A Refined Rough Fuzzy Clustering Algorithm

Sahil Sobti[1], Vivek Shah[2]
School of Computing Sciences and Engineering
VIT University
Vellore, India
sahilsobti92@gmail.com[1], vivek.shah93@gmail.com[2]

B.K. Tripathy, Senior Member IEEE
School of Computing Sciences and Engineering
VIT University
Vellore, India
tripathybk@vit.ac.in

***Abstract*— Clustering is a familiar concept in the realm of Data mining and has wide applications in areas like image processing, pattern recognition and rule generation. Uncertainty in present day databases is a common feature. In order to handle these datasets, several clustering algorithms have been formulated in the literature. The first one being the Fuzzy C-Means (FCM) algorithm and it was followed by the Rough C-Means (RCM) by Lingras. In the paper Lingras has refined his previous algorithm. We combine this algorithm with the fuzzy C-means algorithm to generate a rough fuzzy C-Means (RFCM) algorithm in this paper. Also, we provide a comparative analysis with earlier RFCM algorithm introduced by Mitra et al and establish that our algorithm performs better. We use both numeric as well as image datasets as input and use the performance indices DB and D for this purpose.**

Keywords—Fuzzy set, Rough Set, Clustering, DB-index, D-index

I. INTRODUCTION

Clustering is a process used to agglomerate data which is similar to each other and widen the gap between the dissimilar data [1]. The outcome of this process are clusters of data, resembling their own members and distinguishing themselves from other clusters. The process of clustering has been done since ages using various techniques such as C mean, C nearest neighbors et cetera. But among all the unsupervised techniques used to find the clusters, only C means is widely used for research purposes. The fundamental principle of the objective function of C means is to make clusters such that the inter cluster distance can be minimized and the intra cluster distance can be widened with every iteration [1], [2].

Most of the time, due to ambiguity in the dataset, it is difficult to get a crisp boundary around such clusters and hence rough sets and fuzzy sets are applied to tackle such ambiguity, vagueness and uncertainty in the data.

Rough set is an interval set which helps in assigning data to lower and upper approximation of the clusters generated during the clustering process based on the parameters assigned to these approximations and the threshold value. Also, Fuzzy set proves to be of great significance when it comes to handling ambiguity in the data. It provides the membership value which helps to make better decisions to control the uncertainty in the data [3], [4], [5], [6], [7].

In this paper, we have proposed a refined rough fuzzy C-Means (RFCM) algorithm, which extends the rough C means algorithm by Lingras [8]. The rough and fuzzy combination has proved fruitful to handle the uncertainty in the data [1], [9], [10]. While the fuzzy membership values enable to segregate the overlapping clusters effectively, the Rough sets help in improving this segregation process based on these membership values [3], [4], [5], [6], [7]. Recently some algorithms have been devised to replace Euclidean distance by kernel function [11], [12]. In [13], an application of clustering algorithms which analyzes satellite data is provided. Thus clustering can be extensively exploited to analyze complex scientific data. This paper contains the comparison between the earlier rough fuzzy C means algorithm and the improved rough fuzzy C means to prove the effectiveness of this novel approach. The Davies–Bouldin(DB) [14] and Dunn (*D*) [15] indexes have been used to do a quantitative analysis of the results and to compare them with the old algorithm. The algorithm is also tested on various medical MR images and cancer images to test the strength of the clustering algorithm.

II. BASIC DEFINITIONS AND NOTATIONS

This section consists of few basic definitions and notations which are to be used in the paper. The fuzzy set concept was devised by Zadeh in 1965 in [16].

Definition 2.1: A fuzzy set F from a universe U is defined through its membership function μ_F, which is defined as $\mu_F : U \to [0,1]$ such that for each $x \in U$, $\mu_F(x)$ is in the interval $[0,1]$.

The novel idea of rough sets was introduced by Pawlak [17] in the year 1982 and is defined as follows.

Definition 2.2: Let U be a universal set and E be an equivalence relation defined over U. Then E decomposes U into disjoint equivalence classes. The equivalence class of $x \in U$ is denoted by $[x]_E$. Then any subset Y of U is represented by two crisp sets called the lower and upper approximations of Y, denoted as $\underline{EY}$, $\overline{EY}$ respectively and are defined as

(1) $\underline{EY} = \bigcup \{x \in U \mid [x]_E \subseteq Y\}$

(2) $\overline{EY} = \bigcup \{x \in U \mid [x]_E \cap Y \neq \phi\}$

The boundary of Y represents the uncertainty region associated with Y , denoted as $BN_E(\mathrm{Y})$ and is defined as

(3) $$BN_E(\mathrm{Y}) = \overline{E}Y - \underline{E}Y$$

The set is said to be rough with respect to E if $EY \neq \overline{E}Y$ or equivalently $BN_E(\mathrm{Y}) \neq \phi$. Otherwise, Y is said to be E-definable.

The novel idea of rough fuzzy sets was introduced by Dubois and Prade in 1990 [10], in order to form a hybrid model of the two uncertain concepts. In fact they are the first to remove the wrong conception that rough sets and fuzzy sets are rival models.

III. PROPOSED ALGORITHM

In this proposed algorithm, we have extended the algorithm proposed by Lingras in [8] According to [8], the new cluster centers can be calculated using the lower and boundary approximations of each cluster center while the boundary is not empty and if this case is not true, then find the new clusters directly from the lower approximation which is having the weight equal to 1. Since this novel approach by Lingras, to deal with ambiguous data is only limited to rough set, we propose to implement the same , using modified cluster calculation by combining the effect of rough and fuzzy in C means clustering.

The algorithm starts by initialization of random data points in the dataset and then assigning them as the initial cluster centers v_i where i = 1 to C. Based on these cluster centers, Euclidean distances are calculated between every cluster centers and the data set to measure the similarity between them. The distances so obtained are then used in calculating the membership values for each dataset in the cluster centers. After this step, we examine each data point separately. For each data point, we calculate the difference of its highest membership value and the next highest membership value in the two different cluster centers. If the difference of these membership values is greater than some threshold λ, then assign the data point to lower of the highest membership value cluster center. If the condition is not true, then assign the same data point to the boundary of both the clusters centers in which it has the highest and the next highest membership value. Perform the same procedure for all the data points. After the assignment process is completed, the new cluster center values are computed by refined rough fuzzy c means given by (5). The whole procedure is repeated until the difference of the membership values of each data point in the cluster center in the two consecutive iterations is less than some threshold δ.

The detailed novel refined rough fuzzy Algorithm is explained as follow:

A. The RFCM Algorithm

1. Select some random points v_i in the data set and assign them as the initial *C* cluster center
2. Compute the distance between each cluster center and the data points in the n- dimensional space
3. The distance is then used to find the membership values μ_{ik} of the data point in the cluster center *C* using (4)

(4) $$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

We take m = 2 for the implementation purpose as is done in the literature.

4. If μ_{ik} is the highest membership value and μ_{jk} is the next highest membership value then the difference = $|\mu_{ik} - \mu_{jk}|$

   If difference > λ

   Assign the data X_k to $\underline{BU_i}$

   Else

   Assign the data X_k to $\overline{BU_i}$ and $\overline{BU_j}$ of both the clusters.
5. Determine the new cluster centers using the refined rough fuzzy c means given by (5)
6. If for the t^{th} and $(t-1)^{th}$ iteration $|\mu(t) - \mu(t-1)|$ for final clusters is less than δ, then stop. Otherwise, repeat steps 2 to 6.

Due to the random initialization in the first step, the algorithm tends to give different results in different iterations. The values of weights for the upper and lower approximation are chosen based on the importance of approximations. Similarly, the λ in step 4 is also selected randomly such that it tackles all the ambiguity in the data. The more the λ is, the more is the possibility of data in the boundary of the cluster[3].

δ is another threshold parameter that needs to be decided based on the user preference. It governs the stopping condition for the algorithm. So, it should be as low as possible. In this paper we have taken it to be 0.03.

Also the membership value calculated in step3 of the aforementioned algorithm, should take the value equal to 1 in case of lower approximations to denote the certainty associated with all the data in that region.

(5)
$$V_i = \begin{cases} W_{low}\dfrac{\sum_{X_k \in \underline{BU_i}} \mu_{ik}^m X_k}{\sum_{X_k \in \underline{BU_i}} \mu_{ik}^m} + W_{up}\dfrac{\sum_{X_k \in (\overline{BU_i} - \underline{BU_i})} \mu_{ik}^m X_k}{\sum_{X_k \in (\overline{BU_i} - \underline{BU_i})} \mu_{ik}^m}, & (\overline{BU_i} - \underline{BU_i}) \neq \phi; \\ \dfrac{\sum_{X_k \in \underline{BU_i}} \mu_{ik}^m X_k}{\sum_{X_k \in \underline{BU_i}} \mu_{ik}^m}, & \text{otherwise} \end{cases}$$

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The aforementioned algorithm is tested and compared on various data sets to check the strength of the modifications introduced in the algorithm.
The following data sets are used to compare the results with the old rough fuzzy c means algorithm.

A. Numeric Datasets

We have taken three datasets from the UCI repository for the computation and the comparison purpose of the algorithms.

1) IRIS DATASET

The Iris dataset is a small dataset of size 150 having 4 attributes and 1 class label showing which iris plant it is.The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant [18]. In order to do a comparative analysis with the [1] RFCM, we have partitioned the total set of data into 2 randomly selected equal parts of size 75 each. Then the modified RFCM is applied on both Part A and Part B iris dataset and then compared based on the the Davies–Bouldin (DB) and Dunn (D) indexes [14], [15]. Table 1 demonstrates the results acquired by formal algorithm and modified algorithm for the cluster centers = 3. The λ is taken as 0.1 and $W_{low} = 0.9$ for carrying out the comparative study.

TABLE I: Comparison of DB and D values of existing and refined RFCM

Algorithm/Dataset	**Old RFCM**		**New RFCM**	
	DB	D	DB	D
RFCM/A	0.63	2.63	0.5585	2.9022
RFCM/B	1.97	0.83	0.669	2.265

The results in Table 1 clearly indicates the efficiency of the modified RFCM over the formal algorithm. The observable decrease in DB and increase in D index for RFCM/B is a clear indication that the new RFCM is more efficient and scalable.

2) SOYABEAN DATASET

In this section, we have applied the modified RFCM algorithm and compared it with Soybean dataset [18]. This multivariate dataset consists of 35 attributes and 47 samples. To compare the performance of the old RFCM given in [19] with the new one, for the same dataset, the λ is taken as 0.4, and $W_{low} = 0.9$. Table 2 illustrates the values of DB and D indexes for both the algorithms with center = 4.

TABLE II: Comparison of DB and D values for Soybean data

Algorithm	**Dunn index**	**D index**
Old RFCM	2.37	0.66
New RFCM	1.8317	0.6345

3) ZOO DATA

The performance of the modified RFCM is also tested on various values of cluster centers for the Zoo dataset [18] and the results are shown in Table 3.

TABLE III: Comparison of DB and D values for different number of clusters

Number of Clusters	**DB Value**	**D Value**	**No. of Iterations**
2	1.73	0.806	4
3	1.548	0.952	6
4	1.22	1.01	8
5	1.285	1.017	6
6	1.107	0.968	7
7	1.209	1.257	8
8	1.088	0.52	5
9	1.003	0.6303	5
10	1.04	0.491	8

The zoo data set contains 17 attributes and 101 samples. Since, the first attribute is the name of the animal, it is omitted while using the dataset for clustering. There are 7 class types assigned to the zoo dataset.
It is evident from the DB and D values obtained in Table 3 that the values reach their best results for no. of clusters = 7.

B. Image Datasets

In this section we take two different images, the brain MRI image and the Leukemia cancer image for the computation and comparison of the algorithms.

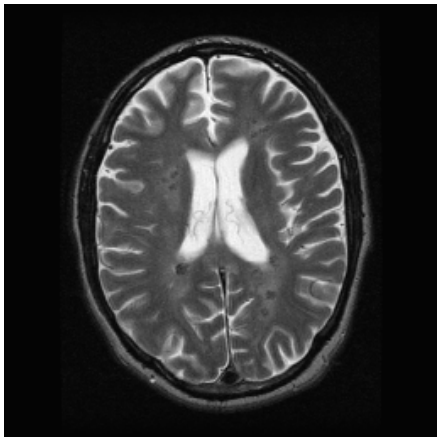

Fig. 1. Original Brain MRI Image

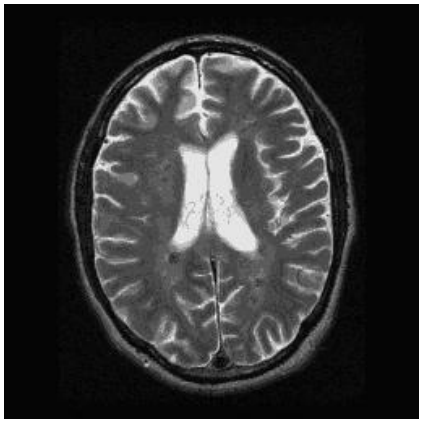

Fig. 2. Image reconstructed using Refined Rough Fuzzy C-Means

As we can see clearly in Fig. 1 and Fig. 2, after applying the algorithm with no. of clusters as 10, exact representation of the image has been reconstructed. In the middle, the faint lines have also become distinct which will help doctors for better analysis of the image.

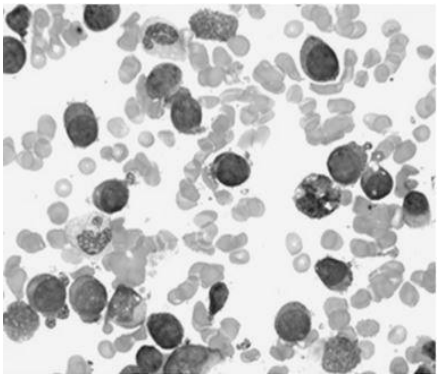

Fig. 3. Leukemia Cell Image

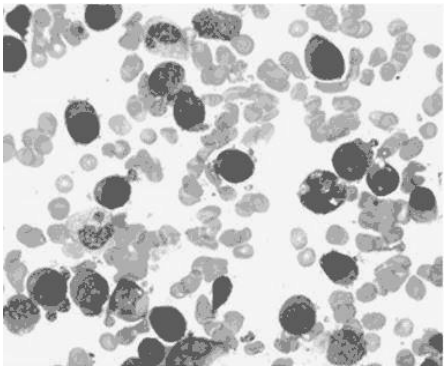

Fig. 4. Image extracted using Refined Rough Fuzzy C- Means

The above image in Fig. 3 is of 382 × 303 pixels with 8 bit gray scales. Therefore the number of pixel objects in the image is 115,746. The threshold values δ and λ are equal to 0.03. After applying the algorithm on the leukemia image, we can see that distinct features of the image have been highlighted and are seen clearly in Fig. 4 which will help in decision making in the field of bioinformatics.

CONCLUSION

Data clustering is the focus of this paper and we have introduced an imprecise hybrid algorithm called the RFCM, which is deviated from the earlier versions of this algorithms in the sense that the definition of the centre is taken in three components in the earlier algorithms whereas we have taken here the original two step computation of the cluster centres. Also, we have made random selection of the first values of the cluster centres, which increases its efficiency in terms of handling large datasets. We have made experimental verification of the efficiency of our algorithm with the existing one and for this purpose two types of datasets, namely numeric and image datasets have been used and also two measures of indices, the DB and the D index have been taken.

REFERENCES

[1] Mitra, S., Banka, H. and Pedrycz, W., "Rough-Fuzzy Collaborative Clustering, System, Man, and Cybernetics", Part B: Cybernetics, IEEE Transactions on 36.4,2006, pp.795-805.

[2] Ruspini, Enrique H., "A new approach to clustering", Information and control, 15, 1, 1969, pp. 22-32

[3] Maji, P. and Pal, S.K., "RFCM: A Hybrid Clustering Algorithm using rough and fuzzy set", Fundamenta Informaticae 80.4, 2007, pp.475-496.

[4] Bezdek, J.C., "Pattern Recognition with Fuzzy Objective Function Algorithms", Kluwer Academic Publishers, 1981.

[5] Atanassov, K.T., "Intuitionistic Fuzzy Sets", Fuzzy sets and Systems 20,1, 1986, pp.87-96.

[6] Klir G. J. and Yuan B. 2002, "Fuzzy sets and fuzzy logic theory and applications", 1997, Prentice Hall of India Private Limited New Delhi.

[7] Yen J. and Langari, R., "Fuzzy Logic Intelligence, Control and Information, 1999, Pearson Education, Inc.

[8] Lingras, P. and West, C., "Interval set clustering of web users with rough k-mean", Journal of Intelligent Information Systems, 23(1), 2004, pp.5-16

[9] Maji, P. and Paul, S., "Microarray Time-Series Data Clustering using Rough-Fuzzy C-means Algorithm", in Proc. 5th IEEE Intl. Conf. on Bioinformatics and Biomedicine, 2011, pp. 269-272.

[10] Dubois, D. and Prade, H. ," Rough fuzzy sets model", International journal of General Systems, vol. 46, no.1, 1990, pp. 191 – 208.

[11] Tripathy, B. K., et al.,"On Kernel Based Rough Intuitionistic Fuzzy C-means Algorithm and a Comparative Analysis.", Advanced Computing, Networking and Informatics,vol 1. Springer International Publishing, 2014,pp.349-359.

[12] Tripathy, B.K. and Bhargav, R., "Kernel Based Rough-Fuzzy C-Means, PReMI, ISI Calcutta, December, LNCS 8251, 2013, pp.148-157

[13] Purushotham, S., & Tripathy, B.,"A comparative study of RIFCM with other related algorithms from their suitability in analysis of satellite images using other supporting techniques", Kybernetes, vol. 43(1),2014,pp.53-81.

[14] Davis, D. L. and Bouldin, D.W., "A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence", vol.PAMI-1, no.2, 1979, pp.224 – 227.

[15] Dunn, J. C., "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", 1973, pp. 32-57.

[16] Zadeh, Lotfi A.,"Fuzzy sets", Information and control,vol 8.3,1965, pp.338-353

[17] Pawlak, Z., "Rough sets", Int. jour. of Computer and Information Sciences, 11, 1982, pp.341-356.

[18] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/

[19] Bhargava, Rohan, et al. "Rough intuitionistic fuzzy C-means algorithm and a comparative analysis", Proceedings of the 6th ACM India Computing Convention. ACM, 2013, pp. 23-33.