their spread. So By considering this fact the structure of this index is based on mean (M) and variance (Σ) of IT-2 fuzzy set.

$$\tilde{X} \approx \alpha.M_i + (1-\alpha).\Sigma_i \qquad (7)$$

$$\tilde{V} \approx \alpha.M_j + (1-\alpha).\Sigma_j \qquad (8)$$

So, for clustering objective data, we present a ' weighted' distance measure that can be applicable for any symmetric fuzzy data. The purposed distance measure is based on comparing each pair of fuzzy approximated variable by considering distances between their means and standard deviations, separately. It should be noted that an efficient weighting system is based on tuning coefficient and its value is between 0.85 and 0.9. A distance measure define as follow:

$$\tilde{d}_{ij}^2(w_\mu, w_\Sigma) = (w_\mu d_{ij\mu})^2 + (w_\Sigma d_{ij\Sigma})^2 \qquad (9)$$

where ,

$$d_{ij\mu} = d(M_i, M_j) = \left\| M_i - M_j \right\| \qquad (10)$$

$$d_{ij\Sigma} = d(\Sigma_i, \Sigma_j) = \left\| \Sigma_i - \Sigma_j \right\| \qquad (11)$$

Notice that w_μ is means distance weight and w_Σ is standard deviation distance between a pair of purposed index. So, we have $w_\mu + w_\Sigma = 1$ and $w_\mu > w_\Sigma$. We should tune these weights via the minimization algorithm. So, we obtain efficient weights through minimizing objective function with respect to optimal values of w_μ and w_Σ.

Then, we can appropriately tune the influence of two components of the fuzzy entities (its mean and standard deviation). The proposed 'weighted' distance measure is used for making comparisons within a set of data rather than looking at a single pair of data. Let us introduce notation that be used :

$U \equiv \{u_{ij} | o \le u_{ij} \le 1, i=1,...,n, j=1,...,c\}$ is the n×c membership matrix where n is the number of data and c present the number of clusters with respect to the following constraint:

$$\sum_{j=1}^{c} u_{ij} = 1, \qquad u_{ij} \ge 0. \qquad (12)$$

So, we can define the objective function of the fuzzy clustering algorithm as follows:

$$J(M,U,V) = \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m d^2(x_i, v_j) \qquad (13)$$

Since our data is IT2 fuzzy so we can use this formulation as objective function of FCM for IT-2 fuzzy set:

$$\tilde{J}(M,\tilde{U},\tilde{V}) = \sum_{j=1}^{c}\sum_{i=1}^{n} \tilde{\mu}_{ij}^m \tilde{d}^2(\tilde{x}_i, \tilde{v}_j) \qquad (14)$$

By using the indices for estimation $\tilde{U},\tilde{V}$ the objective function convert to:

$$\begin{aligned}\tilde{J}(U,V,w) &\equiv \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m \tilde{d}_{ij}^2(w_\mu, w_\Sigma) = \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m \left[(w_\mu d_{ij\mu})^2 + (w_\Sigma d_{ij\Sigma})^2\right] \\ &= \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m \left[w_\mu^2 d_{ij\mu}^2 + (1-w_\mu^2) d_{ij\Sigma}^2\right]\end{aligned} \qquad (15)$$

The objective function should be minimized with respect to the degree of fuzziness (m), membership degree of data into clusters and the mean and variance weights. Therefore, the fuzzy clustering model is characterized as follows:

$$\tilde{J}(U,V,w) \equiv \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m \tilde{d}_{ij}^2(w_\mu, w_\Sigma) = \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^m \left[w_\mu^2 d_{ij\mu}^2 + (1-w_\mu^2) d_{ij\Sigma}^2\right] \qquad (16)$$

St:

$$\sum_{j=1}^{c} u_{ij} = 1, \qquad u_{ij} \ge 0, \qquad (17)$$

$$w_{\mu} \ge 0.5. \qquad (18)$$

As indicated before, the range for the membership exponent is m∈ [1, ∞]. For the case m=1, the distance norm is Euclidean and the FCM algorithm approaches a hard c-means algorithm; i.e., only 0's and 1's come out of the clustering. Conversely, as m→∞, the value of the function Jm→0. This result seems intuitive, because the membership values are usually less than or equal to 1, and the large powers of fractions less than one approach zero. The exponent m thus controls the extant of membership sharing between fuzzy clusters. If all other algorithmic parameters are fixed, then increasing m will result in decreasing Jm. No theoretical optimum choice of m has emerged in the literature. However, the bulk of the literature seems to report values in the range 1.25 to 2. Convergence of algorithm tends to be slower as the value of m increases.

Theorem. The iterative solution of this functional problem is as follows:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{\left[w_{\mu}^{2} d_{ij\mu}^{2} + (1-w_{\mu})^{2} d_{ij\Sigma}^{2}\right]}{\left[w_{\mu}^{2} d_{ik\mu}^{2} + (1-w_{\mu})^{2} d_{ik\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}}} \qquad (19)$$

$$M_{j} = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} M_{i}}{\sum_{i=1}^{n} \mu_{ij}^{m}}. \qquad (20) \qquad\qquad \Sigma_{j} = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} \Sigma_{i}}{\sum_{i=1}^{n} \mu_{ij}^{m}}. \qquad (21)$$

Proof. Let us first get the optimal membership degrees μ_{ij}'s , *i* = 1, . . . , *n*; *j* = 1, . . . , *c*. By considering the Lagrangian function:

$$L(\mu_{ij},\lambda) \equiv \sum_{j=1}^{c}\sum_{i=1}^{n} \mu_{ij}^{m}\left[w_{\mu}^{2} d_{ij\mu}^{2} + (1-w_{\mu}^{2}) d_{ij\Sigma}^{2}\right] - \lambda\left(\sum_{j=1}^{c}\mu_{ij} - 1\right), \qquad (22)$$

We consider the partial deviation of above objective function and set it to 0.

$$\frac{\partial L(\mu_{ij},\lambda)}{\partial \lambda} = \left(\sum_{j=1}^{c}\mu_{ij}\right) - 1 \qquad (23)$$

and

$$\frac{\partial L(\mu_{ij},\lambda)}{\partial \mu_{ik}} = m\mu_{ij}^{m-1}\left[w_{\mu}^{2} d_{ij\mu}^{2} + (1-w_{\mu}^{2}) d_{ij\Sigma}^{2}\right] - \lambda = 0 \qquad (24)$$

Then :

$$\mu_{ij} = \left[\frac{\lambda}{m\left[(w_{\mu}^{2} d_{ij\mu}^{2} + (1-w_{\mu}^{2}) d_{ij\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}} \qquad (25)$$

Then upon restriction (17) describe as follows:

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} \sum_{k=1}^{c} \left[\frac{1}{m\left[(w_{\mu}^{2} d_{ik\mu}^{2} + (1-w_{\mu}^{2}) d_{ik\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}} = 1. \qquad (26)$$

Therefore,

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{k=1}^{c}\left[\frac{1}{\left[w_{\mu}^{2} d_{ik\mu}^{2} + (1-w_{\mu}^{2}) d_{ik\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}}} \qquad (27)$$

then :

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{\left[w_{\mu}^{2} d_{ij\mu}^{2} + (1-w_{\mu})^{2} d_{ij\Sigma}^{2}\right]}{\left[w_{\mu}^{2} d_{ik\mu}^{2} + (1-w_{\mu})^{2} d_{ik\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}}} \qquad (28)$$

By fixing *U*,*V* we can obtain value of w_{μ} .

$$w_{\mu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{m} d_{ij\mu}^{2}}{\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{m}(d_{ij\mu}^{2} + d_{ij\Sigma}^{2})}. \tag{29}$$

$$\tilde{J}(U,V,w) \equiv \sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}\left[w_{\mu}^{2}d_{ij\mu}^{2} + (1-w_{\mu}^{2})d_{ij\Sigma}^{2}\right] = \sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{m}\left[w_{\mu}^{2}\left\|M_{i}-M_{j}\right\| + (1-w_{\mu})^{2}\left\|\Sigma_{i}-\Sigma_{j}\right\|\right] \tag{30}$$

After setting first derivatives with respect to M_j, Σ_j equal to zero:

$$\frac{\partial J(U,V,w_{\mu})}{\partial M_{j}} = w_{\mu}^{2}\sum_{i=1}^{n}\mu_{ij}^{m}(M_{j} - M_{i}) = 0 \tag{31}$$

$$\frac{\partial J(U,V,w_{\mu})}{\partial \Sigma_{j}} = w_{\mu}^{2}\sum_{i=1}^{n}\mu_{ij}^{m}(\Sigma_{j} - \Sigma_{i}) = 0 \tag{32}$$

So, we have:

$$M_{j} = \frac{\sum_{i=1}^{n}\mu_{ij}^{m}M_{i}}{\sum_{i=1}^{n}\mu_{ij}^{m}}. \tag{33}$$

$$\Sigma_{j} = \frac{\sum_{i=1}^{n}\mu_{ij}^{m}\Sigma_{i}}{\sum_{i=1}^{n}\mu_{ij}^{m}}. \tag{34}$$

So, the steps of purposed Type-2 FCM is as follows:

Step1: store unlabled interval type-2 fuzzy data:

$\tilde{x}_i \in \tilde{X}, i = 1,2,...,n$, $\tilde{x}_i$ is an IT-2 fuzzy number for i=1,...,n.

Step2:choose

c : the number of cluster such that 1<c<n

m : weightening exponential as fuzziness measure, m>1

ε : termination criterion ,0< ε <1

initial fuzzy vector $\tilde{v}_{0j} \in \tilde{V}_0, j = 1,2,...,c$

step 3 : approximate each $\tilde{X}, \tilde{V}$ with suggested indices as:

$$\tilde{X} \approx \alpha.\mu_{\tilde{X}} + (1-\alpha).\Sigma_{\tilde{X}} \tag{35}$$

And

$$\tilde{V} \approx \alpha.\mu_{\tilde{V}} + (1-\alpha).\Sigma_{\tilde{V}} \tag{36}$$

Step 4: iterate : for t=1 to T

calculate U_t with V_t and μ_{ij} as:

$$\mu^{t}{}_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{\left[w_{\mu}^{2}d_{ij\mu}^{2} + (1-w_{\mu})^{2}d_{ij\Sigma}^{2}\right]}{\left[w_{\mu}^{2}d_{ik\mu}^{2} + (1-w_{\mu})^{2}d_{ik\Sigma}^{2}\right]}\right]^{\frac{1}{m-1}}} \tag{37}$$

calculate M_j, Σ_j based on U_t as:

$$M_{j} = \frac{\sum_{i=1}^{n}\mu_{ij}^{m}M_{i}}{\sum_{i=1}^{n}\mu_{ij}^{m}}. \tag{38}$$

$$\Sigma_{j} = \frac{\sum_{i=1}^{n}\mu_{ij}^{m}\Sigma_{i}}{\sum_{i=1}^{n}\mu_{ij}^{m}}. \tag{39}$$

IF $E_t = \|(U_{t-1}) - (U_t)\| \le \varepsilon$ (40)
stop, Else Next *t*

Modified Xie-Beni index

Assume $\tilde{x}_i \in \tilde{X}, i = 1,2,\ldots,n$ is unlabeled fuzzy data and $\tilde{v}_{0j} \in \tilde{V}_0,\ j = 1,2,\ldots,c$ is initial fuzzy cluster centers and $1 \prec c \prec n$, m>1, T=iteration limit, $\varepsilon > 0$ is termination criterion and m is a parameter which determines the fuzziness of the resulting clusters then state above we define distance measure as:

$$d_{ij\mu} = d(\mu_i, \mu_j) = \|M_i - M_j\|,\quad d_{ij\mu} = d(\Sigma_i, \Sigma_j) = \|\Sigma_i - \Sigma_j\| \qquad (41)$$

$$d^2 \equiv \quad \|M_i - M_j\|^2 + \|\Sigma_i - \Sigma_j\|^2 \qquad (42)$$

Since the Xie-Beni index measure the compactness and separation of clusters then $V_{XB} = \frac{compactness}{separation}$ we have

$$compactness = \frac{\sigma}{n} \qquad (43)$$

Assume that we have clusters $j = 1,2,\ldots,c$ with $v_1, v_2, \ldots, v_c$ centers of clusters. Because the type of fuzzy data is IT2 then:

$$V_e(clust_j) = \frac{\sum_{i=1}^{n} [\tilde{x}_i - \tilde{v}_j]^2 \tilde{\mu}_{ij}^m}{\sum_{i=1}^{n} \tilde{\mu}_{ij}} \qquad (44)$$

$$\sum_{i=1}^{n} (\mu_{ij}^m [\|M_i - M_j\|^2 + \|\Sigma_i - \Sigma_j\|^2]) \Big/ \sum_{i=1}^{n} \mu_{ij} \qquad (45)$$

$$V_{clust} = \cup_{j=1}^{c} V_{e_e}(clust_j) = [v_l(clust), v_r(clust)] \qquad (46)$$

$$v_l(clust) = Min(v_{clust}) \qquad (47) \qquad\qquad v_r(clust) = Max(v_{clust}) \qquad (48)$$

$$V_{\mathrm{mod}ify} = \frac{v_l(clust) + v_r(clust)}{2} \qquad (49)$$

$$\sigma_{clust} = \sqrt{V_{\mathrm{mod}ify}} \qquad (50)$$

Compactness= $\frac{\sigma_{clust}}{n}$ (51) For separation we should maximize the minimum distance between cluster centers. So, we use:

Separation $\approx \min_{fuzzyranking} \{\tilde{d}^2(\tilde{v}_i, \tilde{v}_j)\}$

$$\chi_i = \frac{\sigma_{clust}}{n(\min_{fuzzyranking} \{\tilde{d}^2(\tilde{v}_i, \tilde{v}_j)\}} \qquad (52)$$

Numerical example

In this section we show the results of our fuzzy c-means clustering algorithm for interval type-2 fuzzy data applied to the risk propensity data that collected from 10 investor prefers. The data are interval type-2 fuzzy sets because it is based on human opinion and human view often contains some ambiguous. On the other hand, we use average of answers given to each question for clustering and it might be another source of vagueness. So, we use IT-2 fuzzy data. The advantage of our work is in that we don't fix the number of clusters in priority. The data related to the answer of 10 questions is obtained by 6 investors. The answers take the linguistically following value: Highly disagree, Slightly disagree, Makes no difference, Slightly agree, Highly agree. Therefore, this ambiguity implies that the data managed hardly. data clustering is take placed based on the average of answers and data is interval type-2 fuzzy set (the membership function is Gaussian (symmetric). so, the purposed FCM algorithm can be applied. By applying the designed FCM algorithm we obtain 3 clusters as follows:

	1	2	3

1	0.9999990	3.3991301E-9	9.3628539E-7
2	2.0592466E-9	0.9999998	1.6611927E-7
3	0.9999990	3.3991301E-9	9.3628539E-7
4	6.3970784E-8	6.6243813E-8	0.9999998
5	5.0466778E-4	7.6911726E-6	0.9994876
6	0.9999884	1.2794790E-7	1.1464319E-5
7	5.0466778E-4	7.6911726E-6	0.9994876
8	2.4691644E-4	0.01347331	0.9862797
9	2.0592466E-9	0.9999998	1.6611927E-7
10	9.5599095E-8	0.9999963	3.5632508E-6

Table 1- membership degree matrix

So, the questions cluster as:

Cluster 1	Cluster 2	Cluster 3
1,3,6	2,9,10	4,5,7,8

Table 2- Clustering results

We preprocessed the data (see Remark 2) and then we performed our fuzzy c-means clustering model. After running several models using different values of m, we decided to set m=1.3 and by using the modified cluster validity index we obtain the optimal number of clusters equal to 3. The membership degree matrix U given in Table 2. From Table 3, we can easily distinguish the clusters of question number 1,3,6 set in cluster 1 ,question 2,9,10 set in cluster 2 and finally we set question number 4,5,7,8 in cluster 3.the questions are as follows:

1	Personal wealth is not an issue and financial loss doesn't keep me awake at night
2	I'd rather be in one investment with the chance of a high return, than in a broad selection with less chance
3	If the market is volatile, it's not a worry - I'll still trade
4	I'm willing to take as much time as I need to oversee my investments
5	I enjoy the excitement of investment trading
6	Price swings in investments I own are of little concern
7	I don't need a steady dividend income and would rather have capital growth
8	It doesn't bother me that my investments are not easily tradable
9	Investment in emerging markets or high-tech research is more appealing to me than traditional blue chips
10	I don't feel the need to frequently check my portfolio and can leave it alone for long periods

Table 3- the questions that implemented the purposed FCM on it

By thinking about meant of question we can found the most related question set in one cluster.

Conclusions and future works

In this paper, we present a fuzzy clustering model for interval type-2 fuzzy data based on weighted distance measure which constructed from two indexes. These two indices are acceptable approximation of interval type-2 fuzzy data which handle appropriate information of initial data. Then distance measure as an index for showing similarity use for optimization lost function that object is minimize the weighted distances between mean of data and mean of center of clusters and their corresponding standard deviations. Mean of data has more critical role than standard deviation. We use standard deviation formulation demonstrated in the paper that specially designed for data with Gaussian interval type-2 fuzzy membership function. For achieve coefficient cluster number we run validity index that specialized for interval type-2 fuzzy data based on introduced distance measure. Using this measure for identifying optimal number of clusters is an advantage that prevents system non optimality. The performance of these clustering method is presented and their results show in the simulation study part that mark its great power in identifying clusters so that the data are related to special cluster with a greet membership function.

Reference

[1] M.R. Anderberg, Cluster Analysis for Application, Academic Press, NewYork, 1973.
[2] S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognition 37 (2004) 481–501.
[3]J.C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Cornell University, Ithaca, NY, 1973.

[4]J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybernet. 3 (1974) 58–73.

[5]J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 57–71.

[6]J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork, 1981.

[7]J.C. Bezdek, Pattern Recognition in Handbook of Fuzzy Computation, IOP Publishing Ltd., Boston, NY, 1998 (Chapter F6).

[8]J.C. Bezdek, R.J. Hathaway, M.J. Sabin, et al., Convergence theory for fuzzy c-means: counter-examples and repairs, IEEE Trans. Systems, Man Cybernet. SMC17 (1987) 873–877.

[9] J.C. Bezdek, N.R. Pal, Cluster validation with generalized Dunn's indices, in: N. Kasabov, G. Coghill (Eds.), Proc. 1995 Second NZ Internat.

[10]J.C. Bezdek, N.R. Pal, Some new indices of cluster validity, IEEE Trans. Systems, Man and Cybernet. 28 (1998) 301–31.

[11]M. Bouguessa, S.R. Wang, A new efficient validity index for fuzzy clustering, in: Proc. Third Internat. Conf. on Machine Learning and Cybernetics, Shanghai, 26–29 August 2004.

[12]R.N. Davé, Clustering relational data containing noise and outliers, Pattern Recognition Lett. 12 (1991) 657–664.

[13]P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982.

[14] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybernet. 3 (1974) 32–57.

[15]Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, in: Proc. Fifth Fuzzy Systems Symp., 1989, pp. 247–250.

[16]I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 773–781.

[17]J.A. Hartigan, Clustering Algorithms,Wiley, NewYork, 1975.

[18]D.W. Kim, K.H. Lee, D. Lee, On cluster validity index for estimation of the optimal number of fuzzy clusters, Pattern Recognition 37 (2004), 2009–2025.

[19]R. Krishnapuram, O. Nasraoui, J. Keller, The fuzzy c spherical shells algorithm: a new approach, IEEE Trans. Neural Networks 3 (5) (1992), 663–671.

[20]Y. Man, I. Gath, Detection and separation of ring-shaped clusters using fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 16 (8) (1994) 855–861.

[21] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, Fuzzy Set and Systems 155 (2) (2005) 191–214.

[22]N.R. Pal, J.C. Bezdek, On cluster validity for fuzzy c-means model, IEEE Trans. Fuzzy Systems 3 (3) (1995) 370–379.

[23]G.E. Tsekouras, H. Sarimveis, A new approach for measuring the validity of the fuzzy c-means algorithm, Adv. in Eng. Software 35 (2004), 567–575.

[24]X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 841–847.

[25]Y. Xie, V.V. Raghavan, P. Dhatric, X.Q. Zhao, A new fuzzy clustering algorithm for optimally finding granular prototypes, Approx. Reason. 40(2005) 109–124.