I.J. Mathematical Sciences and Computing, **2018, 4, 1-18**
Published Online November 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2018.04.01

***Available online at** http://www.mecs-press.net/ijmsc*

A New Credibilistic Clustering Method with Mahalanobis Distance

Ahad Rafati [a], Shahin Akbarpour [b]

[a] *Young Researchers and Elite Club Ilkhchi Branch, Islamic Azad University, Ilkhchi, Iran*
[b] *Department of Computer Engineering Shabestar Branch, Islamic Azad University, Shabestar, Iran*

Received: 06 April 2018; Accepted: 19 July 2018; Published: 08 November 2018

Abstract

The fuzzy c-means (FCM) is the best known clustering and use the degree of membership fuzzy to data clustering. But the membership is not always for all data correctly. That is, at scattered dataset belonging is less and noisy dataset belonging is more assigned and local optimization problem occurs. Possibility c-means (PCM) was introduced to correspond weaknesses FCM approach. In PCM was not self-duality property. In other words, a sample membership for all clusters be assigned more than one and basic condition FCM were violated. One of the new method is Credibilistic clustering and based on credibility theory proposed that is used to study the behavior of fuzzy phenomenon. The aim is to provide new Credibilistic clustering approach with replacing credibility measure instead of the fuzzy membership and Mahalanobis distance use in FCM objective function. Credibility measure has self-duality property and solves coincident clustering problem. Mahalanobis distance used instead of Euclidian distance to separate cluster centers from each other and dens samples of each cluster. The result of proposed method is evaluated with three numeric dataset and Iris dataset. The most important challenge will be how to choose the initial cluster centers in the noisy dataset. In the future, we can be used FCM combined with particle swarm optimization.

Index Terms: Credibilistic clustering, Fuzzy C-Means, Data Mining, Possibility C-Means, Credibility measure, Mahalanobis Distance.

1. Introduction

Clustering is known as one of the effective ways of unsupervised learning and methods of pattern recognition. A cluster is a collection of samples that are similar to each other and different from other samples of clusters. Fuzzy c-means (FCM) clustering method is one of the best known algorithms in the field of

* Corresponding author.
E-mail address: a.rafati.1364@iaushab.ac.ir, akbarpour@iaushab.ac.ir

clustering. The membership belongs to the scattered and noisy dataset are not properly set in this method [12]. Possibility C-Means clustering method are presented to solve Fuzzy c-means (FCM) clustering method`s problems [3].

Possibility C-Means clustering methods have the problem of coincident clusters in the clusters which are close to each other and this is because of absence of self-duality property in the Possibility clustering. In order to solve the problems of Fuzzy c-means (FCM) clustering method and Possibility C-Means clustering method different combined methods are presented with different advantages and disadvantages. And then Credibility theory is presented by Liu [7]. The most important property of this theory is the existence of self-duality which solves the problems of Fuzzy c-means (FCM) clustering method and Possibility C-Means clustering method. The Structure of Article includes: In section two the review of Fuzzy c-means (FCM) clustering method and Possibility C-Means clustering method is presented. In section three suggested methods will be discussed in details. In section four results and analysis of the mentioned methods will be explained with the figures and tables and compared with the results of FCM clustering method and Possibility C-Means clustering method. And then the summary and further researches will be presented in section five.

2. Review of Literature

The aim of cluster analysis is to partition dataset into a cluster (group, and classes). This partition may include two characteristics: The first one: Homogenous among clusters, it means the same data must belong to the same cluster. The second one: Nonhomogeneous among clusters, it means the different data must belong to the different clusters. The main classification of clustering includes: Crisp and fuzzy. Fuzzy clustering is an approach for classifying of data which uses Fuzzy theory. Possibility clustering were introduced after Fuzzy methods. The Proposed method used hybrid approach to reduce the number of outliers [22]. The number of outlier can only reduce by improving the cluster formulation method. The authors used two data mining techniques for cluster formulation i.e. weighted k-means and neural network where weighted k-means is the clustering technique that can apply on text and date data set as well as numeric data set. Weighted k-means assign the weights to each element in dataset.

In [5] credibility measure instead of possibility measure and used Mahalanobis distance to introduce a new clustering method. The goals of model are to minimize average of cluster centers error and separating clusters from each other. Further, the results improved than FCM and PCM. In [11] Zhong-Guo Shi proposed a fuzzy model of scenario planning based on the credibility theory and fuzzy programming. In this article is presented insoluble problems of scenario planning, and proposed a fuzzy model of scenario planning. In [20] Chen and et al proposed a new fluid identification method in carbonate reservoir based on the modified Fuzzy C-Means (FCM) Clustering algorithm. Both initialization and globally optimum of cluster center are produced by Chaotic Quantum Particle Swarm Optimization (CQPSO) algorithm, which can effectively avoid the disadvantage of sensitivity to initial values and easily falling into local convergence in the traditional FCM Clustering algorithm.

The automatic identification of cluster numbers algorithm utilizes a hard partition approach in the process of integration and does not make full use of the membership information from each fuzzy c-means (FCM) clustering result. To address this problem, an automatic fuzzy clustering algorithm is proposed in [21], combining the soft partition method with the membership information from each FCM clustering result. An adaptive neural system which solves a problem of clustering data with missing values in an online mode with a permanent correction of restorable table elements and clusters' centroids is proposed in [23]. The proposed neural system is characterized by both a high speed and a simple numerical implementation.

In [6] Li and et al proposed a new clustering algorithm based on particle swarm optimization and called PSOFCM, and also the new clustering method to improve FCM weakness. It avoids the local optima and has robust to initialization. The experimental results compared with IRIS data. Hesam and Ajith [15] presented a new hybrid fuzzy clustering method based on FCM and FPSO called FCM-FPSO. The FCM algorithm is faster than the FPSO algorithm as result of fewer function evolutions, but it doesn't fall into local optima. The

experimental results show the FCM-FPSO method is efficient.

Conditional FCM was proposed by Pedrycz in [17] in which a conditional variable was propounded for each schema. The structure of these schemas exposes propounding their vicinity in sample space conditioned to the value of these assumed variables. In [18], pedrycz and waletzky modified FCM handling partial supervision. The main idea was operation of labeled data in order to cluster data set. The objective function has been mutated in order to calculate the membership value of labeled and unlabeled data in clusters. In [19], pedrycz proposed algorithm of knowledge-based clustering using FCM. The new clustering method using partial supervision and proximity-based knowledge have been elaborated in [19]. The importance of this paper is on human centricity of the clustering framework.

A. FCM Clustering Algorithm

FCM algorithms tries to find the optimal membership Matrix and The Matrix of cluster`s center and minimize the objective function. The results of the objective function can be calculated from Eq. (1) and the second one must be established. In the mentioned formula m is a real number, greater than one and the best value, m=2, is proven. In Eq. (1), if the value of m is equal to one, non-Fuzzy objective function will be obtained. This method minimizes all the common variables with the limited repetitions. Eq. (4) is an update to membership and updating of clusters center is done with Eq. (5) [12].

$$min\{J_m(U,V,X)\sum_{i=1}^{c}\sum_{k=1}^{n}U_{ik}^{m}d_{ik}^{2}=\sum_{i=1}^{c}\sum_{k=1}^{n}U_{ik}^{m}\|x_k-v_i\|^2\} \tag{1}$$

$$\begin{cases}0\le u_{ik}\le 1, & for\ all\ i,k\\ \sum_{i=1}^{c}u_{ik}=1, & for\ all\ k\end{cases} \tag{1}$$

$$d_{ik}=\|x_k-v_i\| \quad for\ all\ i,k \tag{2}$$

$$U_{ik}=\left(\sum_{k=1}^{c}\frac{d_{ik}^{2/(m-1)}}{d_{ik}^{2/(m-1)}}\right)^{-1} \quad for\ all\ i,k \tag{3}$$

$$V_i=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \quad for\ all\ i \tag{4}$$

B. PCM Clustering Algorithm

Although FCM clustering method is useful, the degree of membership is not always well matched with the belonging data. So Possibility clustering method [3], in Eq. (6), was introduced to cover the weaknesses of FCM method by Krishnapuram and Keller.

$$min\{F_m(U,V;X,\eta)=\sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\,d_{ik}^{2}+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{N}(1-u_{ik})^m\} \tag{6}$$

$$0<\sum_{k=1}^{N}u_{ki}\le N\ for\ all\ i \tag{7}$$

$$u_{ki}\in[0,1]\ for\ all\ i\ and\ k \tag{5}$$

$$\max u_{ik}>0 \ \text{ for all k} \tag{9}$$

$$F_m(U,V;X,\eta)=\sum_{k=1}^{N}\sum_{i=1}^{c}F_m^{ik}\,(U,V;X,\eta) \tag{10}$$

$$F_m^{ik}(U,V) = u_{ik}^m d_{ik}^2 + \eta_i(1-u_{ik})^m \tag{11}$$

$$u_{ik} = \frac{1}{1+\left(\frac{d_{ik}^2}{\eta_i}\right)^{1/(m-1)}} \qquad 1 \le i \le c; \qquad 1 \le k \le N \tag{12}$$

η is the permeability coefficient and a fixed amount. When the rows and columns of the U function are independent, the F objective function must be minimized and it is calculated with the Eq. (11) and then the Eq. (12) and (5) are used to update the membership function and the clusters center. Since this method, U_{ik}, just depends on ith cluster, makes Coincident clusters. To solve the problem of the calculation of η, Eq. (13) was suggested.

$$\eta_i = k\frac{\sum_{k=1}^n u_{ik}^m d_{ik}^2}{\sum_{k=1}^n u_{ik}^m}, k > 0 \ \ (usullu\ k = 1) \tag{13}$$

Krishnapuram and Keller [3] showed that they replaced their formula to solve the problems of clustering. This formula describes Typicality as well as possible and solves the problem of noisiness automatically. They showed that PCM method can produce strong clustering results compared to FCM method in the field of the noisy dataset. However, PCM algorithms can be sensitive to the initial cluster samples values. Another Possibility algorithm was introduced by Young in 2006. After the randomly initializing clusters center, μ membership Matrix and clusters center Matrix are updated and can be calculated through the Eq. (14).

$$J_{PCA06}(\mu, A) = \sum_{i=1}^c \sum_{j=1}^n \left(\mu_{ij}\right)^m \left\|x_j - a_i\right\|^2 + \frac{\beta}{m^2\sqrt{c}} \sum_{i=1}^c \sum_{j=1}^n \left[\left(\mu_{ij}\right)^m ln(\mu_{ij})^m - \left(\mu_{ij}\right)^m\right] \tag{14}$$

$$0 \le \mu_{ij} \le 1 \qquad for\ all\ i.j \tag{15}$$

In this algorithm, updating of membership done with the Eq. (16). The β parameter is defined with the Eq. (17) and Eq. (5) is used to update the cluster center as well as in FCM method.

$$\mu_{ij} = exp\left\{-\frac{m\sqrt{c}d_{ij}^2}{\beta}\right\} \quad for\ all\ \ i,j \tag{16}$$

$$\beta = \frac{\sum_{j=1}^n \left\|x_j - \bar{x}\right\|^2}{n} \quad with\ \ \bar{x} = \frac{\sum_{j=1}^n x_j}{n} \tag{17}$$

Possibility theory is widely used to measure fuzzy events; however it doesn't have self-duality property. Self-duality measure which was introduced by Liu and Liu has been named Credibility measure. This method was introduced as a branch of mathematics to study the behavior of Fuzzy phenomena. And then Credibility clustering algorithms will be evaluated in the next section.

C. CCM Clustering Algorithm

Another group of clustering models based on objective function is Credibility clustering. Credibility clustering is a new approach of clustering. This method which is based on Credibility theory with the self-duality measure was suggested to cover the weaknesses of PCM and FCM methods. Li and Liu presented a sufficient and necessary condition for credibility measure in [7]. Before introducing the related works to Credibilistic clustering, a brief review of credibility measure will be presented. Let θ be a nonempty set, and let $p(\theta)$ be the power set of θ. Each element in $p(\theta)$ is called an event. In order to present an axiomatic definition

of credibility, it is necessary to assign to each event a number Cr {A} which indicates the credibility that A will occur. In order to ensure that the number Cr {A} has certain mathematical properties of credibility, there are following five axioms.

$$\begin{cases} cr\{\theta\} = 1 \\ \text{Cr is increasing, i. e. } cr\{A\} <= cr\{B\} wheever\ A \subseteq B \\ \text{Cr is self} - \text{dual, i. e. }, cr\{A\} + cr\{A^c\} = 1 \quad for\ any\ A \in p(\theta) \\ cr\{U_i A_i\} \wedge 0.5 = SUP_i cr\{A_I\} \quad for\ any\ \{A_I\}\ with\ cr\{A_I\} \leq 0.5 \end{cases}$$

The set function Cr is called a credibility measure if it satisfies the four axioms. When the credibility measure is applied in fuzzy clustering, self-duality of it causes the clustering algorithm to utilize the benefits of FCM without the imposition of its constraint. So, the information about the centers and membership degrees of the other clusters contributes to construct the membership degrees of each cluster. It prevents the creation of coincident clusters which is the shortcoming of PCM. Also, this measure uses the benefit of PCM to deal with the noises. Although the constraint of FCM is relaxed, but the self-duality keeps the advantages of this constraint. Possibility measure was replaced with Credibility measure and the results of clustering were improved and a new approach called Credibility clustering was introduced (Zhou & et al, 2014). Eq. (18) has been defined according to Credibility clustering method and the Eq. (19), (20) and (21) must be established.

$$min\left\{J_{CCA}(Cr, A) = \sum_{j=1}^{n} \sum_{i=1}^{c} \left(Cr_{ij}\right)^m \left\|x_j - a_i\right\|^2\right\} \tag{18}$$

$$sup_{1 \leq i \leq c} Cr_{ij} \geq 0.5 \quad for\ j \tag{19}$$

$$Cr_{ij} + sup_{k \neq i} Cr_{kj} = 1 \quad for\ any\ \ i, j\ \ with\ Cr_{ij} \geq 0.5 \tag{20}$$

$$0 \leq Cr_{ij} \leq 1 \quad for\ \ all\ \ i, j \tag{21}$$

In $\mathrm{Cr} = \left(\mathrm{Cr_{ij}}\right)_{\mathrm{c \times n}}$ presents a credible Matrix and i=c shows the amount of clusters and j which is equal to n, shows the total numbers of samples. To determine Crij updated equation, μ_{ij} memberships' assessment is determined. To μ_{ij} membership, Eq. (22) and (23) must be obtained. To normalize the memberships Eq. (24) must be used and Eq. (25) is used to normalize clusters center. In Eq. $\mathrm{sup}\mu_{\mathrm{kj}}^{*} = 1$ for all j and Eq. $0 \leq \mu_{ij}^{*} \leq 1$ for all i and j, after that, Crij credibility will be calculated and updated with the Eq. (26).

$$0 \leq \mu_{\mathrm{ij}} \leq 1, \ for\ all\ \ i, j \tag{22}$$

$$sup_i \mu_{\mathrm{ij}} = 1\ , \text{for all i, j} \tag{23}$$

$$\mu_{ij}^{*} = \frac{\mu_{ij}}{sup_k \mu_{kj}} \quad for\ \ all\ \ i, j \tag{24}$$

$$v_i = \frac{\sum_{j=1}^{n} (Cr_{ij})^m x_j}{\sum_{j=1}^{n} (Cr_{ij})^m} \quad for\ all\ \ i \tag{25}$$

$$\mathrm{Cr_{ij}} = \frac{1}{2}\left(\mu_{\mathrm{ij}}^{*} + 1 - \underset{k \neq i}{\mathrm{sup}\, \mu_{\mathrm{kj}}^{*}}\right) \tag{26}$$

3. The New Suggested CCM Clustering Algorithm

The suggested method is a Credibilistic clustering method. This method consists of a new objective function with the Credibility measure and Mahalanobis distance. In this approach Mahalanobis distance is replaced instead of Euclidian distance. This method is presented to solve the problems and the weaknesses of PCM and FCM with the use of properties of Credibility theory. The suggested approach is calculated through the Eq. (27).

$$J_{CCA}(\mu, a) = \sum_{j=1}^{n}\sum_{i=1}^{c}\sum_{k=1}^{c}\left(Cr_{ij}\right)^{m}|d_{ij}^{2} - dv_{ik}^{2}| + \frac{\beta}{m^{2}\sqrt{c}}\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\left(Cr_{ij}\right)^{m} ln(Cr_{ij})^{m} - \left(Cr_{ij}\right)^{m}\right] \quad (27)$$

In Eq. (27), the amount of m is fixed and is called fuzzification. In the FCM method the best amount is proved m=2. The numbers of clusters is equal to c and n is the numbers of total samples in dataset.

$$0 < \sum_{j=1}^{N}\mu_{ij} \leq N \; for \; all \; i \; , \; and \; \mu_{ij} \in [0,1] \; for \; all \; i \; and \; j \quad (28)$$

$$dv_{ik}^{2} = \|v_k - v_i\|^2 \; ; \; d_{ij}^{2} = \left\|x_j - v_i\right\|^2 \quad for \; all \; i,j,k \quad (29)$$

Eq. (28) is the same with FCM. Eq. (29) determines the distance from the center of the sample x and the center of clusters from each other which this approach uses more than one cluster center to measure the distance of data`s sample.

$$\mu_{ij} = exp\left\{-\frac{m\sqrt{c}d_{ij}^{2}}{\beta}\right\} \quad for \; all \; \; i,j \quad (30)$$

$$\beta = \frac{\sum_{j=1}^{n}\|x_j - \bar{x}\|^2}{n} \quad with \; \; \bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (31)$$

Eq. (30) is used to determine Fuzzy membership. Eq. (31) is due to obtain variance and to determine scattered data in the range of average value and Eq. (32) is used for updating clusters Center. Eq. (33) shows the relationship between Possibility and Credibility and the Credibility Matrix measure is calculated through Fuzzy membership in this relation.

$$v_i = \frac{\sum_{j=1}^{n}(Cr_{ij})^{m} x_j}{\sum_{j=1}^{n}(Cr_{ij})^{m}} \quad for \; all \; \; i \quad (32)$$

$$Cr_{ij} = \frac{1}{2}\left(\mu_{ij} + 1 - SUP_{i\neq k}\mu_{kj}\right) \quad (33)$$

Eq. (33) also shows belonging credibility of jth cluster sample to ith one. In this relation the relationship among two samples and the others are determined by using of Credibility measure. Self-duality is an instinct property of Credibility measure and the limitation of FCM solved through this method. The steps of this method are shown in the figure 1.

The first purpose of Eq. (27) is showing the relationships between jth sample`s membership into the ith cluster and the differentiation from d_{ik}^{2}, dv_{ij}^{2} and reduce the sum of the squares from the distance between v_i and all other clusters center. In this method there is an attempt to maximize the compression of inside of the clusters and isolation of the clusters from each Other. And the second purpose is to prevent the local minimums.

In this modeling, Crij is Credibility belonging of x_j into ith cluster. In each period, the total square of distance between the center of the ith cluster and other clusters is maximum; it means that trying to separate clusters is done. Eq. (34) is used to reach to this purpose. The distance of jth sample, x_i, is minimum from ith

cluster. Here the aim is to compress samples in each cluster maximally. Eq. (35) shows that:

$$\sum_{i=1}^{c} dv_{ik}^2 = \sum_{i=1}^{c} \|v_k - v_i\|^2 \tag{34}$$

$$d_{ij}^2 = \left\|x_j - v_i\right\|^2 \tag{35}$$

In some models, Validity index is used to separate clusters, in other words, the objective function is not used for this purpose. In the presented model, cluster separation is intended with the objective function and in addition, the compression within the cluster is intended in terms of the objective function.it means that in the objective function and by the use of Credibility theory`s property, clusters evaluation is took place. This features and advantages solve the needs for a separate evaluation of clusters, again. It can be evaluated with the obtained clusters and be reliable.

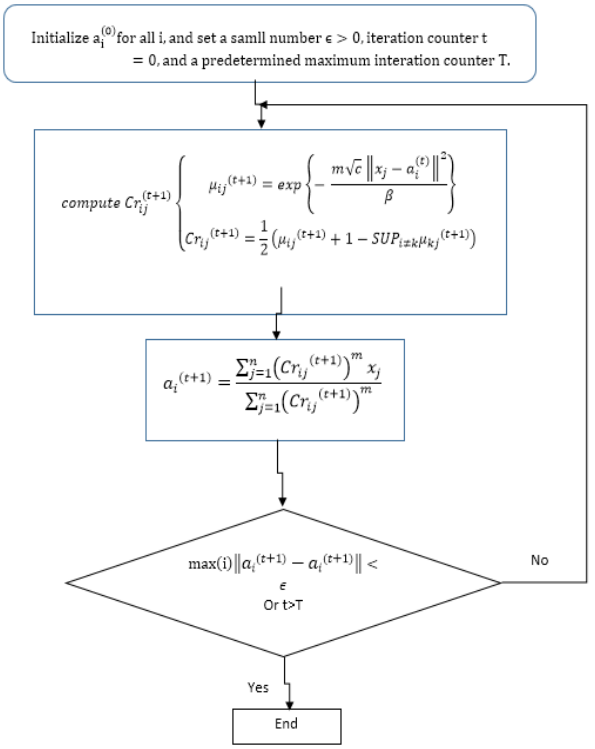


Fig.1. The Steps of Proposed CCM

4. Evaluation Results

Evaluation results of the suggested method done with the three sets of two-dimensional numerical data and Iris four-dimensional set which has been used in previous researches. In [6], the number of samples for each cluster is 200 in three random numerical dataset. It means that the number of samples for the first dataset which contains four clusters is 800, for the second data set which contains six clusters is 1200 and for the third data set with seven clusters is 1400.In the article, 25 samples in each cluster is assumed 1000. It means that, for the number of samples for the first dataset is 4000, the second data set is 6000 and the third dataset is 7000. But

according to this method which has been suggested to use in large datasets, so in all three sets of numerical data for each cluster, 10000 samples are included which are two-dimensional data. Thus, the first data set includes four clusters and the number of all samples is 40000.The second dataset includes 60000 samples. In the third data set, the number of clusters is seven and the number of samples is 70000. The first table shows the way of producing of these numerical set.

For the implementation of the suggested method, 2013 MATLAB software with 64-bits 7 OS and the following characteristics are used. Intel Core i5-2450M CPU 2.50GHZ with RAM DDR3-6G.Firstly,the aim is to minimize the termination measure of the objective function, whatever the measure obtained is minimal, the results for clustering will improve. Another purpose is to decrease the cluster center`s average error in all three datasets.

Table 1. Dataset Production to Evaluate Clustering (Center and Radius)

Data set	**1**	**2**	**3**
Formulation	*The random points*	*The random points*	*The random points*
	V=(-5,3), r=3	V=(-10,-10), r=4	V=(-12,-12), r=3
	V=(5,5), r=3	V=(-3,-2), r=4	V=(-5,-5), r=3
	V=(10,12), r=3	V=(-4,5), r=5	V=(-12,10), r=4
	V=(15,17), r=3	V=(8,2), r=3	V=(-2,3), r=2
		V=(15,17), r=3	V=(4,3), r=3
		V=(18,9), r=4	V=(4,12), r=4
			V=(4,-11), r=4

And the ultimate goal is to compress the inside of the clusters and separate the center of clusters from each other with the using of Mahalanobis distance in the objective function. And then, the results for the suggested method of Mahalanobis distance for all four data collection and is described and evaluated in details.

A. The First Dataset's Results Evaluation

The results of new Credibilistic clustering are shown with Mahalanobis distance with obtained figures. The second figure shows the clustering result after implementation which is contained four clusters and there is 10000 samples in each cluster.

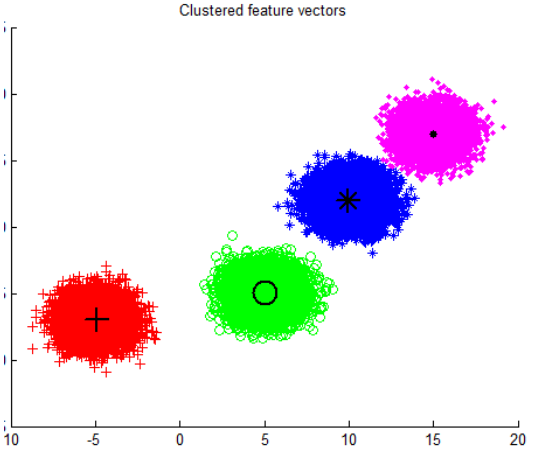


Fig.2. The first 40000 Samples Credibilistic Clustering Result

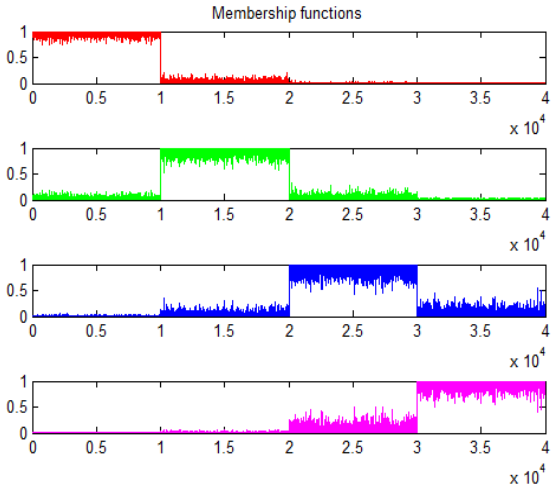


Fig.3. The First 40000 Samples Fuzzy Memberships

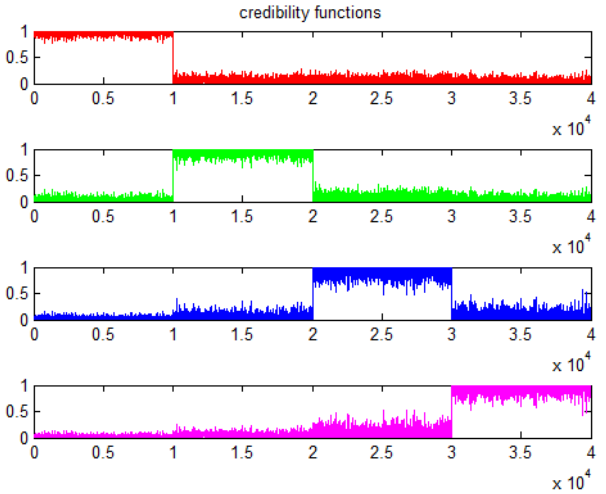


Fig.4. The First 40000 Terminate Credibility Measure

In Figure 3, the termination measure of Fuzzy memberships which is obtained from Eq. (30) for the first 10000 datasets sample is shown. One of the important FCM conditions is that membership values should be between zero and one and is shown in the figure. The figure 4 shows the last Credibility Matrix for the first 400000 datasets which contains 4 clusters and the results obtained from Eq. (33). The amount of Credibility Matrix is between zero and 1 as same as Fuzzy membership but according to Credibility theory must be accomplished with Eq. (36) and this is Credibility measure`s self-duality`s property.

$$Cr_{ij} + sup_{k \neq i} Cr_{kj} = 1 \; for \; any \;\; i,j \; with \; Cr_{ij} \geq 0.5 \tag{36}$$

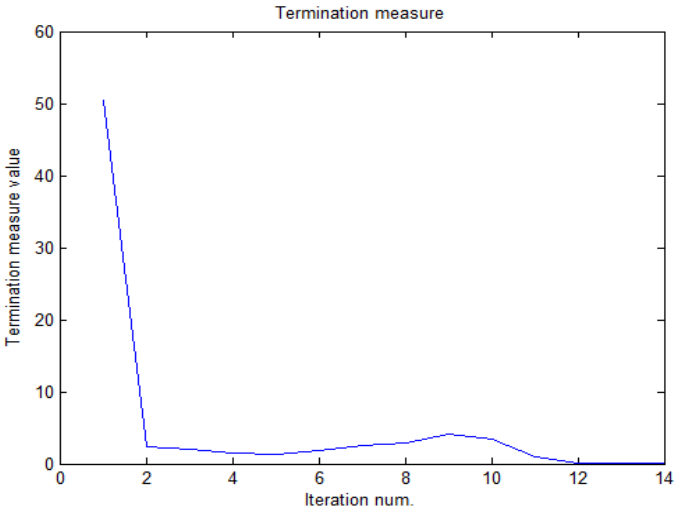


Fig.5. The Terminate Measure of the Objective Function and the Frequency of the First Datasets

The figure 5 shows the process and obtained results which are obtained with Eq. (27) for the first datasets and as it is clear the final results obtained after 14 times repetitions.

B. The Evaluation Results Provided for the Second Datasets

The figure 6 shows the results of clustering after implementation and there are 10000 samples in each cluster. The figure 7 shows the final amount of Credibility memberships.

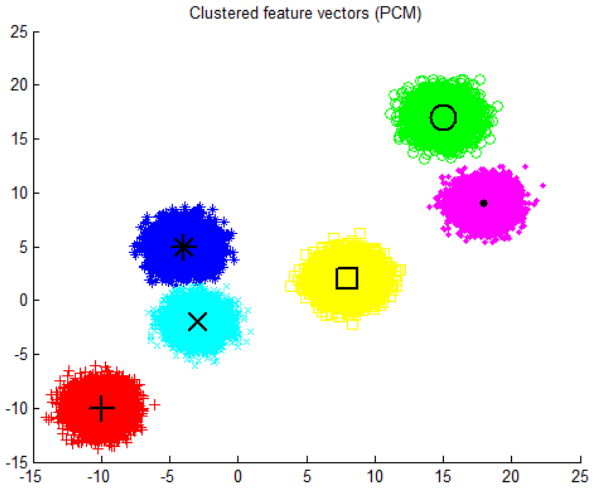


Fig.6. The SECOND 60000 SAMPLES Credibilistic CLUSTERING Results

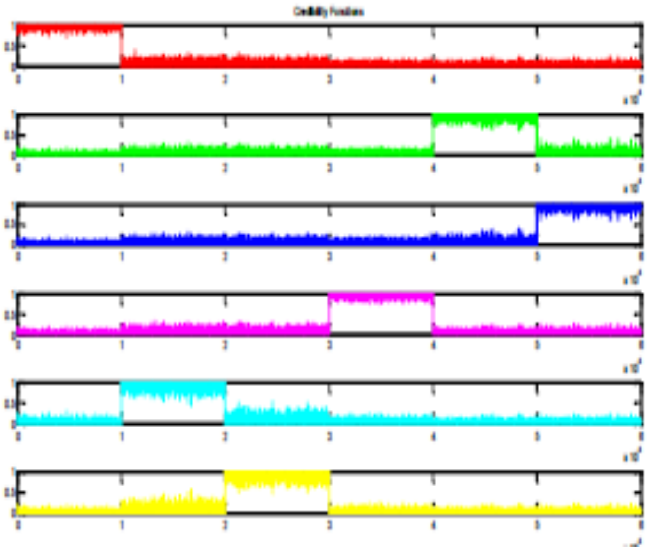


Fig.7. The Second 60000 Credibility Measure

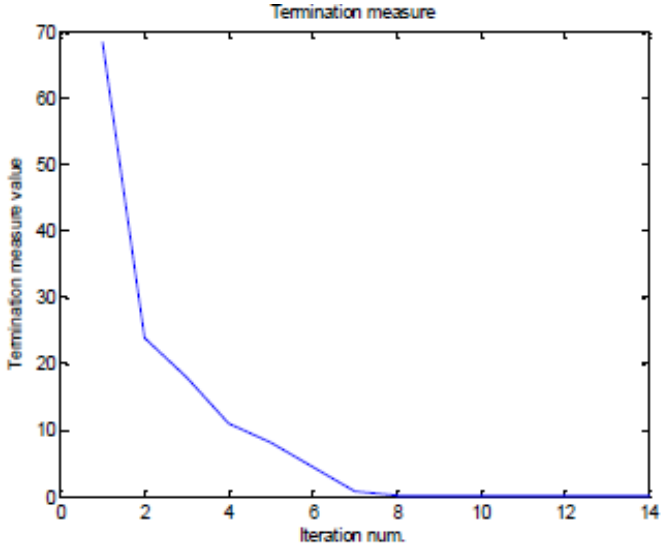


Fig.8. The Final Amount of the Objective Function and the Frequency of the Second Datasets

The figure 8 shows the processes and the obtained amount for the second datasets and as its clear obtained after 14 times repetitions. And then the third dataset results will be shown.

C. The Evaluation Result Provided for the Third Datasets

The figure 9 shows the results of clustering after implementation which contains 7 clusters and there are 10000 samples in each cluster. The figure 10 shows the final amount of Credibility Matrix.

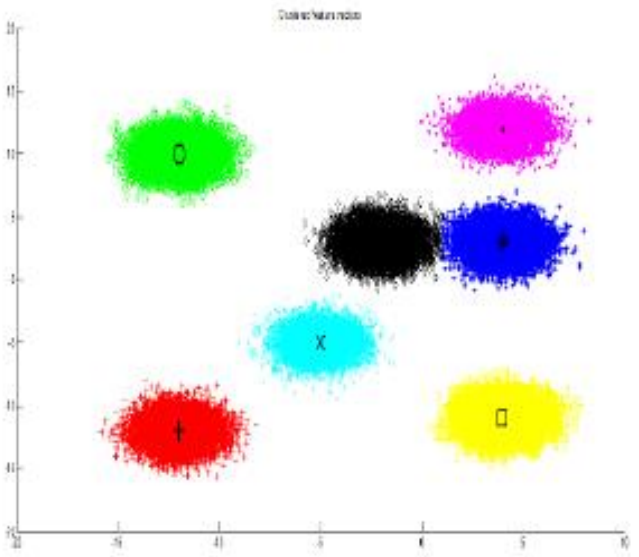


Fig.9. The Result of 70000 Third Sample's Credibilistic Clustering

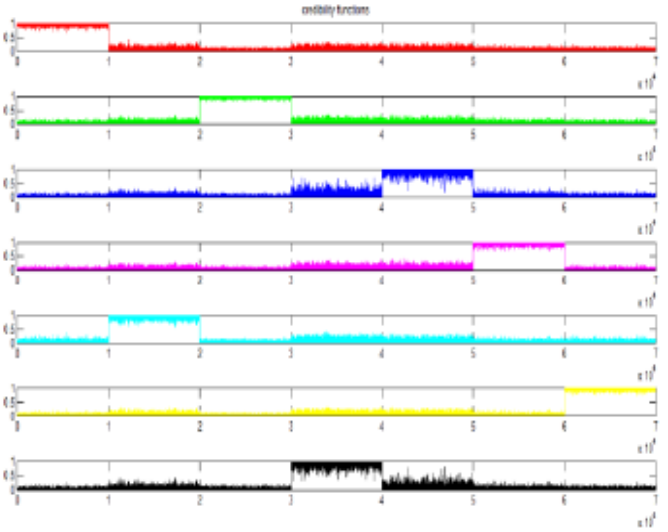


Fig.10. The Third 70000 Samples Credibility Measure

The figure 11 shows the processes and the obtained results for the third datasets and as its clear obtained after 15 times repetitions. And then Iris dataset`s results will be shown.

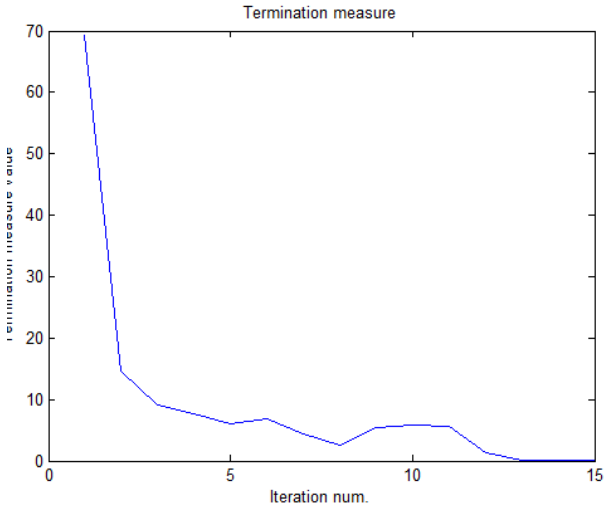


Fig.11. The Final Amount of the Objective Function and Frequency of Repetitions

D. The Assessment Results of Credibilistic Clustering Through Mahalanobis Distance and Iris Dataset

The figure 12 shows the results of Iris dataset after implementation and equipped three clusters correctly and the results of clustering are clear.

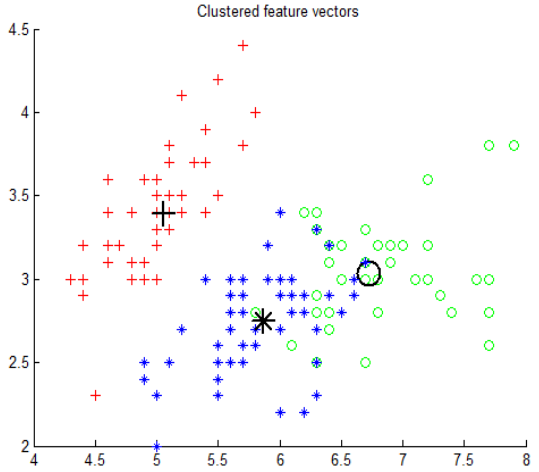


Fig.12. The Results of Iris Dataset Clustering

Figure 13 shows the final amount of Credibility Matrix and as it is mentioned in section 4.1; they have the Fuzzification conditions and Credibility theory property.

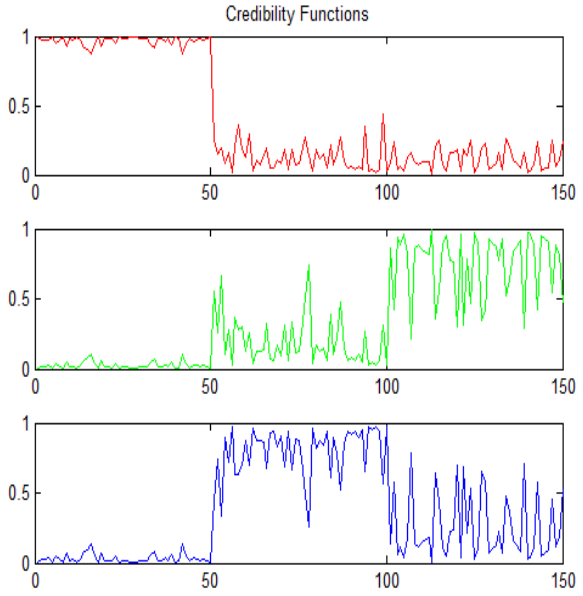


Fig.13. Iris Dataset Termination Measure Value

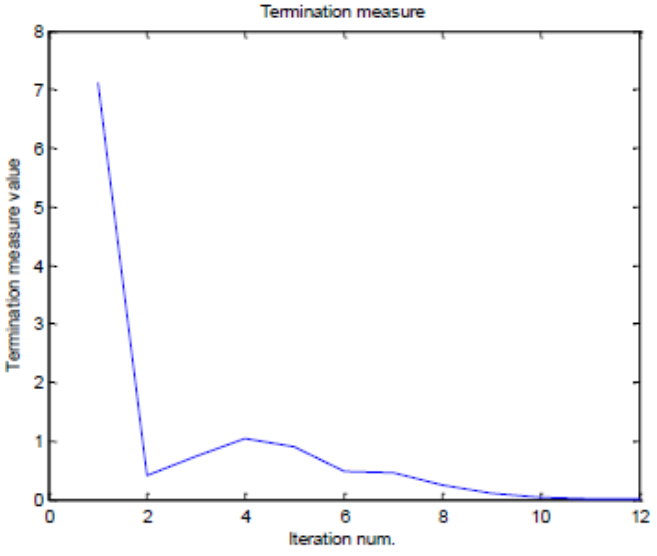


Fig.14. The Objective Function of Termination Measure and Frequency of Iris Dataset

Figure 14 shows the processes and the results of the Iris dataset and shows that the final result has obtained after 12 times repetitions. And then the comparison between the results of FCM and PCM will be expressed. One of the criteria for assessing the quality of the presented method is to obtain cluster`s center`s average errors distance. In three sets of two-dimensional numerical data, clusters center is defined and it will be compared

with the actual results obtained from clusters center. It means that, the accuracy of the clustering depends on the obtained cluster centers.

$$\sum_{i=1}^{c} \|v_i - v_i^*\| / c \tag{37}$$

The mean of clusters center is calculated through Eq. (37). In Eq. (37) c is the number of clusters and vi is the center of real cluster in the two-dimensional dataset and $vi*$ is the amount of actual obtained cluster center. The differences between the two are divided by the number of clusters to obtain average error. whatever the result of Eq. (31) be minimum, the cluster center`s average error will be less and the results obtained are of high quality and clustering well done. In table 2, the clusters center`s average error obtained through the new Credibilistic clustering method will compare with the PCM and FCM.

The new method has the lowest average error of cluster center, on the other hand, In PCM clustering method with the problem of coincident clusters the average difference between clusters is the greatest. In Credebilistic clustering method, according to the precious assigning of Fuzzy membership to the all samples, the impact is evident in the results of clustering and determining accurate clusters center. FCM is as same as Fuzzy c-mean clustering, PCM is as same as Possibility c-mean clustering and CCM is Credibilistic clustering. CCM-N is a presented method of new Credebilistic clustering with the using of Mahalanobis distance.

Table 2. Comparison of the Clusters Center's Average Error in New Method with the FCM and PCM

Data set	**Clustering model**			
	CCM-N	*CCM*	*PCM*	*FCM*
Data set 1	0.0499	0.13	1.54	2.61
Data set 2	0.0375	0.21	1.73	2.69
Data set 3	0.0392	0.33	2.86	2.71

And the other measure is the average distance from the clusters center from each other. In other word, this measure shows the separation of clusters and the clusters must have enough distance as far as possible. But in the Credibilistic clustering methods, besides to compress samples, has been tried to make an enough distance from each other in comparison with the other methods. The average distance of clusters is calculated with relation 38.So that the distance from the center of the cluster obtained and finally divided by the number of clusters to obtain the average distance from the clusters center.

$$\sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} \|v_i - v_j\| / c \tag{38}$$

In this method, the distance of each cluster`s center from each other is maximized in comparison with the other methods. One of that reason is the reduction of the cluster`s center average error. Another reason is the properties of Credibility theory and Credibility measure which the compression of the clusters center maximized. The average distance among clusters center, the clusters distance, has been compared with PCM and FCM in table 3. And lastly, the clusters isolation distance will be reduced when the number of clusters gets more. Because the clusters are in a dense area.

Table 3. The Comparison of the Clusters Average Distance in New Methods with the PCM and FCM

Data set	**Clustering model**			
	CCM-N	*CCM*	*PCM*	*FCM*
Data set 1	6.49	6.24	4.87	5.03
Data set 2	6.76	6.11	4.38	4.97
Data set 3	6.51	5.83	2.94	4.37

Another reasonable indication for better understanding, the obtained number of clusters has been assumed in the three numerical dataset and Iris dataset. Table 4 shows the number of clusters in the presented methods has been compared with FCM and PCM. The obtained results show that there are more or less clusters in PCM, the important reason for this is the lack of Self-duality in PCM. This property makes PCM does not properly belong to Fuzzy membership. In new Credibilistic clustering methods and Credibilistic clustering the number of clusters are equal to the number of expected clusters.

Table 4. Comparison of the number of Clusters with the Previous Methods

Data set	**Clustering model**			
	CCM-N	*CCM*	*PCM*	*FCM*
Data set 1	4	4	4	4
Data set 2	6	6	6	6
Data set 3	7	7	8	7
Data set 4	3	3	2	3

The validity index which presented by Zhu and Huang, is named overall accuracy. Overall accuracy is presented based on the results of clustering and labeled rating of the all samples in dataset. Here, Iris dataset is used. Overall accuracy can be used to evaluate clusters credibility in various algorithms. Because of better efficiency the new Credibilistic clustering is used in Iris dataset and for comparison with PCM and FCM. Table 5 shows the results of evaluation of Iris dataset and it has been compared with the other methods. This table contains the best measure and average measure for the results and also the number of obtained clusters for overall accuracy. The best answer for the new Credibilistic clustering is the Euclidian distance and Credibility measure. On the contrary the least measure is obtained for PCM. The average of overall accuracy means that the algorithm has been carried out 10 times and the obtained average measures has been calculated. According to the table 5, it is clear that the number of obtained clusters in PCM is two and in FCM, CCM are three. It means that, Coincident clusters will be made in PCM.

Table 5. The Evaluation of Iris Dataset through the Overall Accuracy

Model clustering	**Validity accuracy**		**Cluster number**
	Best value	*Average value*	
FCM	89.33	89.33	3
PCM	84.00	66.30	2
CCM	91.33	90.06	3
CCM-N	90.89	90.89	3

And lastly, as it is clear, according to the obtained results of three sets of numerical data and to compare them with the previous FCM, PCM and CCM, the new Credibilistic clustering method is the most efficient one. The evaluation measures show that the number of clusters, the average error from the clusters center and the average isolation of clusters are well answered and resulted. The results presented with Iris dataset and evaluated with the overall accuracy and in comparison with the other methods showed that it is the best method among the previous methods.

5. Conclusion and Further Researches

The advantage of presented method was the elimination of the problems of Coincident clusters in PCM methods and the avoiding of local optimization in FCM methods. The evaluation measures of obtained presented and concluded that the results in terms of the clusters average error`s center was minimum. Also the clusters center is very close in comparison with the previous methods. Then the number of clusters obtained in the presented method is done correctly. Other measure was the average distance of clusters which shows that in

the method, the clusters have the highest distance in comparison with the other methods. And finally, the repetition times of frequency are lower and reaching the answers is more quickly. On the contrary, there are still challenges and problems in Fuzzy clustering and Credibilistic clustering. When the datasets have different features, choosing of the initial cluster center is one of the most important challenging. Choosing of the number of clusters is another challenge in such ways. For the further researches Credebilistic clustering algorithms with the particle swarm methods and the ways of the developments can be analyzed and studied. The new objective can be proposed and the application of those in all of the other sets can be analyzed by changing of the distance measuring function, the use of other evaluation distance measures.

References

[1] P. Wen, J. Zhou, and L. Zheng, "A modified hybrid method of spatial credibilistic clustering and particle swarm optimization," *Soft Comput.*, vol. 15, no. 5, pp. 855–865, 2011.

[2] A. Proceedings, A. F. Shapiro, and M. Koissi, "Credibility Theory in a Fuzzy Environment," 2013.

[3] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, 1993.

[4] M. S. Yang and K. L. Wu, "Unsupervised possibilistic clustering," *Pattern Recognit.*, vol. 39, no. 1, pp. 5–21, 2006.

[5] M. Rostam Niakan Kalhori, M. H. Fazel Zarandi, and I. B. Turksen, "A new credibilistic clustering algorithm," *Inf. Sci. (Ny).*, vol. 279, pp. 105–122, 2014.

[6] L. Wang, Y. Liu, X. Zhao, and Y. Xu, "Particle Swarm Optimization for Fuzzy c-Means Clustering," *Sixth World Congr. Intell. Control Autom. 2006. WCICA 2006*, pp. 6055–6058, 2006.

[7] B. Liu, "A survey of credibility theory," *Fuzzy Optim. Decis. Mak.*, vol. 5, no. 4, pp. 387–408, 2006.

[8] J. Zhou, Q. Wang, C.-C. Hung, and X. Yi, "Credibilistic Clustering: The Model and Algorithms," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 23, no. 04, pp. 545–564, 2015.

[9] D. Wang, B. Han, and M. Huang, "Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics," *Phys. Procedia*, vol. 24, pp. 1186–1191, 2012.

[10] H. Izakian and A. Abraham, "Optimization," pp. 1690–1694, 2009.

[11] Z. Shi, "A Fuzzy Model of Scenario Planning Based on the Credibility Theory And Fuzzy Programming," no. Isora, pp. 211–218, 2011.

[12] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984.

[13] T. a. Runkler and C. Katz, "Fuzzy Clustering by Particle Swarm Optimization," *2006 IEEE Int. Conf. Fuzzy Syst.*, no. 3, pp. 601–608, 2006.

[14] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 517–530, 2005.

[15] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, 2011.

[16] K. E. Permana and S. Z. M. Hashim, "Fuzzy membership function generation using particle swarm optimization," *Int. J. Open Probl. Compt. Math*, vol. 3, no. 1, 2010.

[17] W. Pedrycz, Conditional fuzzy C-means, Patt. Recogn. Lett. 17 (1996) 625–631.

[18] W. Pedrycz, J. Waletzky, Fuzzy clustering with partial supervision, IEEE Trans. Syst. Man Cybernet. B: Cybernet. 27 (1997) 787–795.

[19] W. Pedrycz, Collaborative and knowledge-based fuzzy clustering, Int. J. Innov. Comput. 3 (2007) 1–12.

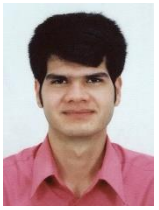
[20] L. Liu, S. Z. Sun, H. Yu, X. Yue, and D. Zhang, "A modified Fuzzy C-Means (FCM) Clustering algorithm and its application on carbonate fluid identification," *J. Appl. Geophys.*, 2016.

[21] C. Hai-peng, S. Xuan-jing, L. Ying-da, and L. Jian-wu, "A novel automatic fuzzy clustering algorithm based on soft partition and membership information," *Neurocomputing*, no. September, pp. 1–9, 2016.

[22] N. Lekhi, M. Mahajan, "Outlier Reduction using Hybrid Approach in Data Mining," *I.J. Modern Education and Computer Science*, 2015, 5, 43-49.
[23] Z. Hu, Y. V. Bodyanskiy, Oleksii K. Tyshchenko and Vitalii M. Tkachov, "Fuzzy Clustering Data Arrays with Omitted Observations," *I.J. Modern Education and Computer Science*, 2017, 6, 24-32.

Authors' Profiles

Shahin Akbarpour was born in Iran in 1972. He received the B.S. degree in Computer science from the University of Isfahan, Iran, in 1996; the M.Sc. degree in Mathematics applied in O.R. from the Islamic Azad University, Iran, in 1999, and the Ph.D. degree in Intelligent Computing from the University Putra Malaysia, Malaysia, in 2011. He joined Islamic Azad University, Shabestar, Iran, in 1999. His main areas of research interest are computer vision, data mining, and web mining.
Dr. Akbarpour is an academic member of Departments of Computer and is currently Assistant Professor.



Ahad Rafati was born in Tabriz, Azarbayjan Sharghi, Iran in 1985. He receied his BSc. degree in 2010 and MSc. Degree in software engineering from Islamic Azad University of Shabestar, in 2016. He is member of Young Researchers and Elite Club Ilkhchi Branch, Islamic Azad University, Ilkhchi, Iran. He is working on many projects in the field of data mining and he has oracle certification. In additional, he published many papers in journals and conferences that are related to data mining approaches. Generally, he most like researching and working with expert teams related to data mining, mobile programming with Xamarin. My profile in linkedin: https://www.linkedin.com/in/ahad-rafati-22243655/.

How to cite this paper: Ahad Rafati, Shahin Akbarpour,"A New Credibilistic Clustering Method with Mahalanobis Distance", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.4, No.4, pp.1-18, 2018.DOI: 10.5815/ijmsc.2018.04.01