ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 3, Issue 10, October 2016

Comparative Efficiency Analysis of K-Means, Fuzzy Class and Rough Class Clustering Algorithm with IRIS Dataset with Multiple Centroid

[1] Arup Kumar Bhattacharjee, [2] Arup Kumar Bhattacharjee,[3] Soumen Mukherjee,[4] Krishnendu Paul,
[5]Dipan Mitra [6] Poulami Mukherjee
RCC Institute of Information Technology

Abstract: **: Data analysis is considered as an efficient and handy tool for processing huge amount of data which is very tough and data mining technology identifies patterns and trends of these data. This technique is used to extract the unknown pattern from a large dataset helping unreal time applications. Raw data from this dataset are classified by using cluster analysis-an important method for classifying data, finding clusters based on similarities with the same cluster and dissimilarities with others. There are various algorithms which are used to solve this problem like, K-Means, Fuzzy C Means (FCM), Rough C means, Rough Fuzzy C means. A comparative study of these algorithms is done in this paper. These algorithms are implemented in MATLAB using a set of real life data sets. So, this paper is a blend of Mathematics, Statistics and Computer Application.**

Key words:— **Data analysis, data mining, K-means, Fuzzy C Means, Rough C Means, Rough Fuzzy C-Means.**

I. INTRODUCTION

Data clustering is the method of data description. It is used as a common technique for data analysis in d data mining, image analysis, machine learning, bioinformatics and pattern recognition. Clustering algorithms are essentially a partitioning method which deals with enormous data and portioned them into different clusters [1].

Amongst the most excellent clustering methods, k-means (or hard C-Means) is used in wide researches [2-4]. It partitions the observations into "K"no. of clusters, each observation belonging to the cluster with close characteristic. On other hand, fuzzy C means (or soft K-Means) relaxes this requirement by allowing gradual memberships. It deals with the data that belong to more than one cluster simultaneously. It assigns with the memberships to an object .It also deals with some uncertainties, like those which are coming from overlapping cluster boundaries [10]. The rough sets provides options of approximation of lower and upper boundaries .the rough fuzzy C means clustering algorithm enhances the objective function and further distributions the membership function for the traditional fuzzy k-means clustering. [5]. Fuzzy-C means tends to run slower than K means, as it does much work. As each point is evaluated with every cluster, the operations are involved more in each evaluation. K-Means just calculates distance, whereas fuzzy c means does an inverse-distance weighting operation [6]. We can combine fuzzy set and rough set to get a significant direction to deal with uncertainty of datasets. Both fuzzy sets and rough sets do provide a mathematical framework to capture uncertainties associated with the data [7,8] which are actually complementary in few aspects. Recently, combining both rough and fuzzy sets, rough-fuzzy c-means (RFCM) has been proposed .In RFCM each cluster consists of a fuzzy lower approximation along with a fuzzy boundary. Each object in (lower) approximation takes a unique weight, which is its fuzzy membership value. On the other hand, the object which has lower approximation value to a cluster has similar control on the subsequent clusters and centroids with their weights independent of other clusters and centroids. Thus, the concept of lower approximation in fuzzy, dealt in RFCM of [8], reduces the weights of objects of lower approximation [11]. This method deviates the cluster prototypes from their desired locations. And to note that it's sensitive to noise and outliers. In this paper, the clustering algorithms are analyzed based on their clustering efficiency.

II OUR WORK

A comparative study between K Means, Fuzzy C Means, Rough C-Means and Rough Fuzzy C-Means algorithms has been implemented. The algorithms along with the equations have been described in brief. A standard data set IRIS has been taken and Comparative Efficiency

Analysis of K-Means, Fuzzy Class and Rough Class Clustering Algorithm with IRIS Dataset with Multiple Centroid Arup Kumar Bhattacharjee, Soumen Mukherjee, Krishnendu Paul, Dipan Mitra, Poulami Mukherjee RCC Institute of Information Technology ABSTRACT: Data analysis is considered as an efficient and handy tool for processing huge amount of data which is very tough and data mining technology identifies patterns and trends of these data. This technique is used to extract the unknown pattern from a large dataset helping unreal time applications. Raw data from this dataset are classified by using cluster analysis-an important method for classifying data, finding clusters based on similarities with the same cluster and dissimilarities with others. There are various algorithms which are used to solve this problem like, K-Means, Fuzzy C Means (FCM), Rough C means, Rough Fuzzy C means. A comparative study of these algorithms is done in this paper. These algorithms are implemented in MATLAB using a set of real life data sets. So, this paper is a blend of Mathematics, Statistics and Computer Application. every algorithm has been implemented on the data set separately and the results have been displayed in tabular formats. The resultant table states the how many data belong to which cluster. This method has been implemented for each of the clustering algorithms.

III CLUSTERING ALGORITHMS

3.1 K-means

Step 1) Set the number of clusters and Initialize the centroid for each clusters and choose the value of threshold (e).

Step 2) Calculate the distance of each data from every

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2}$$

centroid using equation

and group all the data based on minimum distance from each cluster.

Step 3) Set the clusters according to the group values and update the centroids according to their cluster members.

Step 4) Continue until the difference between the old and newly updated cluster is less than threshold (e).

3.2 Fuzzy C-Means

Step 1) Set the number of clusters and Initialize the centroid for each cluster.

Step 2) Calculate the distance of each data from every

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2}$$

centroid using equation

and group all the data based on minimum dis-tance from each cluster.

Step 3) Update the partition matrix using Equa-tion

$$\mu_{ij}^{r-1} = \left(1 \Big/ \sum_{j=1}^{c} (d_{ik}^{r} / d_{jk}^{r})^{2/m-1} \right)$$

Step 4) Calculate group membership for each da-ta and update the centroids according to their cluster members using equation

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ij})^m}$$

Step 5) Stop if ||U (k+1)-U (k) ||<δ

3.3 Rough C-Means

Step 1) Set the number of clusters and Initialize the centroid for each cluster.

Step 2) Calculate the distance of each data from every centroid using equation

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2}$$

and group all the data based on minimum distance from each cluster

Step 3) If (dij − dkj) ≤ δ and dij is minimum for 1 ≤ i ≤ c then then $\in \overline{A}(\beta_i)$ and $x_j \in \overline{A}(\beta_k)$ is minimum for 1 ≤ i ≤ c. In addition, by proper-ties of rough

$x_j \in \overline{A}(\beta_i)$

sets

Step 4) Calculate the new centroids using equa-tion

$$v_i = \begin{cases} w \times \mathcal{A} + \tilde{w} \times \mathcal{B} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) \neq \emptyset \\ \mathcal{A} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) = \emptyset \\ \mathcal{B} & \text{if } \underline{A}(\beta_i) = \emptyset, B(\beta_i) \neq \emptyset \end{cases}$$

$$\mathcal{A} = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j; \text{ and } \mathcal{B} = \frac{1}{|B(\beta_i)|} \sum_{x_j \in B(\beta_i)} x_j$$

3.4 Rough Fuzzy C-Means

Step 1) Set the number of clusters and Initialize the centroid for each clusters.

Step 2) Calculate the distance of each data from every centroid using equation

ion
$$D_{ij} = \left(\sum_{j=1}^{m}(x_{kj}-v_{ij})^2\right)^{1/2}$$

and group all the data based on minimum distance from each cluster.

Step 3) Choose values for fuzzifier m_1 sholds ϵ and Set iteration δ. counter t = 1.

Step 4) Compute μ_{ij} by equation

$$\mu_{ij} = (\sum_{k=1}^{c}(\frac{d_{ij}}{d_{kj}})^{\frac{2}{m_1-1}})^{-1};\ d_{ij}^2 = ||x_j - v_i||^2;$$

$$\text{subject to } \sum_{i=1}^{c}\mu_{ij} = 1, \forall j, 0 < \sum_{j=1}^{n}\mu_{ij} < n, \forall i.$$

Step 5) If μ_{ij} are μ_{kj} the two highest member-ships of x_j and $(\mu_{ij} - \mu_{kj}) \leq \delta$ then $x_j \in \overline{A}(\beta_i)$ A And $x_j \in \overline{A}(\beta_k)$ else $x_j \in \underline{A}(\beta_i)$ .In ad-dition. $x_j \in \overline{A}(\beta_i)$

Step 6) Modify μ_{ij} considering lower and boun-dary regions for c clusters and n objects.

Step 7) Compute new centroids using equation

$$v_i^{RF} = \begin{cases} w \times C_1 + \tilde{w} \times D_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) \neq \emptyset \\ C_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) = \emptyset \\ D_1 & \text{if } \underline{A}(\beta_i) = \emptyset, B(\beta_i) \neq \emptyset \end{cases}$$

$$C_i = \frac{1}{|\underline{A}(\beta_i)|}\sum_{x_j \in \underline{A}(\beta_i)} x_j; \text{ and } D_i = \frac{1}{n_i}\sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j; \text{ where } n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$$

Step 8) Increment t until $|\mu_{ij}(t) - \mu_{ij}(t-1)| > \epsilon$

*** $$\delta = \frac{1}{n}\sum_{j=1}^{n}(\mu_{ij} - \mu_{kj})$$

IV RESULT

Following are the results of the algorithms (Table 1) applied on standard data set IRIS

Table 1: Result comparison of different clustering methods

Methods	Centroid 1	Centroid 2	Centroid 3
Initial	5.1, 3.5, 1.4, 0.2	4.9, 3.0, 1.4, 0.2	4.7, 3.2, 1.3, 0.2
K-Means	6.7, 3, 5.7, 2.1	5.9, 2.8, 2.5, 1.4	5, 3.4, 1.5, 0.2
FCM	6.77, 3.05, 5.65, 2.05	5.89, 2.76, 4.36,1.40	5,3.41,1.48,0.25
RCM	6.85,3.08, 5.75,2.09	5.85,2.74, 4.35,1.41	5.01,3.42 ,1.46,0.24
RFCM	6.17, 2.84, 4.79, 1.62	5.47, 2.48, 3.74, 1.16	4.99, 3.34, 1.58, 0.29

V DISCUSSIONS

The given above table specifies the result of apply-ing the above mentioned algorithms on standard data set IRIS. Initially there are 3 centroids, which have been chosen. Each centroid holds the initial value. Values 5.1, 3.5, 1.4, 0.2 is assigned to centroid 1.Centroid 2 is assigned to values 4.9, 3.0, 1.4, 0.2 initially and values 4.7, 3.2, 1.3, 0.2 is assigned to centroid 3 initially. When we apply the K Means al-gorithm on the data set IRIS we get following results - centroid 1 = 6.7, 3, 5.7, 2.1, centroid 2 =5.9, 2.8, 2.5, 1.4, centroid 3 = 5, 3.4, 1.5, 0.2 . By applying the Fuzzy C-Means algorithm on the data set IRIS we get following results - centroid 1 = 6.77, 3.05, 5.65, 2.05, centroid 2 =5.89, 2.76, 4.36,1.40 , centro-id 3 = 5,3.41,1.48,0.25 . By applying the Rough C- Means algorithm on the data set IRIS we get follow-ing results - centroid 1 =6.85,3.08,5.75,2.09 , cen-troid 2 = 5.85,2.74,4.35,1.41,centroid 3 = 5.01,3.42,1.46,0.24 .Applying the Fuzzy C-Means algorithm on the same data set IRIS we get follow-ing results - centroid 1 = 6.17, 2.84, 4.79, 1.62, cen-troid 2 =5.47, 2.48, 3.74, 1.16, centroid 3 = 4.99, 3.34, 1.58, 0.29.

VI CONCLUSION

A comparative study between K –means, Fuzzy c-means, and Rough c-means and Rough fuzzy c-means algorithms is made in this paper. K-Means partitioning based clustering algorithm requires the number of ultimate cluster (k) beforehand. This type of algorithms also has some problems Susceptibility to local optima, sensitivity to outliers, memory space and iteration steps that are required to form the clusters are unknown. From the above results we can decide that K-means algorithm is slightly better than FCM algorithm but FCM algorithm requires more computation time than K-means. FCM clustering are best suited to handle the situations like to understand ability of patterns, incomplete or noisy data, mixed media information, human interaction and it can provide approximate solutions faster and also gives membership values of each cluster to every data. On the other hand this paper lies in developing a hybrid methodology, which integrates judiciously rough c-means and rough fuzzy c-means algorithm. These algorithms are generated to maximize the utility of both rough sets and fuzzy sets .The performance of RFCM algorithm is compared extensively with that of different c-means algorithms. The effectiveness of the algorithms is demonstrated with a comparison with other related algorithms, on a standard data set.

REFERENCES

[1] V. S. Rao and Dr. S. Vidyavathi, "Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on IRIS Data", Indian Journal of Computer Science and Engineering, vol.1, no.2, 2010 pp. 145-151.

[2] Mitra, S., Banka, H., Pedrycz, W, "Rough-Fuzzy Collaborative Clustering", IEEE Transactions on Systems, Man, and Cybernetics, Part B, Cybernetics, 36, 2006, 795–805.

[3] Soumi Ghosh, Sanjay Kumar Dubey, "Compara-tive Analysis of K-Means and Fuzzy C-Means Algo-rithms", (IJACSA) International Journal of Ad-vanced Computer Science and Applications, Vol. 4, No.4, 2013.

[4] S.Revathy, B.Parvathavarthini, "Comparison of FCM and RFCM On IRIS Dataset using MAT-LAB".

[5]Alan Jose, S. Ravi and M. Sambath, "Brain Tu-mor Segmentation using K -means Clustering and Fuzzy C-means Algorithm and its Area Calcula-tion". International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, issue 2, March (2014).

[6] Pallavi Purohit and Ritesh Joshi, "A New Effi-cient Approach Towards K-means Clustering Algo-rithm", International Journal of Computer Applica-tions, (0975-8887), vol. 65, no. 11, March (2013).

[8] Soumi Ghosh and Sanjay Kumar Dubey, Amity university, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", (IJACSA) Internation-al Journal of Advanced Computer Science and Ap-plications, Vol. 4, No.4, 2013

[9] Mr. Manish Mahajan, "Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm", Tejwant Singh1, Volume 4, Issue 5,May 2014

[10] Maji, P., Pal, S. K. "Rough-Fuzzy C-Medoids Algorithm and Selection of Bio-Basis for Amino Acid Sequence Analysis", IEEE Transactions on Knowledge and Data Engineering, 19(6), 2007, 859–872.

[11] Agrawal Sresht, Tripathy B.K, "Rough Fuzzy C-means Clustering Algorithm using Decision Theoretic Rough Set", accepted for presentation in ICRCICN_2015.

[12] Bhrgava Rohan, Tripathy B.K., Tripathy Anurag, "Rough Intuitionistic Fuzzy C-means Algorithm and A Comparative Analysis", 2015, 6th ACM India Computing Convention.