Refinement of K-Means and Fuzzy C-Means

A. Banumathi
Assistant Professor
Department of Computer Science
Government Arts College,
Karur – 5
Tamilnadu, India

A. Pethalakshmi
Head & Associate Professor
Department of Computer Science
MVM Government Arts College for Women,
Dindigul.
Tamilnadu, India

ABSTRACT
Clustering is widely used technique in data mining application for discovering patterns in large data set. In this paper the K-Means and Fuzzy C-Means algorithm is analyzed and found that quality of the resultant cluster is based on the initial seeds where it is selected either sequentially or randomly. For real time large database it's difficult to predict the number of cluster and initial seeds accurately. In order overcome this drawback we propose two new algorithms Unique Clustering through Affinity Measure(UCAM) and Fuzzy-UCAM clustering algorithm. Both UCAM and Fuzzy-UCAM clustering algorithms works without fixing initial seeds, number of resultant cluster to be obtained. Unique clustering is obtained with the help of affinity measures.

Keywords
Cluster, K-Means, UCAM, Fuzzy C-Means, Fuzzy-UCAM

1. INTRODUCTION
Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. The most popular clustering algorithm used in several field is K-Means since it is very simple and fast and efficient. K-means is developed by Mac Queen[3]. The K-Means algorithm is effective in producing cluster for many practical applications. But the computational complexity of the original K-Means algorithm is very high, especially for large datasets. The K-Means algorithm is a partition clustering method that separates data into K groups. Main drawback of this algorithm is that of a priori fixation of number of clusters and seeds.

Unique Clustering with Affinity Measures (UCAM) clustering algorithm which starts its computation without representing the number of clusters and the initial seeds. It divides the dataset into some number of clusters with the help of threshold value. The uniqueness of the cluster is based on the threshold value. More unique cluster is obtained when the threshold values is smaller. Fuzzy-UCAM uses UCAM clustering algorithm for initial clustering and then the membership matrix is calculated.

2. DATA MINING AND CLUSTERING
Data mining is the process of autonomously extracting useful information or knowledge from large data stores or sets . It involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining consists of more process than collecting and managing data, it also includes analysis and prediction[13]. Data mining is popularly known as knowledge discovery. Fig 1 shows the concept of data mining, which involves three steps:

1. Capturing and storing the data.
2. Converting the raw data into information.
3. Converting the information into knowledge.

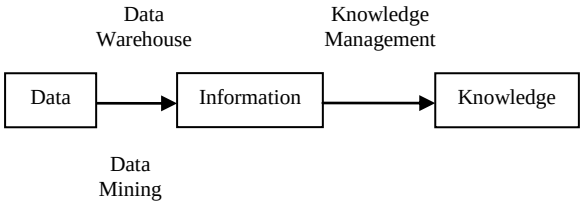

Fig 1: Data mining process

Data in this context comprises all the raw material an institution collects via normal operation. Capturing and storing the data is the first phase that is the process of applying mathematical and statistical formulas to mine the data warehouse. Mining the collected raw data from the entire institution may provide new information. Converting the raw information into information is the second step of data mining and from the information knowledge is discovered.

Clustering is a widely used technique in data mining application for discovering patterns in large dataset. The aim of cluster analysis is exploratory, to find if data naturally falls into meaningful groups with small within-group variations and large between-group variations. Often we may not have a hypothesis that we are trying to test. The aim is to find any interesting grouping of the data. It is possible to define cluster analysis as an optimization problem in which a given function consisting of within cluster similarity and between clusters dissimilarity needs to be optimized. This function can be difficult to define and the optimization of any such function is a challenging task.

The traditional K-Means algorithm is analyzed and found that quality of the resultant cluster is based on the initial seeds but it is difficult to predict the number of cluster and initial seeds accurately. This drawback is rectified through UCAM (Unique Clustering with Affinity Measure) algorithm for clustering which works without giving initial seed and number of clusters to be obtained. Fuzzy clustering concept is appended with UCAM to form Fuzzy-UCAM clustering algorithm. Both UCAM and Fuzzy-UCAM works only for numerical attributes and unique clustering is obtained with the help of affinity measures.

3. K-MEAN CLUSTERING

The main objective in cluster analysis is to group object that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters[3]. One of the most popular clustering methods is K-Means clusters algorithm. It is classifies objects to pre-defined number of clusters, which is given by the user (assume K clusters). The idea is to choose random cluster centers, one for each cluster[14]. These centers are preferred to be as far as possible from each other. In this algorithm Euclidean distance measure is used between two multidimensional data points

$$X = (x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_m)$$

$$Y = (y_1, y_2, y_3, \ldots\ldots\ldots\ldots y_m)$$

The Euclidean distance measure between the above points x and y are described as follows:

$$D(X, Y) = (\sum (x_i - y_i)^2)^{1/2}$$

The K-Means method aims to minimize the sum of squared distances between all points and the cluster centre [8]. This procedure consists of the following steps, as described below

Algorithm1: K-Means clustering algorithm

Input: $D = \{d_1, d_2, d_3, \ldots, d_n\}$ // Set of n data points.

K - Number of desired clusters

Output: A set of K clusters.

K-Means algorithm Steps:

1. Select the number of clusters. Let this number be k.
2. Pick k seeds as centroids of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.
3. Compute the Euclidean distance of each object in the dataset from each of the centroids.
4. Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.
5. Compute the centroids of the clusters by computing the means of the attribute values if the objects in each cluster.
6. Check if the stopping criterion has been met (e.g. the cluster membership is unchanged). If yes, go to step 7. If not go to step 3.
7. [Optional] One may decide to stop at this stage or to split a cluster or combine two clusters heuristically until a stopping criterion is met.

Though the K-Means algorithm is simple, but it has some drawbacks in its quality of the final clustering, since it is highly depends on the initial centroids. K-Means algorithm is implemented in a very small sample data with ten student's information. The process of K-Means clustering is initiated with three initial seeds, which results with three clusters as notated below

$C_1 = \{ S_1, S_9 \}$

$C_2 = \{S_2, S_5, S_6, S_{10}\}$

$C_3 = \{S_3, S_4, S_7, S_8 \}$

Where S_1, $S2 \ldots S_{10}$ Student's details which considers only numeric attributes. In K-Means the initial seeds are randomly selected and hence result of two executions on the same data set will not get the same result unless the initial seeds are same. The main drawback in K-Means is that initial seeds and number of cluster should be defined though it is difficult to predict in the early stage.

4. UCAM CLUSTERING

In cluster analysis, one does not know what classes or clusters exist and the problem to be solved is to group the given data into meaningful clusters. Here on the same motive UCAM algorithm is developed. UCAM algorithm is a clustering algorithm basically for numeric data's. It mainly focuses on the drawback of K-Means clustering algorithm. UCAM algorithm is implemented with the help of affinity measure for clustering. The process of clustering in UCAM is initiated without any centorid and number of clusters that is to be produced. UCAM sets the threshold value for making unique clusters. The step by step procedure for UCAM are given below

Algorithm 2: The **UCAM** algorithm

Input: $D = \{d_1, d_2, d_3 \ldots d_n\}$ // Set of n data points.

S – Threshold value.

Output: Clusters. Number of cluster depends on affinity measure.

UCAM Algorithm Steps:

1. Set the threshold value T.
2. Create new cluster structure if it is the first tuple of the dataset.
3. If it is not first tuple compute similarity measure with existing clusters.
4. Get the minimum value of computed similarity S.
5. Get the cluster index of Ci which corresponds to S.
6. If S<=T, then add current tuple to Ci.
7. If S>T, create new cluster.
8. Continue the process until the last tuple of the dataset.

Implementing UCAM algorithm with the sample data implemented in K-Means. The process is initiated with threshold value T and results with following clusters as shown below

$C1 = \{S_1, S_3, S_7\}$

$C2 = \{S_2, S_5, S_6\}$

$C3 = \{S_4, S_8\}$

$C4 = \{S_9\}$

$C5 = \{S_{10}\}$

Uniqueness of the cluster is depends on the initial setting of the threshold value. If the threshold value increases number of cluster decreases. In UCAM there is no initial prediction on number of resultant cluster. Here, in this algorithm resultant cluster purely based on the affinity measure.

In the above study of K-Means clustering algorithm results with three clusters where low marks and high marks are found in all clusters, since the initial seeds do not have any seeds with the marks above 90. Hence if the initial seeds not defined properly then the result won't be unique and more over it has been constrained that it should have only three clusters.

The UCAM clustering algorithm is initiated with the threshold alone which produces unique result with five clusters.

C_1 → Cluster with medium marks.

C_2 → Cluster with high marks.

C_3 → Cluster with low marks.

C4 = { S_9}

C5 = {S_{10} }

S_9 and S_{10} are found to be having peculiar characteristics for the given threshold value. These two objects have major dissimilarity with the existing clusters and hence it cannot merge with other clusters. By increasing the threshold value it can be merged with other cluster but it reduces the cluster uniqueness and hence it proves that UCAM clustering algorithm has the flexibility of obtaining both approximate clustering and unique clustering.

5. FUZZY C-MEANS (FCM)

The fuzzy c-means clustering algorithm [11] is a variation of the popular k-means clustering algorithm, in which a degree of membership of clusters is incorporated for each data point. The centroids of the clusters are computed based on the degree of memberships as well as data points. The random initialization of memberships of instances used in both traditional fuzzy c-means and k-means algorithms lead to the inability to produce consistent clustering results and often result in undesirable clustering results[9]. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center.

One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm (Bezdek 1981). The FCM algorithm attempts to partition a finite collection of n elements $X = \{x_1,....,x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres $C=\{c_1,...,c_c\}$ and a partition matrix $U=u_{ij} \in [0,1] i=1,...,n,\ j=1,...,c$ where each element u_{ij} tells the degree to which element x_i belongs to cluster c_j. Like the k-means algorithm, the FCM aims to minimize an objective function. The standard function is:

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(center_{k,} x}{d(center_{j,} x} \right)^{2/(m-1)}} \quad (1)$$

This differs from the k-means objective function by the addition of the membership values uij and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large m results in smaller memberships u_{ij} converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2. The basic FCM Algorithm, given n data points (x_1,..., x_n) to be clustered, a number of c cluster with (c1,...,cc) the center of the clusters, and m the level of cluster fuzziness within the cluster.

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. Any point x has a set of coefficients giving the degree of being in the *k*th cluster $w_x(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)} \quad (2)$$

The degree of belonging $w_k(x)$, is related inversely to the distance from x to cluster centre as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest centre. The fuzzy c-means algorithm is very similar to the k-means algorithm:

Algorithm 3: Fuzzy C-Means clustering algorithm

Input: $D = \{d_1, d_2, d_3, ..., d_n\}$ // Set of n data points.

C - Number of desired clusters

Output: A set of C clusters, degree of membership matrix

Fuzzy C-Means algorithm Steps:

1. Choose a C number of clusters.
2. Assign randomly to each point coefficients for being in the clusters.
3. Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ϵ , the given sensitivity threshold):
4. Computer the centroid for each cluster, using the formula (2)
5. For each point , computer its coefficients of being in the clusters, using the formula (1).

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights. The expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Fuzzy C-Means is implemented on same sample data as in K-Means which yields the following membership matrix as shown below

Table 1. Fuzzy C-Means membership matrix

	C1	C2	C3
S1	0.862 *	0.031	0.106
S2	0.343	0.597 *	0.059
S3	0.415	0.076	0.508 *
S4	0.085	0.026	0.888 *
S5	0.116	0.846 *	0.037
S6	0.188	0.742 *	0.069
S7	0.323	0.047	0.629 *
S8	0.099	0.032	0.869 *
S9	0.853 *	0.098	0.047
S10	0.334	0.509 *	0.156

Table.1 indicates the degree of attachment to each cluster by a particular object. Each cluster contains only the stared cells where remaining cells indicates the degree of attachment of the particular object to the corresponding cluster. The classification of the object S_2, S_3, $S_7$and in S_{10} falls into particular cluster by the minor difference in their distance. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improved the accuracy of clustering under noise.

6. FUZZY-UCAM

The fuzzy-UCAM clustering algorithm is a enhanced view of UCAM clustering algorithm, in which a degree of membership of clusters is incorporated for each data point. The centroids of the clusters are computed based on the members of the cluster. The random initialization of the process of traditional fuzzy c-means algorithms leads to cluster error and affects the uniqueness of the cluster. Fuzzy-UCAM algorithm works to rectify the cluster error and increase the uniqueness of Fuzzy C-Means through affinity measure. The Fuzzy-UCAM algorithm is outlined below

Algorithm 4: The **Fuzzy-UCAM** algorithm

Input: D = {d_1, d_2, d_3... d_n } // Set of n data points.

S – Threshold value.

Output: Resultant Clusters, Degree of membership matrix

Fuzzy-UCAM Algorithm Steps:

1. Set the threshold value T.
2. Create new cluster structure if it is the first tuple of the dataset.
3. If it is not first tuple compute similarity measure with existing clusters.
4. Get the minimum value of computed similarity S.
5. Get the cluster index of Ci which corresponds to S.
6. If S<=T, then add current tuple to Ci.
7. If S>T, create new cluster.
8. Continue the process until the last tuple of the dataset.
9. Compute membership matrix for all data points in the resultant cluster using the formula (1).

Fuzzy-UCAM algorithm results with unique clusters which are free from cluster error. The number of resultant cluster is depends up on the threshold value, if the threshold value increases then the number of resultant clusters decreases and on decreasing, the number of resultant cluster increases. Fuzzy-UCAM is implemented on the same sample data as in K-Means which yields the following membership matrix as shown below

Table 2. Fuzzy-UCAM membership matrixes

	C1	C2	C3	C4	C5
S1	0.722 *	0.027	0.042	0.180	0.026
S2	0.073	0.478 *	0.084	0.291	0.073
S3	0.888 *	0.012	0.039	0.043	0.015
S4	0.014	0.017	0.940 *	0.041	0.018
S5	0.005	0.938 *	0.002	0.047	0.006
S6	0.033	0.827 *	0.016	0.085	0.037
S7	0.779 *	0.020	0.091	0.086	0.021
S8	0.012	0.019	0.978 *	0.004	0.021
S9	0.000	0.000	0.000	1.000 *	0.000
S10	0.000	0.000	0.000	0.000	1.000 *

Fuzzy-UCAM clustering algorithm results with the above indicated membership matrix. Table.2 indicates the degree of attachment to each cluster by a particular object. Consider column representation in which each cluster contains only the stared values and where the remaining values indicates the degree of attachment of other objects to the corresponding cluster. On classification of one particular object to a particular cluster has the higher degree of membership and least significant value towards all other clusters.

7. COMPARATIVE ANALYSIS

The comparative study of K-Means, FCM, UCAM and Fuzzy-UCAM clustering are shown in the following table.

Table 2. Comparative study on K-Means, Fuzzy C-Means, UCAM and Fuzzy-UCAM Clustering algorithm

	Initial cluster	Centriod	Threshold value	Cluster result	Cluster Error
K-Means	K	Initial seeds	-	Depend on initial seeds	Yes, if wrong seeds
Fuzzy C-Means	C	Initial seeds	-	Depend on initial seeds	Yes, if wrong seeds
UCAM	-	-	T	Depend on threshold value	-
Fuzzy-UCAM	-	-	T	Depend on threshold value	-

UCAM and Fuzzy-UCAM algorithm produce unique clustering only on the bases of affinity measure; hence there is no possibility of error in clustering. One major advantage of UCAM and Fuzzy-UCAM algorithm is that both rough clustering and accurate unique clustering is possible by adjusting the threshold value. But in K-Means and FCM clustering there is chance of getting error if the initial seeds are not identified properly.

8. MEASUREMENTS ON CLUSTER UNIQUENESS

The cluster representation of K-Mean and UCAM are illustrated through scatter graph as shown below in which each symbol indicates a separate cluster.

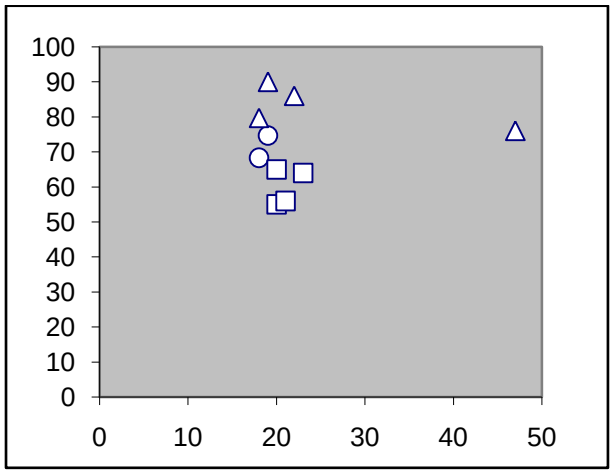

Fig 2 : Clustering through K-Means

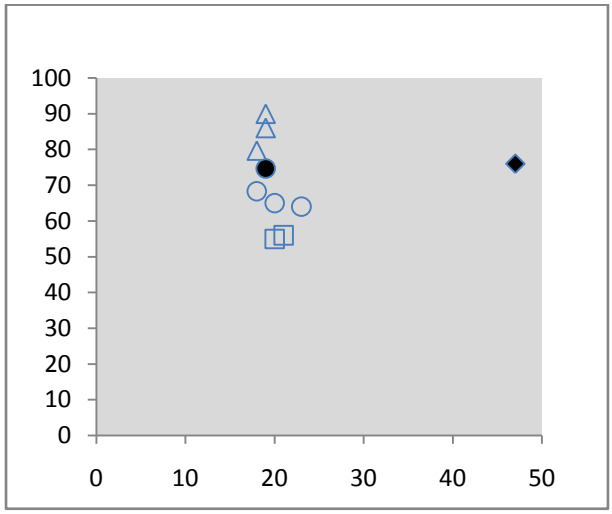

Fig 3: Clustering through UCAM

In the above graph Fig.3 all the cluster are unique in representation compared to K-Means clustering and the dark shaded symbols are peculiar objects, based on the application it can be projected out otherwise it can be merged with nearby cluster by adjusting the threshold value. Both approximate clustering and unique cluster can be obtained by increasing and decreasing the threshold values.

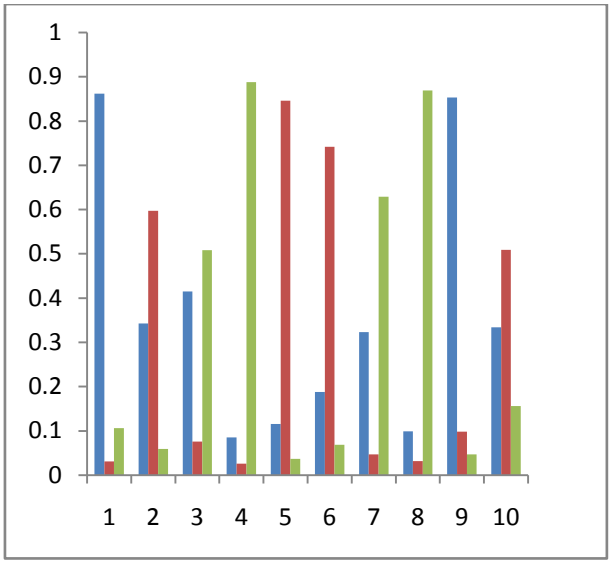

Fig 4: Clustering through Fuzzy C-Means

The table .1 and 2 is represented in the following bar chart which gives clear view on uniqueness of Fuzzy C-Means and Fuzzy-UCAM clustering. Each series indicate the possibility of particular data into all possible clusters.

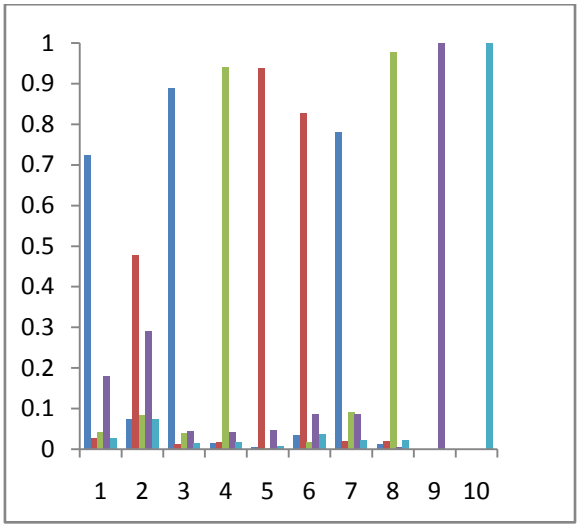

Fig 5: Clustering through Fuzzy-UCAM

Fig.5 gives the clear visualization on cluster uniqueness through Fuzzy-UCAM. If one object is classified into particular cluster then the degree of possibility towards other cluster is least significant. But in Fuzzy C-Means clustering it has the reasonable degree of possibility toward other clusters as shown in Fig.4.

9. CONCLUSIONS

In this paper, new UCAM and Fuzzy-UCAM algorithm is used for data clustering and results with fuzzy membership matrix. This approach reduces the overheads of fixing the cluster size and initial seeds as in K-Means and in FCM. UCAM and Fuzzy-UCAM fixes threshold value to obtain a unique clustering. The proposed methods improves the scalability and reduces the clustering error. This approach ensures that the total mechanism of clustering is in time without loss in correctness of clusters.

10. REFERENCES

[1] Hajing Li,Zaiquing Nie, WangChien Lee.: Scalable community Discovery on Textual Data with relations [http://www.ics.uci.edu/~mlearn/ MLRepository .html] Irvine, CA: University of California, Department of Information and Computer Science.

[2] S. Guha, R. Rastogi, and K. Shim. CURE.: An efficient clustering algorithm for large databases. In Proc. 1998 ACM6SIGMOD Int. Conf. Management of Data (SIGMOD'98), pages 73–84, 1998.

[3] Chen Zhang and Shixiong Xia.: K-Means Clustering Algorithm with Improved Initial center, in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 7906792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Don.: A New Algorithm to Get the Initial Centroids", proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26629, August 2004.

[5] Chaturvedi J. C. A, Green P, "K - Modes clustering." Journals of Classification, (18):35–55, 2001.

[6] Doulaye Dembele and Philippe Kastner, "Fuzzy C means method for clustering microarray data", bioinformatics, vol.19, no.8, pp.9736 980, 2003.

[7] Dongxiao Zhu, Alfred O Hero, Hong Cheng, Ritu Khanna and Anand Swaroop, "Network constrained clustering for gene microarray Data", doi:10.1093 bioinformatics / bti 655, Vol. 21 no. 21, pp. 4014 – 4020, 2005.

[8] G.K. Gupta .: Data mining with case studies.

[9] Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, et al. Inference from clustering with application to gene-expression microarrays. J Comput Biol 2002;9(1):105–26.

[10] Gasch AP, Eisen MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol 2002;3(11) RESEARCH0059.

[11] Bezdek J. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press; 1981.

[12] RM Suresh, K Dinakaran, P Valarmathie, "Model based modified k-means clustering for microarray data", International Conference on Information Management and Engineering, Vol.13, pp 271-273, 2009, IEEE.

[13] Han, Kamber, "Datamining Concepts and Techniques", Elsevier publications, 2005.

[14] Anil K. Jain and Richard C. Dubes, "Algorithms for clustering data", Prentice Hall, New Jersey,1988.

[15] Anirban Mukhopadhyay, Ujjwal Maulik and Sanghamitra bandyopadhyay, " Efficient two stage fuzzy clustering of microarray gene expression data", International Conference on Information Technology (ICIT'06), 2006 IEEE.

[16] Shi Zhong, Joydeep Ghosh, "A unified framework for model based clustering", Journal of Machine Learning Research 4 (2003) 1001-1037

[17] K.Dinakaran, RM.Suresh, P.Valarmathie, "Clustering gene expression data using self organizing maps, Journal of Computer Applications", Vol.1, No.4, 2008.

[18] Han-Saem Park and Sung-Bae Cho, "Evolutionary fuzzy clustering for gene expression profile E. Diday, The symbolic approach in clustering, in: H.H. Bock (Ed.), Classi3cation and Related Methods of Data Analysis, North-Holland, msterdam, 1988.

[19] Y.El-Sonbaty, M.A. Ismail, Fuzzy clustering for symbolic data, IEEE Trans. Fuzzy Systems 6 (2) (1998) 195–204.

[20] K.C. Gowda, E. Diday, Symbolic clustering using a new dissimilarity measure, Pattern Recognition 24 (6) (1991) 567–578.

[21] K.C. Gowda, E. Diday, Symbolic clustering using a new similarity measure, IEEE Trans. System Man Cybernet. 22 (1992) 368–378Z.

[22] Huang andM. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," IEEE Transactions on Fuzzy Systems, vol. 7, no. 4, 1999.