Modifications in K-Means Clustering Algorithm

B. F. Momin, P. M. Yelmar

Abstract: ***In our study, we introduce modifications in hard K-means algorithm such that algorithm can be used for clustering data with categorical attributes. To use the algorithm for categorical data, modifications in distance and prototype calculation are proposed. To use the algorithm on numerical attribute values, mean is calculated to represent centre, and euclidean distance is used to calculate distance. Whereas, to use it on categorical attribute values, proportional representation of all the categorical values (probability) is used to represent center, and proportional weight difference is used as distance measure. For mixed data, we used discretization on numerical attributes to convert these attribute in categorical attribute. And algorithm used for categorical attributes is used.***

Other modifications use the combined fundamentals from rough set theory, fuzzy sets and possibilistic membership incorporated in k-means algorithm for numeric value only data. Same modifications are applied on the algorithm developed for categorical, and mixed attribute data. Approximation concept from rough set theory deals with uncertainty, vagueness, and incompleteness. Fuzzy membership allows dealing with efficient handling of overlapping clusters. Possibi1istic approach simply uses the membership value of data point in a cluster that represents the typicality of the point in the cluster, or the possibility of the point belonging to the cluster. Noise points or outliers are less typical; hence typicality-based (possibilistic) memberships reduce the effect of noise points and outliers. To verify the performance of algorithms DB index and objective function values are used.

Index Terms: Categorical data, clustering, fuzzy membership, k-means, possibilistic membership, rough set.

I. INTRODUCTION

CLUSTERING process groups a set of physical or abstract objects into classes of similar objects. The problem of clustering is defined as follows: Given a set of data objects, the problem of clustering is to partition data objects into groups in such a way that objects in the same group are similar while objects in different groups are dissimilar according to the predefined similarity measurement i. e. data belonging to one cluster are the most similar; and data belonging to different clusters are the most dissimilar [8], [10], [12], [20]. The unsupervised nature of the problem implies that its structural characteristics are not known, except if there is some sort of domain knowledge available in advance. Specifically, the spatial distribution of the data in terms of the number, volumes, densities, shapes, and orientations of clusters (if any) are unknown. Data objects are described by attributes of distinct natures, (binary, discrete, continuous, and categorical). However, finding the optimal clustering result has been proved to be an NP-hard problem. [16], [17]. In the literature, researchers have proposed many solutions for this issue based on different theories, and many surveys focused on special types of clustering algorithm have been presented [4], [5], [9], [10], [11], [13], [15], [16], [19]. Clustering plays an important role in many engineering applications, such as data compression, pattern recognition, image processing [9], system modeling, communication, remote sensing, biology, medicine, data mining [20], machine learning, and information retrieval [16].

Clustering algorithms can be generally classified as: hierarchical, partition-based, density-based, grid-based, and model-based [18], [20]. Most widely used partitional clustering algorithm is hard *c*-means (HCM) [1], where each object must be assigned to exactly one cluster. Whereas fuzzy *c*-means (FCM) [1], [14], [18] relaxes this requirement and allow the data belong to more than one cluster at the same time. The FCM algorithm assigns memberships which are inversely related to the relative distance of data points to the cluster centers. Suppose c=2. If data x_k is equidistant from two centers, the membership of x_k in each cluster will be the same, regardless of the absolute value of the distance of from the two centers (as well as from the other points in the data). This creates the problem that noise points, far but equidistant from the center of the two clusters, can nonetheless be given equal membership in both, when it seems far more natural that such points be given very low (or even no) membership in either cluster. To reduce this weakness of the FCM and to produce memberships that having good degrees of belonging for the data, Krishnapuram and Keller [2], [6] proposed a possibilistic membership approach However, the possibilistic *c*-means (PCM) sometimes generates coincident clusters [6]. Rough-set-based [3], [7] clustering provides a solution that is less restrictive than conventional clustering and less descriptive than fuzzy clustering. Rough set is a mathematical tool for managing uncertainty, vagueness, and incompleteness that arises from the indiscernibility between objects in a set. Lingras [7] proposed a new clustering method called rough *c*-means (RCM), which describes a cluster center and a pair of lower and upper approximations. By combining both rough and fuzzy sets, new *c*-means algorithm(RFCM), is introduced by Mitra [3] where each cluster is consist of a fuzzy lower approximation and a fuzzy boundary. Each object in lower approximation takes a weight corresponding to fuzzy membership value. However, the objects in lower approximation of a cluster should have a similar influence on the corresponding centers, and their weights should be independent of other centers and clusters. So it drifts the cluster centers from their desired locations.

In this paper, we proposed an algorithm termed as rough–fuzzy Possibilistic C-Means (RFPCM). Membership function of the fuzzy sets enables overlapping clusters, and the concept of lower and upper approximations from rough sets handles uncertainty, vagueness, and incompleteness; Whereas possibilistic membership functions generate memberships which are compatible with the center of the class and not coupled with centers of other classes. The algorithm modified to use on categorical data by using probability distribution of

Manuscript received on July 09, 2012.

Dr. Bashirahamad F. Momin , Computer Science and Engineering Department , Walchand College of Engineering Sangli-416415, India,

Prashant M. Yelmar, Computer Engineering Department, S. B. Patil College of Engineering, Indapur, Maharashtra-486103, India.

categorical values.

II. ALGORITHMS

A.Hard C- Means (HCM)

In HCM [20] each object is assigned to exactly one cluster. The main steps of the *c*-means algorithm [1] are as follows.

1) Assign initial means v_i (also called centers) for each cluster.
2) Assign each data object x_k to the cluster U_i with the closest mean.
3) Compute new mean for each cluster using

$$v_i = \frac{1}{n_i} \sum_{x_k \in U_i}^{n} x_k \qquad (1)$$

4) Iterate Steps 2) and 3) **until** criterion function

$$J_{HCM} = \sum_{i=1}^{c} \sum_{x_j \in U_i} \left\| x_j - v_i \right\|^2 \qquad (2)$$

Converges, i.e., there are no more new assignments of objects.

B. Fuzzy C-Means (FCM)

FCM [1], [2], [18] allows one data object to belong to two or more clusters at the same time. The memberships are inversely related to relative distance of object x_k to the center v_i.They are calculated by using

$$u_{ik} = \left\{ \sum_{j=1}^{j=c} \left\{ \frac{d_{ik}}{d_{jk}} \right\}^{\frac{2}{m-1}} \right\}^{-1} \quad \forall\ i,k \qquad (3)$$

Where ${d_{ik}}^2 = \|x_k - v_i\|^2$;

$1 \le m \le \infty$ (ideally selected as 2);

$\mu_{ik} = [0,1]$, probabilistic membership of x_k to cluster β_i. FCM partitions data set into c clusters by minimizing o objective function

$$J_{FCM} = \sum_{i=1}^{i=c} \sum_{k=1}^{k=n} \mu_{ik}^m \|x_k - v_i\|^2 \qquad (4)$$

subject to $\sum_{i=1}^{i=c} \mu_{ik} = 1;\ \ k = 1, \dots, n.$ and $0 < \sum_{k=1}^{k=n} \mu_{ij} < n, \forall\ i,j.$

Steps in FCM:-

1) Randomly choose c objects as centers of c clusters.
2) Calculate membership based on relative distance.
3) Calculate new centers using

$$v_i = \frac{\sum_{k=1}^{k=n} (\mu_{ik}^m x_k)}{\sum_{k=1}^{k=n} (\mu_{ik}^m)} \quad \forall\ i \qquad (5)$$

4) Iterate until criterion function converges.

C.PCM

FCM becomes very sensitive to noise and outliers because data point memberships are inversely related to the relative distance of the data to the cluster centers. In addition, for compatibility with the center, the membership of an object x_k in a cluster β_i should be determined solely by center v_i of the cluster and should not be coupled with its similarity with respect to other clusters. To handle this problem, Krishnapuram and Keller [2], [6] proposed PCM. For PCM objective function is formulated as

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} (v_{ij})^{m2} \left\| x_j - v_i \right\|^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - v_{ij})^{m2} \qquad (6)$$

Where $1 \le m_2 \le \infty$ is the fuzzifier, and η_i represents the scale parameter. The update equation of v_{ij} is given by

$$v_{ij} = \frac{1}{1 + D} \qquad (7)$$

Where $D = \left(\frac{\|x_j - v_i\|^2}{\eta_i} \right)^{1/m2-1}$

subject to $v_{ij} \in [0,1], \forall\ i,j$; and $0 < \sum_{j=1}^{n} v_{ij} \le n, \forall i$;and $\max_i v_{ij} > 0, \forall j$.

The Scale parameter represents the zone of influence or size of the cluster β_i. The update equation for η_i is

$$\eta_i = K . \frac{P}{Q} \qquad (8)$$

Where $P = \sum_{i=1}^{n} (v_{ij})^{m2} \left\| x_j - v_i \right\|^2$ and $Q = \sum_{i=1}^{n} (v_{ij})^{m2}$ Value of K is chosen to be one. In each iteration, the new value of v_{ij} depends only on the similarity between the object x_j and the center v_i.The resulting cluster of the data can be interpreted as a possibilistic cluster, and the membership values may be interpreted as degrees of possibility of the objects belonging to the cluster, i.e., the compatibilities of the objects with the center.

D. Rough C-Means (RCM).

The rough set [1], [3], [7] is a mathematical tool for managing uncertainty that arises from the indiscernibility between objects in a set. It approximates a rough (imprecise) concept by a pair of exact concepts, lower and upper approximations. The lower approximation is the set of objects definitely belonging to the vague concept, whereas the upper approximation is the set of objects possibly belonging to the same. RCM views each cluster as an interval or rough set [3], [7]. A rough set X is characterized by its lower and upper approximations $\underline{B}X$ and $\overline{B}X$, respectively, with the following properties.

i. An object x_k can be part of at most one lower approximation.
ii. If x_k $\underline{B}X$ of cluster X, then simultaneously x_k $\overline{B}X$.
iii. If x_k is not a part of any lower approximation, then it belongs to two or more upper approximations.

This permits overlaps between clusters. The center computation is modified by incorporating the concepts of upper and lower approximations. Since objects in the lower approximation definitely belong to a rough cluster, they are assigned a higher weight by parameter w_{low}. The objects lying in the upper approximation are assigned a relatively lower weight by parameter w_{up} during computation. The center of cluster is calculated by equation

$$v_i^R = \begin{cases} w_{low} \times C_1 + w_{up} \times D_1 & if\ \underline{B}U_i \neq \emptyset\ and \\ & \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ C_1 & if\ \underline{B}U_i = \emptyset\ and \\ & \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ D_1 & otherwise \end{cases} \qquad (9)$$

Where the parameters w_{low} and w_{up} correspond to the relative importance of the lower

and upper approximations, respectively such that $w_{low} + w_{up} = 1$. Here, $|\underline{B}U_i|$ indicates the number of patterns in the lower approximation of cluster U_i while $|\overline{B}U_i - \underline{B}U_i|$ is the number of patterns in the rough boundary. RCM is found to generate three types of clusters, such as those having objects:

i. In both the lower and upper approximations;
ii. Only in lower approximation;
iii. Only in upper approximation.

The condition for an object belonging to the lower or upper bound of a cluster is explained as next. Let x_k be an object at distance d_{ik} from centroid v_i of cluster U_i. The difference $d_{ik}-d_{jk}$, $i \neq j$, used to determine whether x_k should belong to lower or upper bound of a cluster. The algorithm steps are as follows

1) Assign initial means v_i (also called centers) for each cluster.
2) For each data object x_k compute difference $d_{ik}-d_{jk}$, $i \neq j$, from center pairs v_i and v_j.
3) Let d_{ik} be minimum and d_{jk} be next to minimum. If difference (d_{jk} - d_{ik}) is less than some threshold, then x_k belong to upper approximations of both clusters else x_k belong to lower approximation of cluster U_i such that distance d_{ik} is minimum over all c clusters.
4) Compute new center for each cluster using (9).
5) Iterate Steps 2)-4) until criterion function converges. Objective function is given by

$$J_R = \begin{cases} w_{low} \times A_1 + w_{up} \times B_1 & if\ \underline{B}U_i \neq \emptyset\ and\ \overline{B}U_i - \underline{B}U_i = \emptyset \\ A_1 & if\ \underline{B}U_i \neq \emptyset\ and\ \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ B_1 & otherwise \end{cases} \quad (10)$$

Where $A_1 = \sum_{i=1}^{c} \sum_{x_k \in \underline{B}U_i} \|x_k - v_i\|^2$ and

$B_1 = \sum_{i=1}^{c} \sum_{x_k \in B(U_i)} \|x_k - v_i\|^2$, $B(U_i) = \overline{B}U_i - \underline{B}U_i$.

The performance of the algorithm is dependent on the choice of w_{low}, w_{up}, and threshold. Used combinations are $w_{up} = 1 - w_{low}$, $0.5 < w_{low} < 1$, and $0 <$ threshold < 0.5. An optimal selection of these parameters is an issue of research interest.

E. Rough, Fuzzy, Possibilistic C-Means (RFPCM)

RFPCM [19] adds both probabilistic and possibilistic memberships and the lower and upper approximations of rough sets into *c*-means algorithm. While the membership of fuzzy sets enables efficient handling of overlapping partitions, the rough sets deal with uncertainty, vagueness, and incompleteness in class definition. Integration of both probabilistic and possibilistic memberships avoids the problems of noise sensitivity of the FCM and the coincident clusters of the PCM. Fig. 1 provides a schematic diagram of a rough set X within the upper and lower approximations, consisting of granules from the rectangular grid.

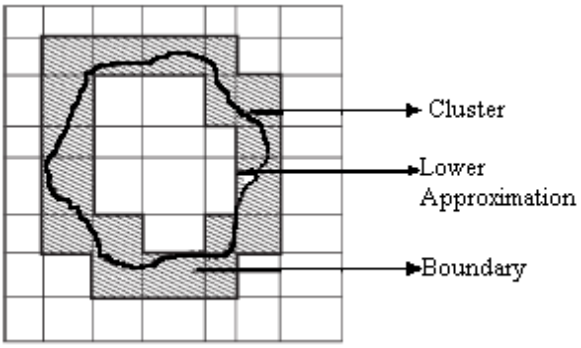

Fig. 1 RFPCM. Cluster is represented by lower bound and fuzzy boundary.

RFPCM algorithm steps are outlined as follows:-

1) Randomly assign c objects as centers of c clusters.
2) Calculate probabilistic and possibilistic memberships for all objects using (3) and (7) respectively.
3) The scale parameters η_i for c clusters are calculated by using (8). According to Krishnapuram and James M. Keller [6] the value of η_i can be fixed for all iterations or it may be varied in each iteration. In our experimentation we used fixed value of η_i.
4) Compute
$$u_{ij} = \{a\mu_{ij} + bv_{ij}\}$$
for all clusters and all data objects where i=1,…,c and j=1,…,n.
5) A Sort all u_{ij} and the difference of two highest memberships of x_j are compared with threshold δ.
6) A Let μ_{ij} and μ_{kj} highest and second highest memberships of x_j respectively. If $(\mu_{ij} - \mu_{kj}) > \delta$ then $x_j \in \underline{B}U_i$, as well as $x_j \in \overline{B}U_i$; otherwise $x_j \in \overline{B}U_i$ and $x_j \in \overline{B}U_k$. Now modify the membership values μ_{ij} and v_{ij}.
7) Calculate new centers by

$$V_i^{RFP} = \begin{cases} w_{low} \times C1 + w_{up} \times D1 & if\ \underline{B}U_i \neq \emptyset\ and\ B(U_i) \neq \emptyset \\ C1 & if\ \underline{B}U_i \neq \emptyset\ and\ B(U_i) = \emptyset \\ D1 & otherwise \end{cases} \quad (11)$$

The δ represents the size of granules of rough–fuzzy clustering and selected as $0 < \delta < 0.5$.
8) Iterate Steps 2)-7) **until** criterion function converges.

Objective function for RFPCM is calculated as:

$$J_{RFP} = \begin{cases} w_{low} \times A1 + w_{up} \times B1 & if\ \underline{B}U_i \neq \emptyset\ and\ B(U_i) \neq \emptyset \\ A1 & if\ \underline{B}U_i \neq \emptyset\ and\ B(U_i) = \emptyset \\ B1 & otherwise \end{cases} \quad (12)$$

Where

$$A1 = \sum_{i=1}^{k} \sum_{x_j \in \underline{B}(\mu_i)} \{a(\mu_{ij})^{m1} + b(v_{ij})^{m2}\} \|x_k - v_i\|^2 + \sum_{i=1}^{k} \eta_i \sum_{x_j \in \underline{B}(\mu_i)} (1 - v_{ij})$$

And

$$B1 = \sum_{i=1}^{k} \sum_{x_j \in B(\mu_i)} \{a(\mu_{ij})^{m1} + b(v_{ij})^{m2}\} \|x_k - v_i\|^2 + \sum_{i=1}^{k} \eta_i \sum_{x_j \in B(\mu_i)} (1 - v_{ij})$$

III. C-MEANS FOR CATEGORICAL DATA CLUSTERING

In this section we introduced C-Means algorithm for categorical data clustering. Earlier Ralambondrainy [4], [5] presented k-means to cluster categorical data by converting multiple categorical attributes into binary attributes, each using one for presence of a category and zero for absence of it, and then treats these binary attributes as numeric ones in the k-means algorithm. This needs to handle a large number of binary attributes when data sets have attributes with many categories increasing both computational and storage cost. The other drawback is that the cluster means given by real values between zero and one do not indicate the characteristics of the clusters. The k-modes algorithm introduced by Zhexue Huang [4] extends the k-means algorithm by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. These extensions have removed the numeric-only limitation of the k-means algorithm. We further extended this idea by using by using probability distribution for distance calculation as well as center representation. In our algorithm, for center representation we count number of instances in cluster for particular value (instance) of particular categorical attribute. To calculate distance from center over particular attribute mathematical formula (1-probability of instance value on that category) is used. Objective function is formulated as

$$E = \sum_{i=1}^{c} \sum_{k=1}^{n} d(X_k, Q_i) \quad (13)$$

Where $d(X_k, Q_i)$ is the distance of data object X_k from cluster center Q_i. This distance measure is formulated as equation

$$d(X_k, Q_i) = \sum_{a=1}^{m} 1 - q_{ia} \ , \forall\, d, d = 1, \dots, m \quad (14)$$

cluster center or prototype or is representative vector for cluster *i* is defined as

$$Q_i = [q_{i1}, q_{i2}, \dots, q_{im}] \quad (15)$$

In which *m* is the number of attributes.

$$q_{im} = \left[\left[[Fr_{im1}, Fr_{im2}, \dots, Fr_{imd}]\right]/N\right] \quad (16)$$

Where Fr_{imd} is frequency of value *d* for the attribute *m* in the cluster *i*. N are number of data objects present in cluster *i*. This process is explained with following example. Suppose Attribute1 has domain values {R,G,B}; Attribute2 has domain {A,B,C,D,E}; Attribute3 has domain {X,Y}; Attribute4 has domain {L,M,N,O}.

TABLE I SAMPLE INSTANCE FOR CLUSTER

Sr. No.	Attributes			
	Attribute1	Attribute2	Attribute3	Attribute4
1	R	A	X	L
2	G	B	Y	L
3	G	A	X	O
4	R	C	X	N
5	R	E	X	M
6	B	D	Y	N
7	G	D	X	O
8	B	A	X	L
9	G	C	Y	N
10	B	D	X	L

So center prototype which is calculated by (15) and (16) will be

Q=[(0.3,0.4,0.3);(0.3,0.1,0.2,0.3,0.1);(0.7,.0.3);(0.4,0.1,0.3,0.2)]. Calculation of $(X_3, Q_i) = (1 - 0.4) + (1 - 0.3) + (1 - 0.7) + (1 - 0.2) = 2.4$. Iterative steps are same as that for numeric data.

IV. ALGORITHM RESULT EVALUATION CRITERION

To evaluate the performance of algorithm on various data sets objective function value and DB index are used in case of numeric data sets whereas objective function value is used in case of categorical data sets. The DB is a function of the ratio of the sum of within-cluster distance to between-cluster separation. Let $\{x_1, \dots, x_{|ci|}\}$ be data objects in a cluster U_i ,and then average distance between objects within the cluster U_i is given by

$$S(U_i) = \frac{\sum_{k,k'} \|x_k - x_{k'}\|}{|c_i|(|c_i|-1)} \quad (17)$$

Where$x_k, x_{k'} \in U_i$ and $\neq k'$. The between-cluster separation is defined as

$$d(U_i, U_l) = \frac{\sum_{i,j} \|x_k - x_j\|}{|c_i||c_l|} \quad (18)$$

Where$x_k \in U_i,\ x_j \in U_l$ such that$k \neq l$.

The optimal results minimizes following formula for DB

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{j \neq i} \left\{ \frac{S(U_i) + S(U_j)}{d(U_i, U_j)} \right\} \quad (19)$$

for $1 \leq i, j \leq c$.

V. RESULTS

Experimentation is performed on various datasets from http://www.ics.uci.edu/~mlearn. Runs are performed with c=3. Other parameters are $w_{low} = 0.99$, m_1=m_2=2.0; *a*=*b*=0.5. The parameters are held constant across all runs. To run the algorithm on mixed data sets, the attributes with numerical values are discritized and converted to categorical form. For numeric value data sets results are tabulated as

TABLE II. PERFORMANCE OF ALGORITHMS (NUMERIC VALUE DATA)

ALGORITHM	IRIS DATA		GLASS		WINE	
	DB	Obj. Fun.	DB	Obj. Fun.	DB	Obj. Fun.
HCM	0.565	78.94	3.33	727.19	7.91	11217.24
FCM	-	60.57	-	363.16	-	7411.55
RCM	0.487	64.86	3.14	656.48	8.21	10384.73
RFPCM	0.462	43.28	0.69	53.23	2.24	915.68

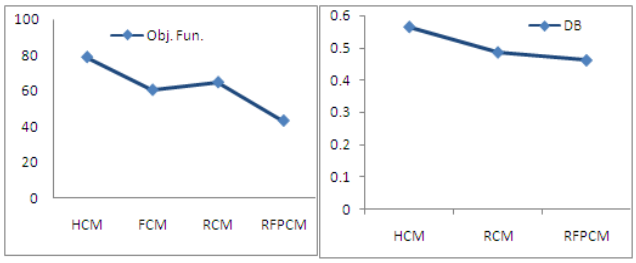

Fig. 2 Objective Function Value and DB index for Iris Data Set.

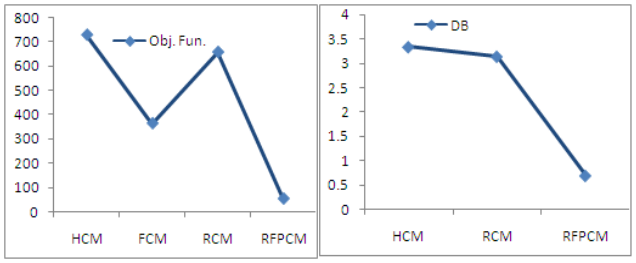

Fig. 3 Objective Function Value and DB index for Glass Data Set.

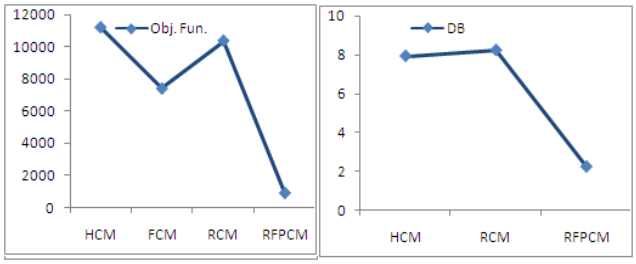

Fig. 4 Objective Function Value and DB index for Wine Data Set.

Lower values of DB and objective function indicate improvement in performance of algorithm. For each data set value of DB and objective function is lowest for RFPCM. So we can say RFPCM significantly performs over HCM, FCM and RCM by removing limitations of particular individual algorithm for numeric data sets.

TABLE III. PERFORMANCE OF ALGORITHMS (CATEGORICAL DATA)

	Teaching Evaluation Data Set			Contraceptive Method Data Set		
	Obj. Fun. (Max)	Obj. Fun. (min)	Obj. Fun. (converg)	Obj. Fun. (Max)	Obj. Fun. (min)	Obj. Fun. (converg)
HCM	785	574.55	574.55	12463.4	7958.79	7959.07
FCM	582.90	349.29	353.05	9360.31	5794.21	5794.21
RCM	269.38	256.21	268.53	5790.81	5681.30	5765.47
RFPCM	102.68	61.32	61.88	2594.76	2174.99	2184.20

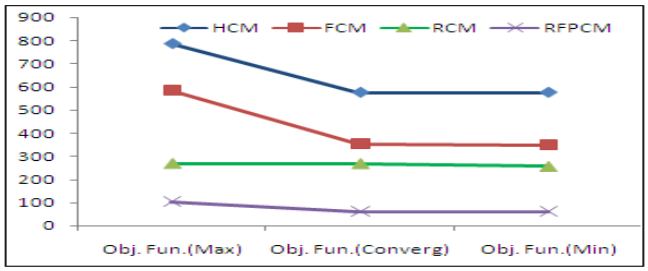

Fig. 5 Objective Function Value for Teaching Evaluation Data Set.

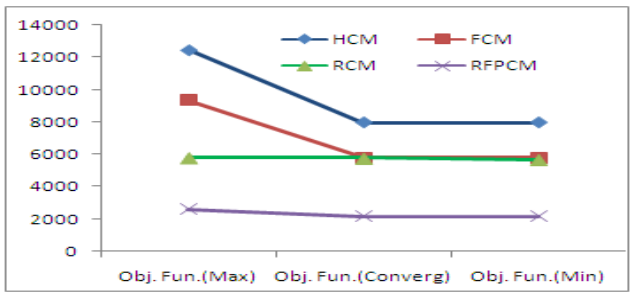

Fig. 6 Objective Function Value for Contraceptive Method Use Data Set.

Above results show that modified k-means algorithm gives reduced value of objective function for categorical data clustering. If we observe stability of algorithm in terms of objective function value for minimum value and converged value, these values are equal or almost equal. Results show that there is significant reduction in objective function value from maximum (which occur at first iteration) to local minimum or converged value of objective function for each algorithm. Whereas values are decreasing in sequences from HCM, FCM, RCM to RFPCM. So we can say RFPCM for categorical data performs better over other c-mean variants. Among these algorithms RFPCM gives improved results over other variations of k-means algorithm.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their insightful comments and suggestions to make this paper more readable. The authors would like to thank Dr. P. J. Kulkarni Dy. Director, Walchand college of Engineering, Sangli for his continuous encouragement for research work.

REFERENCES

1. P. Maji and S. K. Pal, "Rough–fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis," IEEE Trans. Knowl. Data Eng., vol. 19, no. 6, pp. 859–872, Jun. 2007.
2. Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek," A Possibilistic Fuzzy c-Means Clustering Algorithm," IEEE Trans. Fuzzy Syst, Vol. 13, no. 4, Aug 2005.
3. S. Mitra, H. Banka, and W. Pedrycz, "Rough–fuzzy collaborative clustering," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 4, pp. 795–805, Aug. 2006.
4. Zhexue Huang, Michael K. Ng.," A fuzzy k-modes algorithm for clustering categorical data," IEEE Trans. on fuzzy systems., Vol 7 No 4. August 1999.
5. Chen Ning, Chen An, Zhou Long-xiang ," Fuzzy k-prototypes algorithm for clustering mixed Numeric and categorical valued data," Journal of software Vol.12 No. 8,2001.
6. R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98–110, May 1993.
7. Pawan Lingras, Min Chen, and Duoqian Miao," Rough Cluster Quality Index Based on Decision Theory," IEEE Trans. Knowl. Data Eng., vol. 21, no. 7, July 2009.
8. Tapas Kanungo,David M. Mount, Nathan S. Netanyahu,Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 24, NO. 7, PP. 881-892, 2002.
9. Sankar K. Pal , Pabitra Mitra ," Multispectral Image Segmentation Using the Rough-Set-Initialized EM Algorithm", IEEE Transactions on Geoscience and

Remote Sensing", VOL. 40, NO. 11, PP. 2495-2501, 2002.

10. Jacek M. Leski ," Generalized Weighted Conditional Fuzzy Clustering", IEEE Trans. on Fuzzy Systems , VOL. 11, NO. 6, PP. 709-715, 2003.
11. Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li," Automated Variable Weighting in k-Means Type Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, NO. 5, PP. 657-668, 2005.
12. Jian Yu," General C-Means Clustering Model", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, NO. 8, PP.1197-2111, 2005.
13. Carlos Ordonez ," Integrating K-Means Clustering with a Relational DBMS Using SQL", IEEE Trans. Knowl. Data Eng.,, VOL. 18, NO. 2, PP. 188-201, 2006.
14. Francesco Masulli, Stefano Rovetta," Soft Transition From Probabilistic to Possibilistic Fuzzy Clustering", IEEE Trans. on Fuzzy Systems, VOL. 14, NO. 4, PP.516-527, 2006.
15. Michael K. Ng, Mark Junjie Li, Joshua Zhexue Huang, and Zengyou He," On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm," IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 29, NO. 3, PP. 503-507, 2007.
16. Hung-Leng Chen, Kun-Ta Chuang, and Ming-Syan Chen, "On Data Labeling for Clustering Categorical Data", IEEE Trans. Knowl. Data Eng.,VOL. 20, NO. 11, PP.1458-1471, 2008.
17. Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho ," A Survey of Evolutionary Algorithms for Clustering", IEEE Trans. Syst., Man, Cybern.—Part C: Appl. And Review, Vol. 39, No. 2,PP.133-155,2009.
18. Lin Zhu, Fu-Lai Chung, and Shitong Wang," Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions", IEEE Trans. Syst., Man, Cybern. B, Cybern ,VOL. 39, NO. 3, PP.578-591, 2009.
19. Pradipta Maji and Sankar K. Pal," Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices," IEEE Trans. Syst., Man, Cybern. B, Cybern , vol. 37, no. 6, Dec 2007.
20. J iawei Han, Micheline Kamber,"Data Mining:Concepts and Techniques",Second Edition,Elesvier Publications,2006.

AUTHORS PROFILE

Dr. Bashirahamad F. Momin is working as Associate Professor & Head, Dept. of Computer Science & Engineering, Walchand College of Engineering Sangli, Maharashtra State India. He received the B.E. and M.E. degree in Computer Science & Engineering from Shivaji University, Kolhapur, India in 1990 and 2001 respectively.

In February 2008, he had completed his Ph.D. in Computer Science and Engineering from Jadavpur University, Kolkata. He is recognized Ph.D. guide in Computer Science & Engineering at Shivaji University, Kolhapur. His research interest includes pattern recognition & its applications, data mining and soft computing techniques, systems implementation on the state of art technology. He was a Principal Investigator of R & D Project titled "Data Mining for Very Large Databases" funded under RPS, AICTE, New Delhi, India. He had delivered a invited talk in Korea University, Korea and Younsea University, Korea. He had worked as "Sabbatical Professor" at Infosys Technologies Ltd., Pune. He is a Life Member of "Advanced Computing and Communication Society", Bangalore INDIA. He was a student member of IEEE, IEEE Computer Society. He was a member of International Unit of Pattern Recognition and Artificial Intelligence (IUPRAI) USA.

Prashant M. Yelmar is working as Assistant Professor at S. B. Patil College of Engineering, Indapur, Maharashtra, India. He completed his BE in Information Technology from Mumbai University and M. Tech in Computer Science and Engineering from Walchand college of Engineering Sangli. His research interest includes Data Mining, Information Retrieval, Geographic Information Systems, Time Series Data Mining, and Soft Computing.