Fuzzy C-means clustering algorithm

HALIMA S. TALHI
Computer Science. Faculty of Information Technology
Misurata, Libya
Halima.sanoussi@gmail.com

***Abstract* - In this work, we given introduction to the Data Mining (DM) concept which is the most essential step in Knowledge Discovery in Database (KDD) process. Also we reviewed the tasks of DM, in particular the cluster analysis task which is one of the old and well-studied concepts in data mining. The different methods of cluster analysis from the view point of Partitioning/ Hierarchal are presented too. In this paper, the focus on the partitioning C-means method in the sense of fuzzy. Our attention is on the fuzzy C-means algorithm. Fuzzy C-means algorithm is implemented in a system developed in visual basic.net programming language. A number of data sets in the form of experiments test our system. Our study concluded with analysis and discussion of the experiments' result on the bases attention is of several criteria.**

***Index Terms:* C-means, Cluster analysis, Data Mining (DM), Knowledge Discovery in Database (KDD).**

I. Introduction

Advances in computer technology and data acquisition tools and techniques have resulted in the generation of terabytes of data. These huge amounts of data have been stored in database or other repository systems. These massive volumes of data exceeded human and traditional data analysis tools in transforming them into useful knowledge. The implicit knowledge in these volumes of data necessitates the invention of some automatic new data analysis tools and techniques to discover the implicit knowledge in one form or another. These tools and techniques had been implemented in a new emerging field known as Knowledge Discovery in Databases (KDD).

The term KDD also known by other names such as; data mining, knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging Ref. [1]. Many people treats the term Data Mining (DM) as a synonym to KDD while others view DM as one step in the KDD process as depicted in Fig. 1 Ref. [2]. DM is the extraction of useful interesting knowledge from large data sets, databases or other repository system Ref. [3].

According to Ref. [1], the KDD process consists of:

1) Selection of the relevant date for the current mining task and make it ready for analysis.
2) Pre-processing is to clean the data from noise or outliers or irrelevant data.
3) Data mining is to apply intelligent methods in order to extract patterns or regularities from the data.
4) Pattern evaluation step is to identify the truly interesting patterns and/or regularities via some interestingness measures.
5) Knowledge presentation is to present the mined knowledge to the user.

KDD defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data Ref. [4]. The DM field is a confluence of many disciplines such as; database systems, data warehousing, statistics, machine learning, and other areas. Data mining can be applied to a wide variety of problems. Recently, KDD and DM has become one of the most active research areas that have attracted the attention of many researchers. Many successful applications been reported from different sectors such as marketing, finance, banking, manufacturing, security, medicine, multimedia, education telecommunications, etc…Ref. [4].

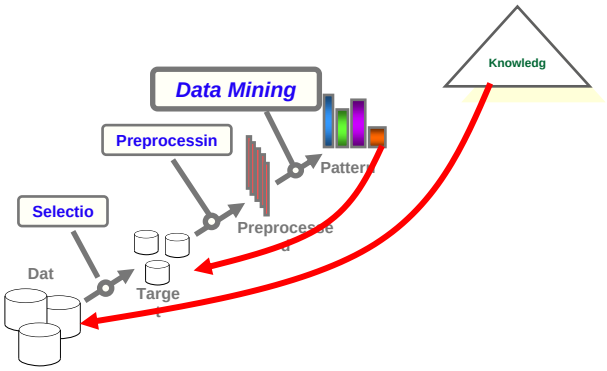

Figure 1. The KDD process.

II. Data mining tasks

Generally, DM has many tasks that are classified into two main categories namely: predictive and descriptive. Predictive mining tasks are those processes used to formulate a model described by the current data set and then the model is used to make prediction for new data objects. Classification and Prediction are some examples of predictive mining tasks. Descriptive mining tasks are those characterize the general properties of the current data set. Characterization, Discrimination, Cluster Analysis, Outlier Analysis and Association Analysis are examples of descriptive mining tasks.

III. Cluster analysis

Cluster analysis is the process of partitioning data objects into groups or clusters, so that objects in one group are as similar to each other as possible, and objects in different clusters are as dissimilar as possible. Cluster analysis has been in use since a long time ago in many fields such as: Biology, Chemistry, Botany and others. In the past, clusterings were conducted in a subjective manner and researchers relied on their perception and judgment in interpreting the results, this was done quite easily done due to the low dimensionality (at most 2-D) and size of the data.

Nowadays, old clustering methods are not suitable any more due to the increase of extent and/or dimensionality of the data to be clustered. Automatic classification of data is a very young scientific discipline in vigorous development Ref. [5]. Automatic classification has establishing itself as an independent scientific discipline due to the thousands of articles scattered over many periodicals such as; journals of statistics, biology, psychometrics, computer science, and marketing. Since the mid of the 80's cluster analysis has been established as an independent discipline where some scientific periodicals such as; the Journal of Classification and the International Federation of Classification Societies are dedicated for such discipline. Computer scientists consider cluster analysis as a branch of pattern recognition and artificial intelligence.

A Cluster analysis methods

Generally speaking, most of the clustering methods can be classified into one of two categories; partitioning methods or hierarchical methods.

B Partitioning clustering method

A partitioning clustering method constructs k mutual-exclusive clusters out of n data objects. The partitioning method must satisfy the following conditions:

1) Each cluster must contain at least one data object.
2) Each data object must belong to only one cluster.

The first condition implies that there are as many clusters as there are objects: $k \leq n$. The second condition states that two different clusters cannot have any objects in common.

Partitioning algorithms creates clusters or groups of data objects in such a way that objects within a cluster have high similarity (Intra-cluster) and objects in different clusters have high dissimilarity (Inter-cluster) as depicted in Fig. 2.

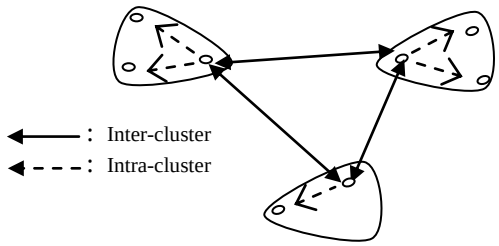

Figure 2. Intra-cluster and Inter-cluster similarities.

C Hierarchical clustering methods

Hierarchical methods are unlike partitioning methods in the sense that all of the values of k (number of clusters) are present in the constructed dendrogram (tree like hierarchy). Starting from the partition of k = 1 (all objects are together in one cluster) and the partition with k = n (singleton clusters) there are, all values of k = 2, 3, ..., n – 1.

Hierarchical clustering methods construct a tree-like hierarchy of data objects. There are two types of hierarchical techniques namely; agglomerative and divisive as depicted in Fig. 3. These two techniques construct the tree-like hierarchy in the opposite directions.

1) The agglomerative technique
   Starting with n singleton clusters and in each step two of the clusters will be combined into one cluster until all objects are merged into one cluster with n data objects.
2) The divisive technique
   Starting with one cluster with n data objects and in each step, one of the clusters will be split up into two clusters, until there are n singleton clusters.

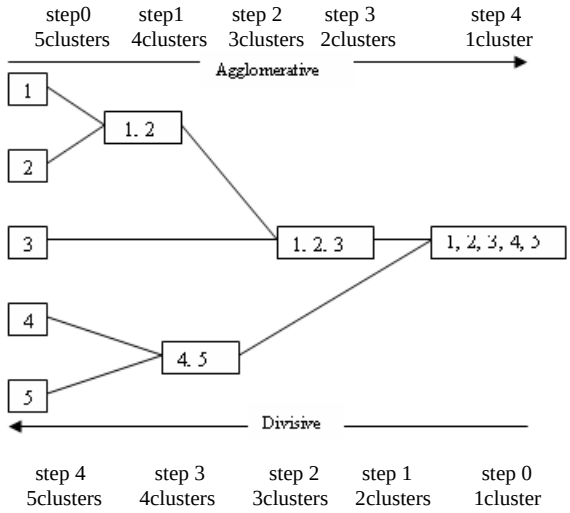

Figure 3. Agglomerative and divisive techniques.

IV. K-means methods

Here, we give some details of the K-means methods that are of relocation partitioning type. This type of clustering depends on the calculation of the mean of each cluster in the process of assigning objects to clusters. The purpose of focusing on the K-means methods is due to the main concern of this work. The K-means methods can be of hard clustering type (called Crisp C-means) and can be of soft clustering type (called Fuzzy C-means).

The *Fuzzy* C-means clustering algorithm is based on the concept of the fuzzy sets approach. The *Fuzzy* C-means clustering algorithm aims to find fuzzy partitioning of a given data set, where each data object

belongs to several clusters with the degree of belongingness between 0 and 1.

According to Ref. [6, 7], the Fuzzy C-means clustering algorithm was proposed by Bezdek in 1981. The steps of the Fuzzy C-means clustering algorithm are as the follows:

Preparations:

1) Choose the number of clusters k where $2 \le k \le n$.
2) Choose a value for the exponent weight m', where ($m' \in (1,\infty)$) and the default value of m' is 2.
3) Choose a value for the stopping criterion ε, where the default value is 0.01.
4) Choose the distance function to be used. The default is the Euclidean distance.
5) Initialize the membership matrix $\underline{U}$ with random values $\underline{U} = \underline{U}^{(0)}$; so that:

$$\sum_{i=1}^{c} \mu_{ib} = 1 \tag{1}$$

$$0 < \sum_{b=1}^{n} \mu_{ib} < n \tag{2}$$

$$0 \le \mu_{ib} \le 1 \tag{3}$$

Where $b = 1, 2 \ldots n$ and $i = 1, 2 \ldots c$, n represents the number of objects and c represents the number of clusters. Therefore μ_{ib} is membership of object b in clusters i.

Each step in this algorithm will be labelled with r, where r is the number of iterations and take the values of: 0, 1, 2, …

Step 1: Calculate the fuzzy clusters centers using:

$$v_{ij} = \frac{\sum_{b=1}^{n} \mu_{ib}^{m'} x_{bj}}{\sum_{b=1}^{n} \mu_{ib}^{m'}}$$

Where $j = 1, 2, \ldots m$ and m represents the number of properties. X_{bj} is the value of the b^{th} object of the j^{th} property.

Step 2: Find the distance between each object and cluster centers by the use of:

$$d_{ib} = d(x_b - v_i) = \left[\sum_{j=1}^{m} (x_{bj} - v_{ij})^2 \right]^{1/2}$$

Step 3: Update the membership matrix for the r^{th} step $\underline{U}^{(r)}$ by the use of:

$$\mu_{ib}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ib}^{(r)}}{d_{jb}^{(r)}} \right)^{2/(m'-1)} \right]^{-1}$$

Where i, j are any two different clusters.

Step 4: Check if $\max \left| \underline{U}^{(r+1)} - \underline{U}^{(r)} \right| \le \varepsilon$ then

Stop (the resulted clustering is good enough), otherwise set $r = r + 1$ and return to step 1.

The steps of the standard Fuzzy C-means clustering algorithm are as the follows:

Input: A data set, k: number of clusters, n: number of data objects, $2 \le k \le n$.

Output: Clusters

```
While (true) do:   [Outer loop].
    if  P < 2 then  (P is number of iteration)   [Check].
       Ran_Create ( )   [Initialize Matrix].

         for I⟵ 1 to k [Loop].
              Randomize ( )  (choose random membership
                              coefficients).
       End for

    Else
          Vectors ( )  [Calculate the fuzzy clusters centers].
        Dis ( ) [find the distance]
     f clustering ( )  [update the membership matrix].
           if max < 0.01 (true) then [Test fuzzy clustering
                                      quality].
         Exit while (find best fuzzy clustering this data set).
       Else   [Repeat].

         Set P⟵ P+1   [Iterations counter].
       End if
    End if
End while
```

V. Design and implementation

Here, we will demonstrate the design and implementation of our Fuzzy Clustering System (FCS). We have used the Unified Modeling Language (UML) Ref. [8] in the analysis and design phases, and Visual Basic.NET 2005 programming language in the actual implementation of the system.

Here, we will demonstrate only the activity diagram, class diagram and the sequence diagram, of the FCS system, due to their importance and capabilities to give a clear view of the system. Fig. 4 depicts the activity diagram for the FCS system that shows the system processes and helps to define the sequence of tasks, the needed conditions and the concurrency of the system.

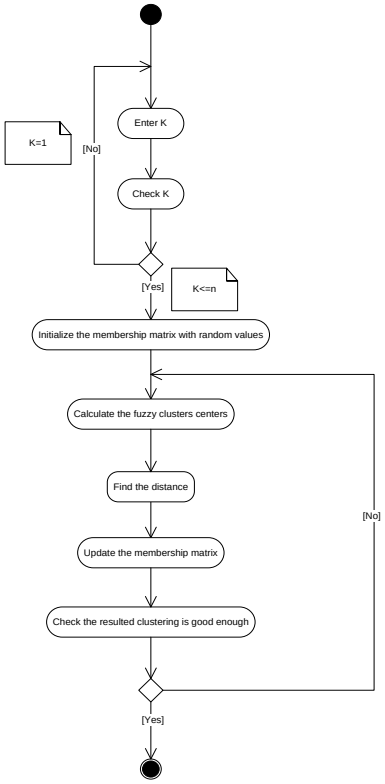

Figure 4. Activity diagram for the FCS system.

The class diagram gives an overview of a system via describing its attributes and operations of each of the classes, and the various types of static relationships among those classes. Fig. 5 depicts the class diagram of the FCS system, which includes the classes and the relationships between them.

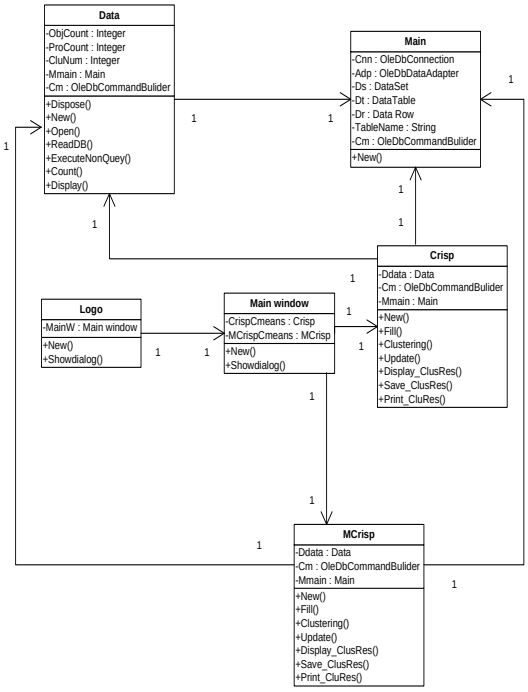

Figure 5. The class diagram of the FCS subsystem.

There is a method, which is common to all classes; this method is New (constructor), which initializes the objects' parameters when it is created. The Data Class is to establish the appropriate connection with the database, read its content, to execute SQL commands and display the data. The Main window Class is to show the system to execute. The Logo Class is the starting point to perform the system. The Fuzzy Class has methods, to read the actual data from the database and update it after some operations have been carried out. And to display the results, save it and print after performance of the sought clustering.

The sequence diagram describes the behaviour of the system and depicts the messages passed between the objects during the period of the system execution. Fig. 6 depicts the sequence diagram for the FCS subsystem.

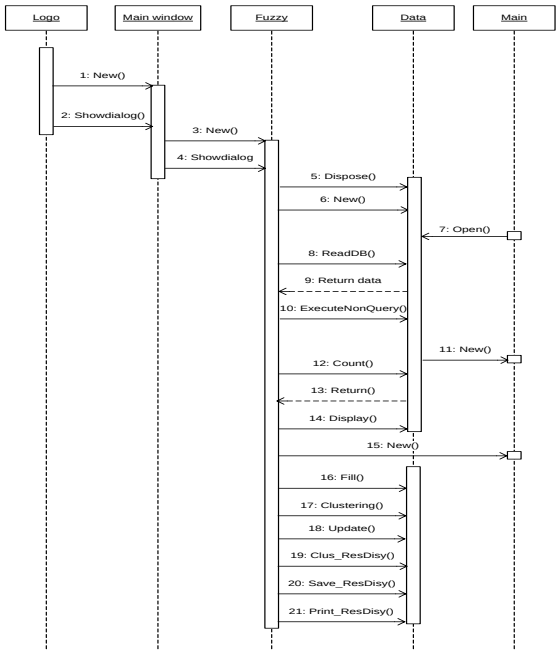

Figure.6: The sequence diagram for the FCS sub-system.

VI. Experiments and results

Here, we will demonstrate the obtained results from two experiments on the FCS system using two different data sets. The interpretation and comparisons of results are presented. Here we had conducted datasets that are used in testing data mining systems and algorithms are used to test our system performance.

A First experiment

The experiment has been conducted on a database Ref. [10] that consists of 4 objects each of which is described by two variables as depicted in table 1. This experiment is set to form two clusters (k = 2).

TABLE 1. Data objects for first FCS system experiment.

SN	Object name	X	Y
1	A	1.0	3.0
2	B	1.5	3.2
3	C	1.3	2.8
4	D	3.0	1.0

Since the FCS system initializes the membership coefficients with random values, so this experiment has been conducted ten times to get better results. Table 2 lists the final obtained results of those 10 runs. Table 3 and Fig 7 show details of analyzing the performance of the results for the first run of table 2.

Interpretation of the results as follows:

1) From table 2 all ten runs gave the same results except fifth one.
2) Table 3 and Fig 7 show the details of the 2 passes of the first run, which indicate the reduction in the error value to single the occurrence of the stopping criterion.

TABLE 2: The FCS system results for first experiment.

Test runs	Pass number	Error value
1	2	0.00
2	2	0.00
3	2	0.00
4	2	0.00
5	2	0.00
6	2	0.00
7	2	0.00
8	2	0.00
9	2	0.00
10	2	0.00

TABLE 3: Performance of the FCS system for the first run of table-2.

Pass No	Error value
1	0.014
2	0.000

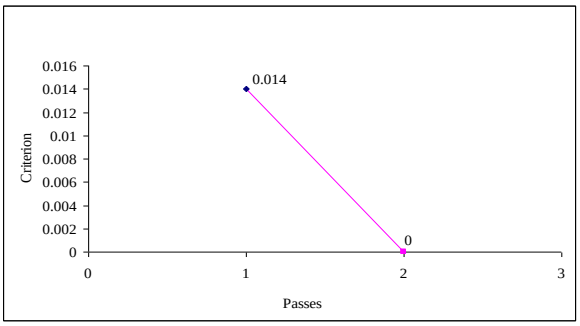

Figure 7. Analyzing the performance of the FCS system in table 3.

Comparison of results

The original results from Ref. [7] are depicted in table 4 where there are 4 objects belongs to two clusters with the degree of membership.

In comparing our ten runs with original results, we have found that the results are identical.

TABLE 4. The original results of FCS system for first experiment.

	C1	C2
A	0.991	0.008
B	0.987	0.013
C	0.993	0.006
D	0.000	1.000

B Second experiment

This experiment has been conducted on the database Ref. [5] that consists of 22 objects each of which is described by two variables as depicted in table 5.

TABLE 5. The original results of FCS system for second experiment.

Object name	C_1	C_2	C_3
O_1	0.87	0.06	0.07
O_2	0.88	0.05	0.07
O_3	0.93	0.03	0.04
O_4	0.86	0.06	0.08
O_5	0.87	0.06	0.07
O_6	0.42	0.35	0.23
O_7	0.08	0.82	0.10
O_8	0.06	0.87	0.07
O_9	0.06	0.86	0.08
O_{10}	0.06	0.87	0.07
O_{11}	0.06	0.86	0.08
O_{12}	0.07	0.84	0.09
O_{13}	0.36	0.27	0.37
O_{14}	0.12	0.08	0.80
O_{15}	0.08	0.07	0.85
O_{16}	0.10	0.10	0.80
O_{17}	0.08	0.06	0.86
O_{18}	0.04	0.04	0.92
O_{19}	0.07	0.07	0.86
O_{20}	0.10	0.08	0.82
O_{21}	0.07	0.06	0.87
O_{22}	0.09	0.09	0.82

This experiment is set to form three clusters (k = 3). We have run the experiment ten runs to find best possible results. Table 6 lists the final obtained results of those 10 runs. Table 7 and Fig 8 show details of analyzing the performance of the results for the first run of table 6.

Interpretation of the results is as follows:

1) From table 6, we can see that all the ten runs gave different results.
2) In table 7 and Fig 8 show the details of the 2 passes of the first run, which indicate the reduction in the error value to single the occurrence of the stopping criterion.

TABLE 6: The FCS system results for second experiment.

Test runs	Pass number	Error value
1	2	0.00
2	2	0.00
3	2	0.00
4	2	0.00
5	2	0.00
6	2	0.00
7	2	0.00
8	2	0.00
9	2	0.00
10	2	0.00

Table 7. Performance of the FCS system for the first run of table 6.

Pass No	Error value
Pass 1	0.686
Pass 2	0.000

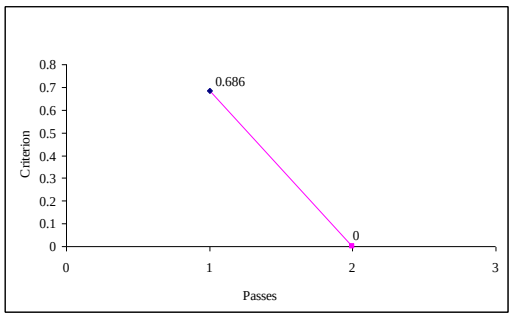

Figure 8. Analyzing the performance of the FCS system in table 6.

Comparison of results

The original results from Ref. [5] are depicted in table 5 where there are 22 objects belongs to three clusters with the degree of membership.

In comparison with our results of the ten runs we have found that the first run had the highest clustering value which is nearly identical with the original results.

VII. Conclusion

The objectives of this work were set to implement a fuzzy version of the C-means algorithm a system developed in visual basic.net programming language and to test the fuzzy version by two databases.

The results we had obtained from the experiments are summarized in the following:

- The FCS system experiments results were identical to the original results for the first data set and nearly identical to the original results for the second data set.

From the humble experience gained during the time of carrying out this work, the author would like to make the following recommendations for further research:

1) More experiments with different data sizes are needed because of the nature of cluster analysis.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Canada: Morgan Kaufmann publishers, 2000, ch. 1, pp. 6 – 21.

[2] U. Fayyad, G. Piatetsky-shapiro and P. Smyth, *From Data Mining to Knowledge Discovery in data base*, American, Association for Artificial Intelligence, 1996, pp. 1 – 54.

[3] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining*, Cambridge, Massachusetts London England: Massachusetts Institute of Technology Press, 2001, pp. 2 – 21.

[4] S. Mitra, and T. Acharya, *Data Mining Multimedia*, New Jersey: John Wiley & Sons, Inc, 2003, pp. 1 – 28.

[5] L. Kaufman and P. Rousseeum, *Finding Groups in Data*, United States of America: John Wiley & Sons, Inc, 1990, pp. 1 – 66.

[6] S. Miyamoto, H. Ichihashi, and K. Honda, *Algorithms for Fuzzy Clustering Methods*, Verlag Berlin Heidelberg: Springer, 2008, pp. 16 – 39.

[7] T. Ross, *Fuzzy logic engineering applications*, New York: McGraw-Hill, 1995, pp. 371 – 410.

[8] T. Weilkiens, *Systems engineering with SysML/UML: modeling, analysis, design*, United States of America: Morgan Kaufmann Publishers, 2007, pp. 1– 320.

Biographies

Halima Sanoussi Talhi was born in Benghazi /Libya. She received Bsc degree in Computer Science /Hardware from University of Benghazi, in 1997. She got Msc degree in Artificial Intelligence/ Specialization Cluster analysis from Academy of Graduate Studies, a branch of Benghazi in 1/1/2009. She is currently a collaborator lecturer in Faculty of Information Technology at Misurata University/Libya. Her research field is Artificial Intelligence /Data Mining.