A Novel Approach in Clustering via Rough Set

A. Pethalakshmi[1], A. Banumathi[2]

[1]Associate Professor, Department of Computer Science,
MVM Government Arts College for Women, Dindigul, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science,
Government Arts College, Karur, Tamilnadu, India

Abstract: ***Clustering is a widely used technique in data mining application for discovering patterns in large dataset. K-Means and Fuzzy C-Means clustering algorithm are the traditional approach for clustering. Above mentioned algorithm are analyzed and found two drawbacks that the quality of resultant cluster is based on prior fixation of cluster size K and on sequentially or randomly selected initial seed. Our earlier proposes namely UCAM (Unique Clustering with Affinity Measures) and Fuzzy-UCAM over bridged the drawbacks of K-Means and Fuzzy C-Means. UCAM and Fuzzy-UCAM clustering algorithm works without initial seed and prior fixation on number of clusters, where the unique clustering is obtained with the help of affinity measures. In this paper Rough Set Attribute Reduction (RSAR) is hybridized with UCAM and Fuzzy-UCAM, which reduces the computational complexity, increases the cluster Uniqueness, and retains the originality of the data.***

Keywords: Cluster, UCAM, Fuzzy-UCAM, Rough Set.

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. 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. To overcome this drawback UCAM (Unique Clustering with Affinity Measures) where introduced.

Unique Clustering with Affinity Measures (UCAM) is a 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 [15]. The uniqueness of the cluster is based on the threshold value. More unique cluster is obtained when the threshold values is smaller. Fuzzy-UCAM is an enhancement on UCAM to have fuzzy measure for the object in the cluster [18]. In this article UCAM and Fuzzy-UCAM is upgraded with Rough Set by hybridized with RSAR to reduce the attributes count in the dataset without affecting the originality of the dataset. Reduction on the attribute count helps to reduce computational complexity, reduce the processing time and increase the cluster uniqueness.

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. These tools can include statistical models, mathematical algorithms and machine learning methods such as neural networks or decision trees etc. Data mining is popularly known as knowledge discovery. Figure.1 shows the concept of data mining, which involves three steps:

- Capturing and storing the data.
- Converting the raw data into information
- Converting the information into knowledge.

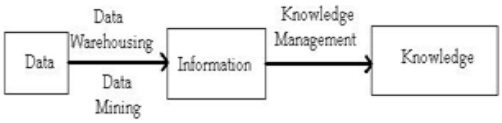


Figure 1: Data mining process

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. In this paper the UCAM (Unique Clustering with Affinity Measure) and fuzzy-UCAM clustering algorithm is enhanced with Rough set by using RSAR approach to reduce the computational complexity and processing time.

3. Related Work

3.1 K-Means

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. 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. 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. This procedure consists of the following steps, as described below

Algorithm 1: K-Means clustering algorithm

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

K – Number of desired cluster.
Output: A set of K clusters.

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.

3.2 UCAM

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. In K-Means algorithm, the process is initiated with the initial seeds and number of cluster to be obtained. But the number of cluster that is to be obtained cannot be predicted on a single view of the dataset. The result may not unique if the number of cluster and the initial seed is not properly identified.

UCAM algorithm is implemented with the help of affinity measure for clustering. The process of clustering in UCAM initiated without any centroid and number of clusters that is to be produced [17]. But it set 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.

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.

The UCAM clustering algorithm is initiated with the threshold value alone but it produces unique result. In unsupervised clustering to increase the cluster uniqueness UCAM is applied for clustering overcome the drawbacks of K-means.

3.3 Fuzzy C-Means

The fuzzy c-means clustering algorithm [3] 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[2]. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm [3]. The FCM algorithm attempts to partition a finite collection of n elements $X = \{x_1,\ldots,x_n\}$ 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 centre's $C=\{c_1,\ldots,c_c\}$ and a partition matrix $U=u_{ij} \in [0,1] i=1,\ldots,n,\ j=1,\ldots,c$ 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

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

which differs from the k-means objective function by the addition of the membership values u_{ij} 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. 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 center 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

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.

3.4 Fuzzy-UCAM

The fuzzy-UCAM clustering algorithm is a variation of the 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 as follows
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).

On implementing Fuzzy-UCAM clustering algorithm in the table I, which produces the result with the M x N matrix, where M is number of resultant cluster and N is the total number of objects in the initial set which is to be clustered [18]. The each row of the matrix indicates the degree of membership of the particular object towards all the clusters. The sum of each row should be the values between 0 and 1.

3.5 Rough Set

Rough set theory, first proposed by Pawlak in 1982 [19], employed mathematical modeling to deal with class data classification problems, and then turned out to be a very useful tool for decision support systems, especially when hybrid data, vague concepts and uncertain data were involved in the decision process. To use the rough set process, one begins with a relational database, a table of objects with attributes, and attributes values for each object. One attribute is chosen as the decision attribute, then the rest of the attributes are the condition attributes [19]. Rough sets address the continuing problem of vagueness, uncertainty and incompletion by applying the concept of equivalence classes to partition training instances according to specified criteria. Two partitions are formed in the mining process. The members of the partition can be formally described by unary set- theoretic operators or by successor functions for lower approximation and upper approximation spaces from which both possible rules and certain rules can be easily derived. Vague and imprecise data sets have no clear-cut boundaries. Thus, the rough set theory approach is based on refusing certain set boundaries, implying that every set will be roughly defined using a lower and an upper approximation. Let B # A and X # U be an information system. The set X is approximated using information contained in B by constructing lower and upper approximation sets:

BX ¼ fxj½x B # Xg (Lower approximation)
And BX ¼ fxj½x B \ X–;g (Upper approximation)

The elements in BX can be classified as members of X by the knowledge in B. However, the elements in BX can be classified as possible members of X by the knowledge in B. The set BNB ðxÞ ¼ BX BX is called the B-boundary region of X and it consists of those objects that cannot be classified

with certainty as members of X with the knowledge in B. The set X is called ‘‘rough” (or ‘‘roughly definable”) with respect to the knowledge in B if the boundary region is non-empty. Rough sets theoretic classifiers usually apply the concept of rough sets to reduce the number of attributes in a decision table [19] and to extract valid data from inconsistent decision tables. Rough sets also accept discretized (symbolic) input.

4. Methodology

This section briefly introduces the newly proposed method which enhances UCAM and Fuzzy-UCAM clustering algorithm with Rough Set for attribute reduction. The basic idea residing behind this hybridization of Rough with above mentioned clustering algorithm are to reduce the computational complexity, increase cluster uniqueness and to reduce the processing. In this proposed method Rough Set Attribute Reduction (RSAR) is carried as the initial step before clustering, so that it provides attribute reducted data set. Then UCAM clustering algorithm is used for unique clustering and Fuzzy-UCAM is used to supply membership degree to the resultant cluster. The algorithm representation of the new method is given in the following lines.

Algorithm 5: RSAR with UCAM and Fuzzy-UCAM

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

S – Threshold value.

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

Steps**:**

1. Set the threshold value T
2. Rough set attribute reduction is carried out for dimensionality reduction.
3. Clustering through UCAM.
4. (optional) Fuzzy measure is computed using Fuzzy-UCAM.

Here, the computational complexity is reduced by RSAR, which also helps to reduce the processing time. Uniqueness of cluster is provided by UCAM clustering algorithm. And fuzzy measure is provided by the Fuzzy-UCAM algorithm, where it is notated as optional since it may or may not be needed which differs based on the application.

5. Experimental Analysis

The newly proposed algorithm and UCAM and Fuzzy-UCAM clustering method is basically designed for numerical data. The following section gives the clear view on the efficiency and effectiveness of above listed algorithms when it is applied in a small numerical data set

5.1 K-Means and UCAM clustering

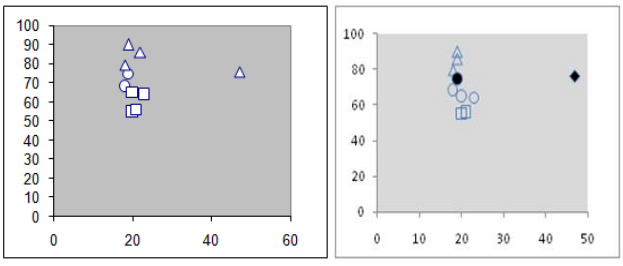
K-Means algorithm is implemented in a very small sample data with ten student’s information. The process of K-Means clustering is initiated with three initial seeds, which results with three clusters as notated below

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…S_{10} Student’s details which considers only numeric attributes. 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. 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 it, in the early stage. 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

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



(a) **(b)**

Figures 2: (a) Clustering through K-Means (b) Clustering through UCAM

In the above graph each symbol represents a separate cluster. In Figure 2 (a) shows the clustering overlaps with each other but in Figure 2 (b) 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.

5.2 Fuzzy C-Means and Fuzzy-UCAM

Uniqueness of the clusters for fuzzy c-means and fuzzy-UCAM is measured by using the same data that were used in k-means and UCAM. The membership matrix of fuzzy c-means and fuzzy-UCAM is illustrated in the following bar chart representation. Figure.3 gives the clear visualization on cluster uniqueness of numerical data representation on Fuzzy C-Means and Fuzzy-UCAM. Each series indicate the possibility of particular data into all other possible clusters. If one object is classified into particular cluster then the degree of possibility towards other cluster is least significant in the case of clustering through Fuzzy-UCAM figure.3 (b). But in Fuzzy C-Means clustering it has the reasonable degree of possibility toward other clusters as shown in figure.3 (a). The following chart gives clear view on the higher degree of uniqueness in clustering by Fuzzy-UCAM compared to Fuzzy C-Means.

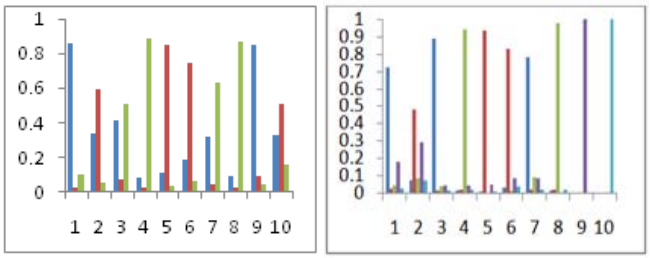


Figure 3: (a) Clustering through Fuzzy C-Means (b) Clustering through Fuzzy-UCAM

5.3 Analysis on RSAR with UCAM and Fuzzy-UCAM

In this section, the validity measure for the newly proposed method RSAR with UCAM is analyzed. Measures of inter cluster and Intra cluster distance is validated and succeed with positive result. The inter cluster distance or the distance between cluster is to be as big as possible. The intra distance measure is simply the distance between a point and its cluster center, which should be as less as possible. These inter and intra cluster distance is measure for the students sample data which is found with positive response as notated below table 1.

Table 1: Validity Measure

Validity Measure	*Number of cluster =4*		
	K-Means	*UCAM*	*RSAR with UCAM*
Inter Cluster Distance	*148.9325*	*153.7929*	*155.2471*
Intra Cluster Distance	*12.1167*	*11.0854*	*9.0781*

Table 1: (a)

Validity Measure	*Number of cluster =5*		
	K-Means	*UCAM*	*RSAR with UCAM*
Inter Cluster Distance	*143.9025*	*148.9375*	*150.0118*
Intra Cluster Distance	*8.9167*	*7.8812*	*6.4821*

Table 1: (b)

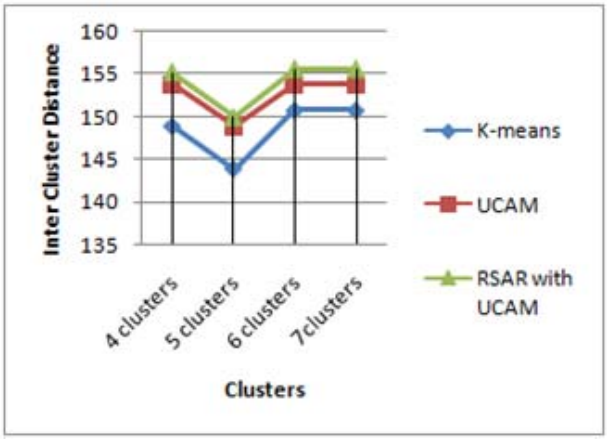
Validity Measure	*Number of cluster =6*		
	K-Means	*UCAM*	*RSAR with UCAM*
Inter Cluster Distance	*150.7929*	*153.7726*	*155.6273*
Intra Cluster Distance	7.8102	5.7938	4.8123

Table 1: (c)

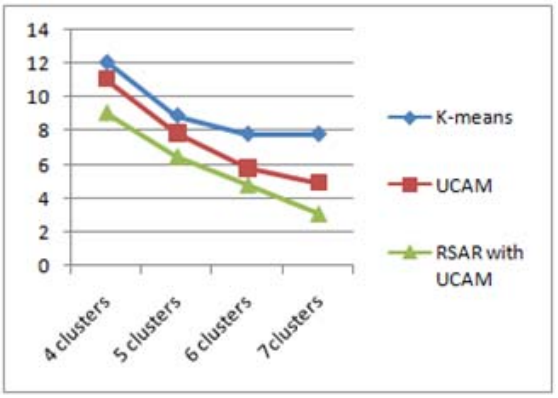
Validity Measure	*Number of cluster =7*		
	K-Means	*UCAM*	*RSAR with UCAM*
Inter Cluster Distance	*151.0919*	*154.2752*	*157.7442*
Intra Cluster Distance	*7.8102*	*4.9497*	*3.0981*

Table 1: (d)

In the table 1: k-Means, CAM and RSAR with UCAM is listed out on the bases of inter and intra cluster distance, which shows the maximum value for inter cluster distance and the minimum value for intra cluster distance for the proposed method than the other two methods. Following graph shows the clear view for the table values.



Figures 3: (a) Inter Clustering Distance



Figures 3: (b) Intra Clustering Distance

This section evidenced with proof that the enhancement of UCAM with RSAR have enriched the total environment of clustering..Rough set provides feature reduction and UCAM provides uniqueness, so that the computational complexity is reduced along with the reduction in processing time and moreover uniqueness is verified through inter and intra cluster distance measure. Based on the application if fuzzy measure is needed than Fuzzy-UCAM is applied, since Fuzzy-UCAM is enhanced view of UCAM, it retains the result of unique as in UCAM. All the three methods are purely works on affinity measure by setting the threshold value. On increasing the threshold value the number of cluster decreases and by decreasing, the number of cluster increases.

6. Comparative Analysis

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

Table 2: Comparison of clustering algorithms

	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				Depend on threshold value	
UCAM with RSAR	-	-	T	Depend on threshold value	-

The proposed method UCAM with RSAR, UCAM and Fuzzy-UCAM clustering 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 with RSAR, 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.

7. Conclusion

In this research paper, enhanced vision of UCAM algorithm is represented by hybridizing with RSAR, which reduces attributes in the data set without affecting its originality. The hybridization of UCAM with RSAR helps to reduce computational complexity, processing time and to increase the cluster uniqueness. This approach also reduces the overheads of fixing the cluster size and initial seeds as in K-Means. It fixes threshold value to obtain a unique clustering. The proposed method 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.

References

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

[2] 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.

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

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

[5] 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.

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

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

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

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

[11] 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.

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

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

[14] 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.

[15] A.Pethalakshmi. A.Banumathi,"Refinement Of K-Means And Fuzzy C-Means1**,** International Journal of Computer Applications,Volume 39, Paper Number : 17, Feb 2012.

[16] A.Pethalakshmi. A.Banumathi," Increasing Scalability through Affinity Measure in Clustering", SPRINGER CCIS , part II,ccis 270, p.302.

[17] A.Pethalakshmi. A.Banumathi," Novel Approach for Upgrading Indian Education by Using Data Mining Techniques", IEEE digital library.

[18] A.Pethalakshmi. A.Banumathi,"Increasing Cluster Uniqueness in Fuzzy C-Means through Affinity Measure", IEEE digital library.

[19] Pawlak,Z.(1991).Rough sets: Theoretical aspects of reasoning about data. Boston: Kluwer Academic Publishers

[20] 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

Author Profile

A. Pethalakshmi was born in 1965 at Karaikudi, Tamil Nadu (TN), India. She received her Master of Computer Science from Alagappa University, Tamilnadu, India in 1988 and the Master of Philosophy in Computer Science from Mother Teresa Women's University, Kodaikanal, Tamilnadu, India in 2000. She obtained her Ph.D degree from the Mother Teresa Women's University, Kodaikanal, Tamilnadu, India in 2008. Currently she is working as Associate Professor and Head, Department of Computer Science, M.V.M Government Arts College (W), Dindigul,

Tamilnadu, India. Her area of interests includes Data Mining, Rough Set, Fuzzy Set, Grid Computing and Neural Network.

A. Banumathi was born in Karur, Tamil Nadu (TN), India, in 1978. She received the Bachelor of Computer Science (B.Sc.) degree in 1998, Master of Computer Applications (M.C.A.) degree in 2001 and Master of Philosophy in Computer Science in 2005 from Bharathidasan University, Trichy, Tamilnadu, India. She is currently pursuing her Ph.D. degree in Computer Science from Manonmaniam Sundaranar University. Her research interests include data mining, Fuzzy Set and Rough Set.