Comparison of FCM and FISODATA

B. Fergani
MISC Lab, Mentouri
University of Constantine
Algeria

M.K. Kholladi
MISC Lab, Mentouri
University of Constantine
Algeria

M. Bahri
Mentouri
University of Constantine
Algeria

ABSTRACT
In fuzzy clustering, the fuzzy c-means (FCM) clustering algorithm is the best known and used method. An interesting extension of FCM is the fuzzy ISODATA (FISODATA) algorithm; it updates cluster number during the algorithm. That's why we can have more or less clusters than the initialization step. It's the power of the fuzzy ISODATA algorithm comparing to FCM. The aim of this paper is to compare FCM and FISODATA results.

General Terms
Machine Intelligence, Fuzzy Systems.

Keywords
Fuzzy clustering, FCM, FISODATA.

1. INTRODUCTION
Clustering is a process for classifying objects or patterns in such a way that objects of the same group are more similar to one another than objects belonging to different groups. In other words, it's a process for clustering a dataset $X = \{x_1 \ldots x_n\}$ in a p dimensional space R^P into $1<c<n$ subgroups of similar objects. This is done by assigning labels to the vectors in X, and hence, to the objects generating X.

Many clustering methods have been used (see [1]), such as the hard clustering methods and the fuzzy clustering methods. The hard clustering methods restrict each object of the data set to exactly one cluster. Since Zadeh proposed fuzzy sets that produced the idea of partial membership described by a membership function, fuzzy clustering has been studied and applied in many fields, for example: image segmentation (see [1], [2-5]), content based image retrieval system (see [6]), e-mail filtering (see [7]),etc.

In the literature on fuzzy clustering, there are various fuzzy clustering techniques proposed by researchers, like: the Fuzzy C-Means (FCM), the Possibilistic C-Means (PCM), and Fuzzy Possibilistic C-Means (FPCM). Although FCM is the most known and used method, it is not perfect because of its drawbacks: 1. The FCM algorithm assigns in some cases equal memberships to the same object x_i , in other words x_i is equidistant from two clusters. To overcome this problem Krishnapuram and Keller in [8] proposed a new clustering algorithm called Possibilistic C-Means (PCM), which helps to identify outliers (noise individuals). The PCM algorithm is very sensitive to initialization and sometimes generates coincident clusters. In order to enhance fuzzy clustering results, several researchers propose several algorithms. Yang et al. propose another PCM algorithm in [9]. Pal et al. propose a clustering algorithm that combines the characteristics of both fuzzy and possibilistic c-means (FPCM). A modified fuzzy possibilistic clustering algorithm is presented in [10]; it is developed to obtain better quality of clustering results. The objective function is based on adding new weight of data individuals in relation to every cluster and modifying the exponent of the distance between an object and a class. The MFPM algorithm increases the cluster compactness and the separation between clusters. For further improvement in clustering accuracy, Vanisri in [11] introduces a repulsion term in the objective function.

2. The FCM algorithm is sensitive to initializations; clustering algorithms typically require the user to specify the number of cluster centers and their locations. The quality of the solution depends strongly on the choice of the initial values. Fuzzy ISODATA (FISODATA) which is an extension of FCM algorithm updates cluster number during the algorithm; it has a capability of self organizing by splitting and merging clusters. The purpose of this paper is to compare FCM and FISODATA results.

This paper is organized as follows. Section2 presents the FCM algorithm. Section3 describes the fuzzy ISODATA clustering algorithm (FISODATA). Section4 briefly describes the IRIS dataset; it presents experimental results and discussions. Finally, section 5 concludes this paper.

2. FUZZY C-MEANS (FCM)
The fuzzy C-means (FCM) can be seen as the fuzzy version of the K-means algorithm. It's a method of clustering which allows one object of the data set to belong to two or more clusters [10]. This method was proposed by Dunn [12] and it was modified by Bezdek.

The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error J_{FCM} (the objective function).

$$J_{FCM}=\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^{m}d^{2}(x_j,v_i) \quad (1)$$

Where $1<m<\infty$ is the fuzzifier, $X=\{x_1, x_2 \ldots x_n\} \subset R^P$ is the data set in the p dimensional vector space, p is the number of data items, $2<=c<=n-1$ is the number of clusters. $V = \{v_1, v_2 \ldots v_c\}$ is the c centers of the clusters; V_i is the p dimension center of the cluster i, and $d^2(x_j, v_i)$ is a distance measure between object x^j and cluster center V_i. Note that $d^2(x_j,v_i) = \|x^j - v_i\|^2$ is the most used. $U = \{u_{ij}\}$ Represents a fuzzy partition matrix with $\{u_{ij}\}$ is the degree of membership of x_j in the i^{th} cluster.

The fuzzy partition matrix satisfies:

$$\forall\, i,j \;\; 1 \le i \le c, 1 \le j \le n \quad U_{ij} \in [0,1] \quad (2)$$

$$\forall\, i,j \;\; 1 \le i \le c, 0 \le \sum_{j=1}^{n} U_{ij} < n \quad (3)$$

$$\forall \;\; 1 \le j \le n, \sum_{i=1}^{c} U_{ij} = 1 \quad (4)$$

The following can be derived by optimizing the objective function in (1) with respect to U and V.

$$U_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{1}{d_i^2(x, v_i)}\right)^{2/(m-1)}} \quad (5)$$

$$v_i = \frac{\sum_{j=1}^{N}(U_{ij})^m \, x^k}{\sum_{j=1}^{N}(U_{ij})^m} \quad (6)$$

2.1. FCM algorithm

Step1: initialize: the number of clusters c, the threshold ε, the fuzzifier m, iteration counter I=0. The initialization of membership function and centers is done in two ways (see section 4).

Step2: Compute U_{ij}^{I+1} using equation (5)

Step3: Compute v_{ij}^{I+1} using equation (6)

Increment I until $\|U^{I+1} - U^{I}\| < \varepsilon$

2.1.1. Cluster Validity

Since most of the fuzzy clustering algorithms need to pre-assume clusters number, a validity criterion for finding an optimal c becomes the most studied topic in cluster validity. For a given cluster number range validity measure is evaluated for each given cluster number and then an optimal number is chosen for these validity measures. The following four indexes are the most cited for fuzzy clustering [9]:

- *Partition coefficient PC(c):* it is defined by:

$$PC(c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}^2. \quad (7)$$

Where: $1/c \le PC(c) \le 1$.
In general, the optimal cluster number c is found by solving: $max_{\,2\le c\le n-1} PC(c)$.

- *PE index:* it is defined by

$$PE(c) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij} log_2 U_{ij}. \quad (8)$$

Where $0 \le PE(c) \le log_2 c$.
In general, the optimal c is found by solving $min_{\,2\le c\le n-1} PE(c)$

- *FS index:* it is a validity function proposed by FUKUYAMA and SUGENO, it's defined by:

$$FS(S) = \sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}^m \left\|x_j - V_i\right\|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}^m \|Vi - \bar{V}\|^2$$

$$FS(S) = J_{FCM}(U,V) + K(U,V) \quad (10)$$

Where $\bar{V} = \sum_{i=1}^{c} Vi/c$, $J_{FCM}(U,V)$ is the FCM objective function which measures the separation. In general, an optimal c is found by solving $min_{\,2\le c\le n-1} PS(c)$ to produce a best clustering performance for the dataset X.

- *XB index:* it is a validity function proposed by XIE and BENI with m=2, and then generalized by PAL and BEZDEK. It's defined by:

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}^m \|x_j - V_i\|^2}{n \, min_{ij}\|Vi - V_j\|^2} \quad (11)$$

$$XB(c) = \frac{J_{FCM}(U,V)/n}{Sep(V)} \quad (12)$$

$J_{FCM}(U,V)$: is a compactness measure and $Sep(V)$ is a separation measure.

In general, an optimal c is found by solving $min_{\,2\le c\le n-1} XB(c)$ to produce a best performance for the dataset X.

3. FUZZY ISODATA (FISODATA)

Fuzzy ISODATA algorithm is based on the FCM algorithm; it employs processes of merging and splitting, it's an extension of ISODATA [13]. Clusters are merged if either objects number in a cluster is less than a certain threshold (avoiding too small clusters) or if the centers of two clusters are closer than a certain threshold (so closed clusters are merged). Clusters are split into two different clusters if the cluster standard of variation exceeds a predefined threshold, so dissimilar clusters are split.

3.1. FISODATA algorithm

Step1: initialize the number of clusters c, the threshold ε, the fuzzifier m and iteration counter I=0. The initialization of membership function and centers is done in two ways (see below).

Step2: Compute U_{ij}^{I+1} using equation (5)

Step3: Compute v_{ij}^{I+1} using equation (6)

Step4: Merge similar clusters.

Step5: Split dissimilar clusters.

Increment I until $\|U^{I+1} - U^{I}\| < \varepsilon$

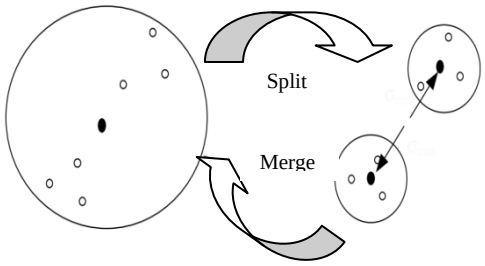

Fig 1: Merge and split.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we perform some experiments to compare the performance of FCM and FISODATA algorithm with IRIS dataset. The experiments are performed using MATLAB.

The Iris flower dataset (Fisher's Iris data set) is a multivariate data set. The data set comprises 50 samples from each of the three species of Iris flowers: SETOSA, VIRGINICA and VERSICOLOR (see Figure.2). Four features were measured from every sample; they are the length and the width of sepal and petal, in centimeters; so each flower is represented by 4 dimensional vectors. Based on the combination of the four features,

Fisher has developed a linear discriminant model to distinguish the species from each other. It is used as a typical test for many classification techniques [11].

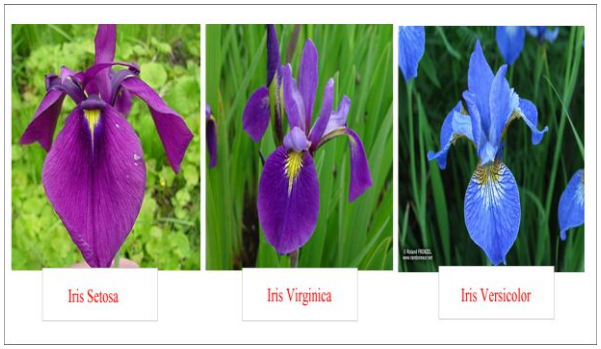

Fig 2: Iris flowers.

4.1. Parameter selection

The weighting exponent m is called the fuzzifier [14]; it has influence on the clustering performance of FCM (see Table1, Table2, Table3). When m=1, the FCM is reduced to the hard c-means, when m tends to infinity $U_{ij} = 1/c$ for all i, j and the sample mean is a unique optimizer ofJ_{FCM}.

Table1. Confusion matrix m=1.2

	C_1	C_2	C_3
C_1	***50***	*0*	*0*
C_2	*0*	***49***	*0*
C_3	*0*	*0*	***51***

Table2. Confusion matrix m=1.4

	C_1	C_2	C_3
C_1	**50**	0	0
C_2	0	**54**	0
C_3	0	0	**46**

Table3. Confusion matrix m=2

	C_1	C_2	C_3
C_1	**50**	0	0
C_2	0	**54**	0
C_3	0	0	**46**

Another parameter which also has an influence on U_{ij} is the cluster number c; FCM assumes that cluster number is known a priori contrary to FISODATA which updates c during the algorithm, that's why FISODATA is more flexible than FCM.

The initialization of the membership function and the centers influence the results. There are two ways to initialize them. The first one is called initialization by gravity centers. It is based on the search of the best centers than we compute membership function. The second one is called initialization by membership function. It is based on a random initialization of the membership function, centers are then computed. The first method gives better results than the second one. Figure 3 shows that time execution of FCM with the initialization by gravity center is less than FCM based on the initialization by membership function. Figure 4 shows the mean square error (MSE) of FCM based on the two initializations. MSE is defined by:

$$\text{MSE} = \sqrt{\|v_c - v_t\|^2} \qquad (13)$$

V_c is the resulting center and V_t is the correct one.

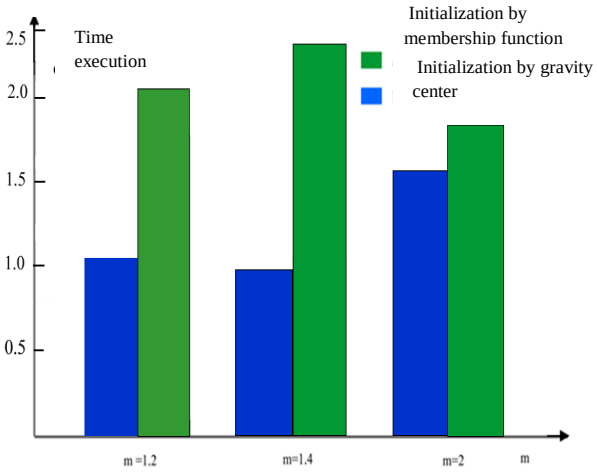

Fig 3: Execution times of both initializations.

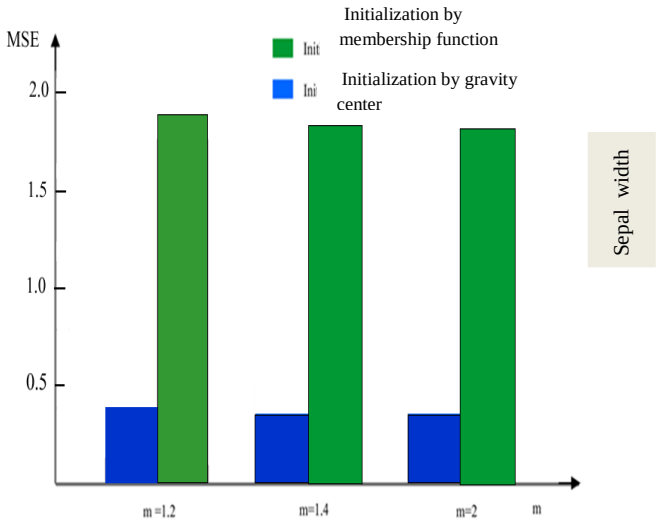

Fig 4: MSE of both initializations

Clustering's result using FISODATA with m=1.2 is shown in Figure 5. Figure 6 shows a comparison between FCM's result and FISODATA's result; FISODATA's results are better than FCM's results. But the time execution of FCM's algorithm is less than the time execution of FISODATA's algorithm, because of the splits and merges steps of the FISODATA's clustering (see Figure 7).

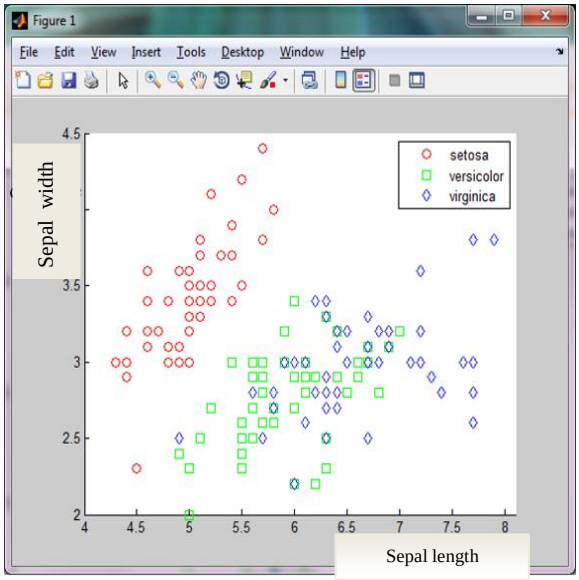

Fig 5: FISODATA's clustering with m=1.2

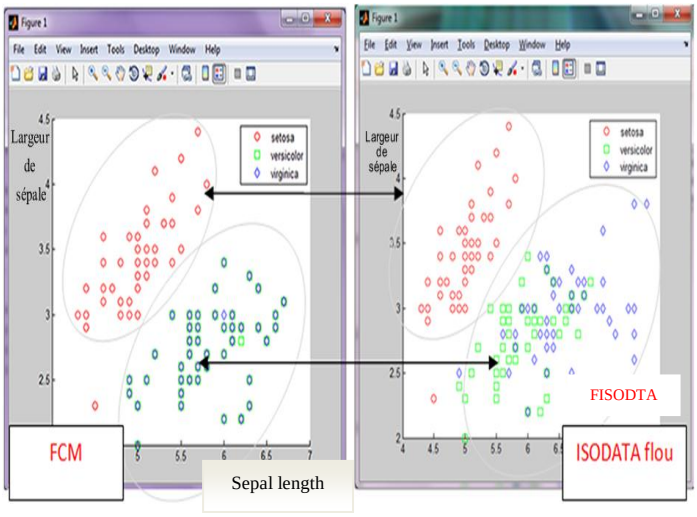

Fig 6: Comparison of FCM's and FISODATA's clustering.

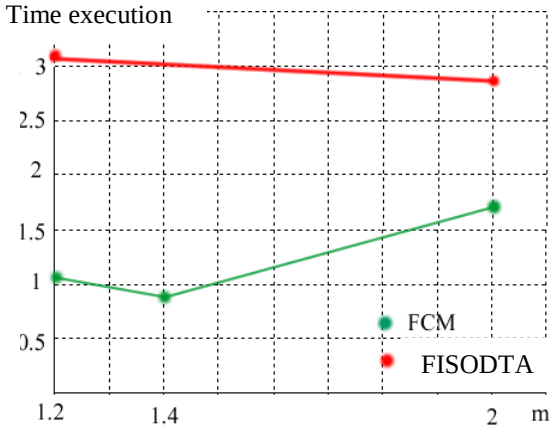

Fig 7: Comparison of FISODATA's and FCM's times execution.

5. CONCLUSION

In this paper we presented a comparison between FCM clustering algorithm and FISODATA clustering algorithm. A comparison of the two algorithms shows that the advantages of FISODATA comparing to FCM algorithm are its self organizing capability, its flexibility in eliminating small clusters, its ability to split dissimilar clusters, and its ability to merge similar clusters. But FISODATA is not perfect; it's slower than FCM because of the splits and merges steps.

6. ACKNOWLEDGMENTS

The authors are grateful to S. Kerdous and W. Boudjelal, without their assistance the simulation experiments would not have been possible. The authors would also to thank A. LAYEB and A. DRAA.

7. REFERENCES

[1] Ameer Ali, M., karmakar, G. C. and Dooley, L. S. 2008. Review on fuzzy clustering algorithms. IETECH journal of advanced computations, VOL.2, NO.3, 169-181. IETECH publications.

[2] Yang, Y. Zheng, C. and Lin, P. 2005. Fuzzy c-means clustering algorithm with a novel penalty term for image segmentation. OPTO - ELECTRONICS. Review 13 (4), 309 – 315.

[3] Ayech, M.W., El- kalti, k. and El Ayeb. B. 2010. Image segmentation based on adaptive Fuzzy –C-M clustering. International conference on Pattern recognition.

[4] Moumen, E – M., Zanaty, E.A., Walaa, M.A-E. and Aly, F. 2007. On cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation. ICIP. 1-4244-1437-7/07/2007 IEEE.

[5] TZafesta, S.G. and Raptis, S. N .2000. Image segmentation via iterative fuzzy clustering based on local space-frequency multi-feature coherence criteria. Journal of intelligent and robotics systems 28: 21-37.

[6] Shanbharkar, S. and Tripude, S.2011. Fuzzy c-means clustering for content based image retrieval system. International conference on advancements in information technology with workshop of ICBMG. IPCSIT, vol.20, Singapore.

[7] Sun, J., Zhang, H. and Yuan, Z. 2009. Fuzzy clustering algorithm based on factor analysis and its application to e-mail filtering. Journal of software, Vol.4, no.1.

[8] Krishnapuram, R. and Keller, K. M.1993. A possobilistic approach to clustering. IEEE transaction on fuzzy systems, vol.1, n°.2.

[9] Yang, M.-S. and Wu, k. L. 2006. Unsupervised Possibilistic clustering**.** Pattern Recognition 39(2006) 5-21. Published by Elsevier.

[10] Saad, M.F. and Alimi, A. M. 2009. Modified fuzzy possibilistic c-means**.** Proceeding of the international Multi conference of Engineers and computer scientists Vol I. IMECS 2009, Hong Kong.

[11] Vanisri, D. and Loganathan, C. 2011. An enhanced Fuzzy Possibilistic C-means with Repulsion and Cluster Validity Index. IJCSNS international Journal of Computer Science and Network Security, VOL.11 No 2.

[12] Dunn, J.C. 1974. A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters**.** J. Cybern. Vol3, no.3, pp32-57.

[13] Ball, G. and Hall, D. 1965. ISODATA, a novel method of data analysis and pattern classification**.** Technical Report, Stanford Research Institute.

[14] Wu, K.L. 2010. Parameter selections of fuzzy C-means based on robust analysis**.** World Academy of science, Engineering and technology 65.