ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 7, July 2017

Iris Data Classification Using Fuzzy Clustering With Varying Parameters

[1] Nisha Singh
[1] Dept Of CSE, RAMA University, Kanpur

***Abstract* - In the field of various real environment, there is problem of clubbing the data according to their behavior or working techniques. Fuzzy clustering can be used where any data belongs to more than one class or bucket formed anywhere. That means the decision to keep them in any bucket is done by applying some similarity measurements. According to this the data points of any data set can belong to more than one class, even having different membership function value to different class. Fuzzy clustering is comprising two very dissimilar data types as fuzzy data and usual (crisp) data. It is a kind of function working on probabilistic mode of evaluating the values. Where the whole process is done without training of values to that system is done. In this paper the data used is iris flower data based problems are used to be clustered with the proper usage of fuzzy clustering model.**

Index Terms: Centroid, Fuzzy Clustering, Fuzzy C- Means Algorithm (FCM), Membership Function.

INTRODUCTION

This document provides the evaluation of Iris data based problem using Fuzzy clustering, an unsupervised classification technique. Where clustering is an important tool for numerous fields as analysis of statistical data, compression of data, data mining, and vector quantization, where to gather the data in class is its aim and data in each class shows higher degree of similarity, while being different to data belonging to another cluster (Jain and Dubes, 1998 ; Sarkar at al., 1997 ; Han and Kamber , 2001). Where the aim is to group the data as their similarity level is maximal within the group's data and similarity is minimal when compared with other group's data.

FUZZY CLUSTERING

Fuzzy clustering is also referred as soft clustering or overlapping logic based clustering. As according to overlapping clustering, a data item may not exclusively belong to only one cluster. As per the value of membership value calculated it may belong to more than one cluster. In the clustering algorithm, called K – Means, it is based on center, very easy, and a speedy algorithm that goals to clubbing of n data into k cluster where every data is belonging to the cluster having nearest mean (MacQueen, 1967). However, in case of real application there are not found the sharp boundaries within a cluster so the data belongs partially to multiple clusters (Jain et al.,1999). In Fuzzy Clusters the membership degree is related to the closeness of data to any cluster center defined over there.

Fuzzy C-Means (FCM) clustering is given by Bezdec(1981) and also it is the most popular fuzzy clustering based algorithm. However, FCM is an algorithm which can be used effectively while selecting the center point at random making iterative process to fall under local optimal solution with ease of accessing. To tackle this problem the evolutionary algorithms used here are as differential evolution based algorithm (DE), genetic algorithm (GA).

FUZZY C - MEANS CLUSTERING

Fuzzy C-means (FCM) is a soft clustering of proposed dataset where each datapoint belongs to a cluster to some degree which is given by membership value. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods[4][5].Fuzzy matrix μ with n rows and c columns are used to describe fuzzy clustering of different objects where n means numbers of data and c for the numbers of clusters. In the matrix μ the element in ith row and jth column, the element is μij .

$$\mu_{ij} J [0,1] , i=1,2,\ldots,n ; j=1,2,\ldots,c \quad (1)$$

$$\sum_{j=1}^{c} \mu_{ij} =1 \quad i=1,2,\ldots,n \quad (2)$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \quad j=1,2,\ldots c \quad (3)$$

The objective function is minimization of fuzzy clustering equation :

$$J_m= \sum_{j=1}^{N} \sum_{i=1}^{C} \mu_{ij}^{m} \| x_i - c_j \|^2 \quad (4)$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering Engineering (IJERCSE)
Vol 4, Issue 7, July 2017

Where, cj is the cluster and m is the fuzzy index governing the influences of membership grades, where m is set to 2.

$$\mu_{ij}=\frac{1}{\Sigma_{k=1}^{C}\left(\frac{||x_i-c_{j||}}{||x_i-c_{k||}}\right)^{\frac{2}{m-1}}} \quad (5)$$

$$cj=(\Sigma_{i=1}^{N}\mu_{ij}^{m}.x_i)/(\Sigma_{i=1}^{N}\mu_{ij}^{m}) \quad (6)$$

Where, μ_{ij} is used to evaluate membership values. And it depends on value of m, high value of m will provide the lower value of μ_{ij} .

From the sample points x_i to the cluster center a_j , the Euclidean distance is measured by the term used here in equation (4) as $\| x_i - a_j \|^2$.

Here in given equation the iteration gets terminated when, $\max_{ij} |\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}| < \epsilon$, where ϵ is a termination criterion between 0 and 1, whereas k are the iteration steps.

Algorithm for the defined strategy to follow is :

1. Initialize $U=[\mu_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$
   $cj=(\Sigma_{i=1}^{N}\mu_{ij}^{m}.x_i)/(\Sigma_{i=1}^{N}\mu_{ij}^{m})$
3. Update $U^{(k)}$, $U^{(k+1)}$
   $\mu_{ij}=\frac{1}{\Sigma_{k=1}^{C}\left(\frac{||x_i-c_{j||}}{||x_i-c_{k||}}\right)^{\frac{2}{m-1}}}$
4. If $\| U^{(k+1)} - U^{(k)} \| < \epsilon$ then STOP ; otherwise return to step 2.

For FCM algorithm value of the time complexity is given by $O(ndc^2i)$ [9][11].

VALIDATION

The validation step is related to the procedure for the verification of fuzzy zone, as it fits best to the whole database. Usually, the cluster validity indexes are calculated in this step measure statistical properties of clustering results, usually the distance within cluster or among clusters. In this step fitting includes other fields also as a fixed number of clusters and the shapes of cluster found.

This study has validated the set of objects via two types of validity indices described as following:

(a)Partition Coefficient (PC): calculates the value of "overlapping" between clusters it is defined by bezdec et al. (1984) as the given equation:

$$PC(c)=\frac{1}{N}\Sigma_{i=1}^{c}\Sigma_{j=1}^{N}(\mu_{ij})^2 \quad (7)$$

Where μij is membership function of data joint j in cluster i.

(b)Classification Entropy (CE): according to the study of cheng et al. (1998), CE measures the fuzziness of the cluster partition only , which is same as measuring the previous coefficient.

$$CE(c)=-\frac{1}{N}\Sigma_{i=1}^{c}\Sigma_{j=1}^{N}\mu_{ij}\log(\mu_{ij}) \quad (8)$$

DATA SET DESCRIPTION:

Iris Data Set: The data set has been taken from UCI Repository of Machine Learning Databases. It has 3 classes, each having 50 instances.Where evry class refers to a type of iris plant. The attribute to be predicted is class of iris plant with 150 instances total and 4 attributes in data set- sepal length in cm, sepal width in cm, petal length in cm, petal width in cm.

Dataset	Size	No. of attributes	Classes
Iris	**150**	**4**	**3**

Table -1 Iris Data Set

RESULTS AND DISCUSSION:

Now on varying the value of exponent from 1.5 till 2.1 by unit increment in the value of exponent following cluster graphs are generated :

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering Engineering (IJERCSE)
Vol 4, Issue 7, July 2017

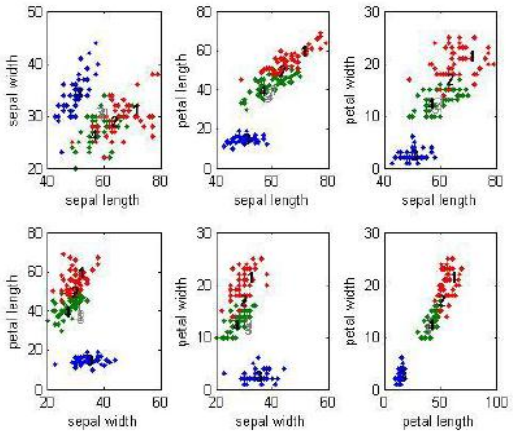

Fig (1) : For the value of exponent 1.5

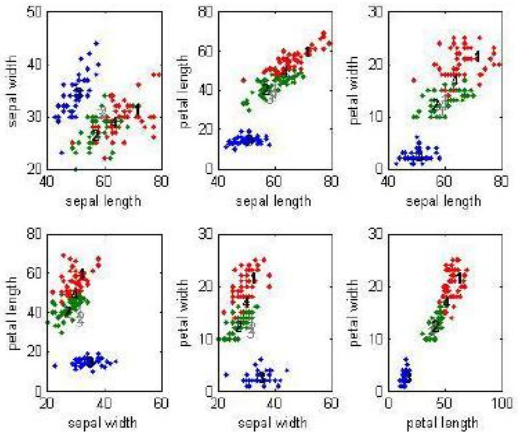

Fig (2) : For the value of exponent 1.6

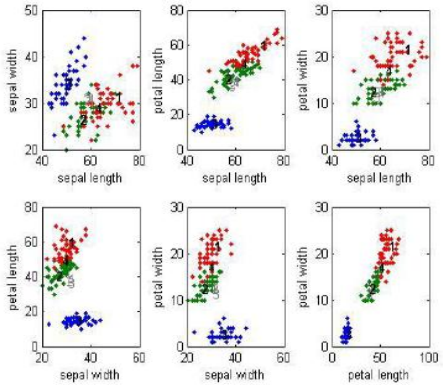

Fig (3) : For the value of exponent 1.7

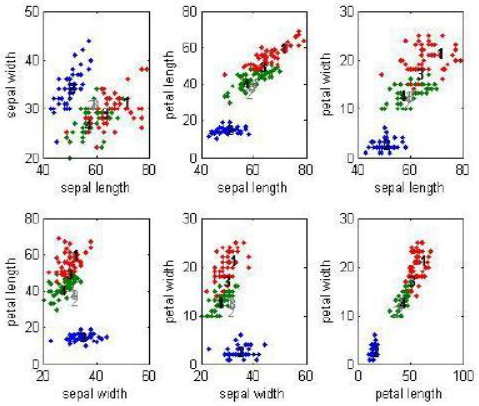

Fig (4) : For the value of exponent 1.8

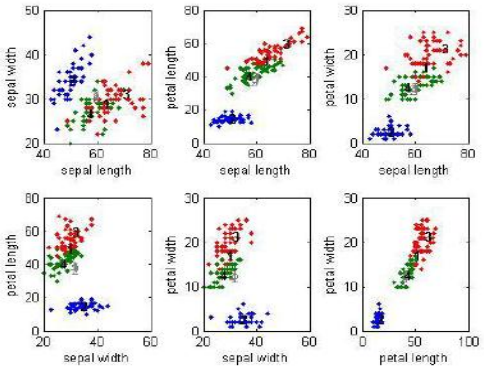

Fig (5) : For the value of exponent 1.9

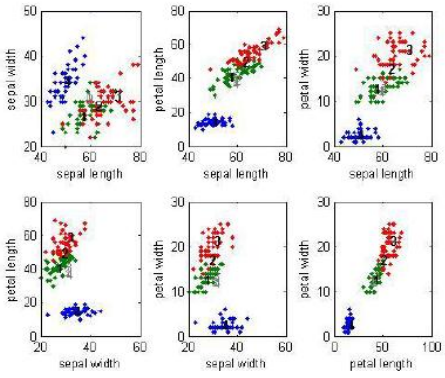

Fig (6) : For the value of exponent 2.0

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering Engineering (IJERCSE)
Vol 4, Issue 7, July 2017

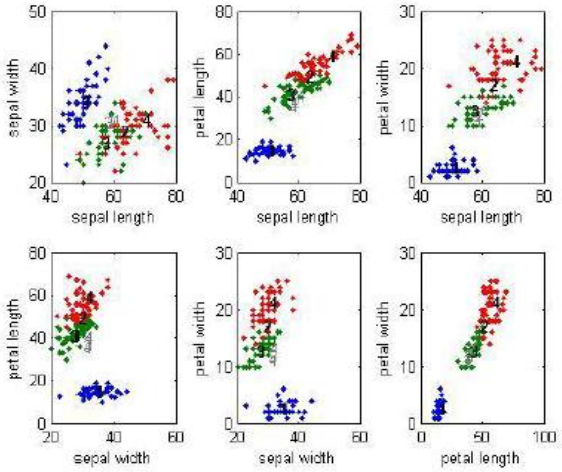

Fig (7) : For the value of exponent 2.1

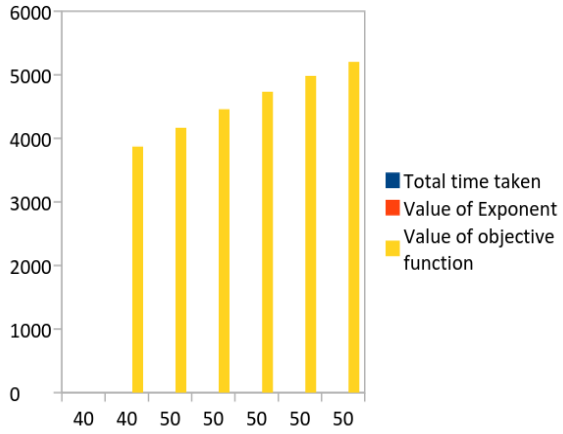

Fig (8): Max. Iteration Count & Value of Objective Function for varying Exponent Value

The overall results are shown below in the Table -2:

Value of Exponent	Max. Iteration count	Total time taken	Value of objective function
1.5	49	1.7188	5384.478684
1.6	47	1.5938	5203.513296
1.7	45	1.0938	4984.149622
1.8	46	1.2031	4733.021835
1.9	45	1.1875	4458.383501
2.0	39	1.0469	4168.707060
2.1	36	1.000	3871.837094

Table -2 Experimental result after taking the different value of exponent

So, It can be clearly declared that result for the value of exponent as 2.1 FCM algorithm can give the better performance in terms of minimum no obtained of maximum iteration count, minimum time and minimum objective function value.

TIME COMPLEXITY :

The time complexity of FCM is O(ndc2i) [9]. Now taking no. of data points as constant and n=150, d=2, i=10 for this. And now on varying no. of clusters we can find a table with a graph where n is no. of data sets, c stands for no. of clusters, d is for dimension, and i stands for iterations.

Time Complexity when no. of cluster varying :

Sr. No.	No. of Cluster	Time complexity
1	1	3000
2	2	12000
3	3	27000
4	4	48000

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering Engineering (IJERCSE)
Vol 4, Issue 7, July 2017

Table -3 Time Complexity when Number of cluster varying

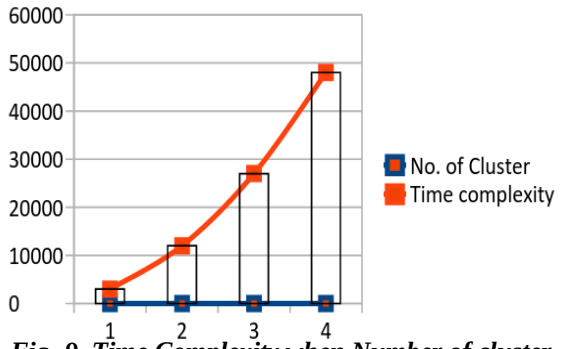

Fig -9 Time Complexity when Number of cluster varying

Now taking no. of clusters as constant, and assuming n=150, d=2, c=2 and varying the no. of Iterations, following table and graph is obtained :

Sr. No.	No. of Iteration	Time complexity
1	05	6000
2	10	12000
3	15	18000
4	20	24000

Table -4 Time Complexity when Number of iteration varying

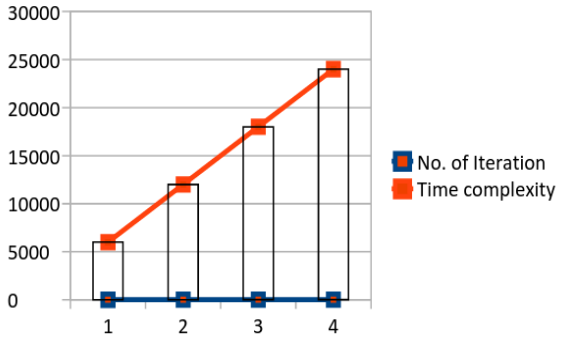

Fig -10 Time Complexity when Number of iteration varying

SPACE COMPLEXITY :

Space complexity of FCM method is O(nd+nc). Now taking no. of data points as constant value, lets assume n=150, d=2 and varying the value of no. of clusters we obtain following table with its corresponding graph.

Sr. No.	No. of Cluster	Space complexity
1	05	450
2	10	600
3	15	750
4	20	900

Table -4 Space Complexity when Number of cluster varying

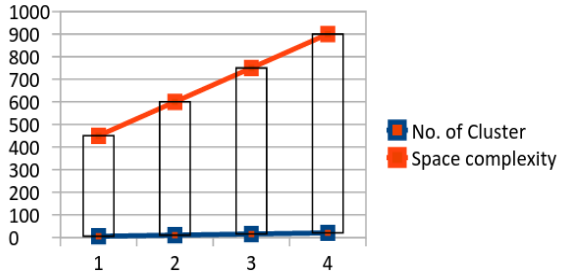

fig -11 Space Complexity when Number of cluster varying

CONCLUSIONS:

In Fuzzy clustering, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in discovering association rules and functional dependencies and image retrieval. The time complexity of the Fuzzy C Mean algorithm is

O(ndc2i). The memory complexity of FCM is O(nd + nc),and the disk input output complexity will be O(ndi) and there is effect in the output also if we vary the value of exponent as 2.0 is standard value assigned to it but in some cases it is not compulsory always to take similar value of exponent. It means we can get effective result on varying the value of exponent in comparison to being stuck to a constant value.

REFERENCES:

[1] Bezdek, J., R. Ehlich, W. Full. 1984 FCM: The fuzzy c-means clustering algorithm. Computers and geosciences. 10(2) , 191-203.

[2] Cheng, H. D., Chen, J.R, Li., J.1998. Thresold selection based on fuzzy c-partition entropy approch. Pattern recognition, 31(7), 857-870.

[3] Frank Hoppner, Fuzzy cluster analysis: methods for classification, data analysis, and image recognition.

[4] S.N. Sivanandam, S. Sumathi, Introduction to fuzzy logic using MATLAB.

[5] J. C. Bezdek, "Pattern recognition with Fuzzy Objective Function Algorithms", Plenum Press,New York, 1981.

[6] J.C. Dunn, "A fuzzy Relative of the ISODATA Process and Its Use in detecting Compact Well-Separated Clusters", Journal of Cybernetics 3:1973,32-57.

[7] Dibya Jyoti Bora and anil Kmumar Gupta, "A comparative study between Fuzzy Clustering Algorithm and Hard clustering Algorithm", InternationalJournal of Computer Trends anmd technology (IJCTT), V10(2), Apr 2014. ISSN:2231-2008.pp. 108-113.

[8]A. Rui and J.M.C. Sousa, "Comparison of fuzzy clustering algorithms for classification", Inetrnational Symposium on evolving Fuzzy Systems, 2006.pp.112-117.

[9] Prodip Hore, Lawrence O. Hall, and Dmitry B. Goldgof "Single Pass Fuzzy C Means", CSEEE, vol.28, 2000.

[10] M. Alata, M. Molhim, and A. Ramini, "Optimizing Fuzzy C Means clustering algorithm using GA", Proceedings of World Academy of Science, Engineering and Technology, vol. 29, 2008.