ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Paneer and Tripathy.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

TUNED HYBRID SOFT CLUSTERING ALGORITHM FOR UNCERTAIN INFORMATION SYSTEM

Prabhavathy Paneer[1*] and Balakrishna Tripathy[2]

[1]*School of Information Technology and Engineering, VIT University, Vellore, INDIA*
[2]*School of Computing Science and Engineering, VIT University, Vellore, INDIA*

ABSTRACT

Clustering is an important mission in the field of machine learning, pattern recognition and web mining. Handling uncertain data in the information system is one of the key research topics in the vicinity of knowledge representation. Number of clustering algorithms are available [23][6][12]27]; but many of those algorithms are challenging when dealing with uncertain data. The aim of the paper is to tune two existing rough c-means and fuzzy c-means and integrate them into a tuned hybrid soft clustering algorithm termed as the tuned rough-fuzzy c-means algorithm. Rough c-means is extremely sensitive to the initial placement of the cluster centers. The proposed algorithm is enhanced by introducing dynamic centroid computation. The proposed algorithm performance is compared with the existing rough c means, fuzzy c-means, and rough fuzzy c –means approaches. The effectiveness of the algorithm is verified on real and synthetic datasets.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 16th –Aug-2015

KEY WORDS

Clustering; Uncertain; RoughSet; Fuzzy; C-means

***Corresponding author: Email:** pprabhavathy@vit.ac.in, **Tel.:** +91-9486236259; **Fax:** +91-416-2243092

INTRODUCTION

Cluster analysis [1] is a technique for finding natural groups present in the data. It divides a given data set into a set of clusters in such a way that two objects from the same cluster are as similar as possible and the objects from different clusters are as dissimilar as possible. Clustering techniques have been effectively applied to a wide range of engineering and scientific disciplines such as pattern recognition, machine learning, psychology, biology, medicine, computer vision, communications, and remote sensing. A number of clustering algorithms have been proposed to suit different requirements. Clustering is categorized as hard or soft in nature .Soft clusters may have fuzzy or rough boundaries. In hard clusters, the elements which are similar to each other are placed in the same cluster. The elements whose natures differ with each other drastically are placed in different clusters. Soft clustering [12] helps researchers to discover overlapping clusters in many applications such as web mining and text mining. Hence soft clusters may have two types of boundaries 1) Fuzzy boundary 2) Rough boundary. Fuzzy clusters need an association degree to distinguish each element present in the cluster. The elements in the rough clusters are distinguished with the help of boundary region. The relations between rough sets and fuzzy sets were compared [2, 4]. On the whole, both theories deal with the difficulty of information granulation: the theory of fuzzy sets is centered upon fuzzy information granulation, where as rough set theory is paying attention on crisp information granulation.

Data generation methods create uncertain, incompleteness, and granularity in information system which provides inaccurate result in data analysis. Rough set theory is a valuable tool for data mining. In the past few years the concept of basic rough sets has been extended in many different directions. The original rough set theory proposed by Pawlak [18-20] is based upon equivalence relations defined over a universe. It is the simplest formalization of indiscernibility. However, it cannot deal with a number of uncertain problems in real information systems. This has direct to numerous significant and motivating extensions of the original concept. Bezdek's fuzzy c-means [10, 11] is the another most popular soft clustering algorithm for many real life applications in a very diverse range of domains. K-means is one of the most extensively used partitioned based clustering algorithms and it is extremely sensitive to the initial placement of the cluster centers. Numerous initialization methods have been proposed [15] to deal with this problem. Efficient hybrid evolutionary data clustering algorithm K-MCI [9] has been presented to handle high dimensional data and large cluster. Fuzzy clustering is suitable to classify ordered sequences in human activity pattern analysis [22]. However, the majority of the present fuzzy clustering modules [3, 4, 16] packaged in both open source and commercial products

THE IIOAB JOURNAL
ISSN: 0976-3104

have lack of enabling users to explore fuzzy clusters extremely and visually in terms of examination of different relations among clusters.

Attribute weighted fuzzy clustering has became a very active area of research and interval number has been introduced for attribute weighting in the weighted fuzzy c-means (WFCM) clustering approach [13, 25]. The existing fuzzy and rough clustering approaches have been refined based on the concept of shadowed sets. Shadowed clustering [26] has been presented which serves as a conceptual and algorithmic bridge between the FCM and RCM. Much work has been carried out using rough c-means, fuzzy c-means and rough fuzzy c-means in data clustering. The extensive survey of the significant extensions and derivatives of soft clustering approaches have been studied [5, 7]. In this paper, tuned rough fuzzy c-means clustering approach is proposed to resolve the uncertainty of information system.

This paper focuses on traditional rough fuzzy c-means and tuned rough fuzzy c-means approaches for handling uncertainty presents in the information system. The remainder of this paper is organized as follows. The introduction about the work is discussed in section 1. In Section 2, traditional soft clustering algorithms are discussed under materials and methods. Section 3 investigates experimental analysis of tuned soft clustering algorithm for uncertain data. Section 5 discusses performances of the proposed algorithm and this paper concludes in section 6.

MATERIALS AND METHODS

Traditional soft clustering

Fuzzy sets and roughsets [27-31] were incorporated in the c-means framework to develop the fuzzy c-means (FCM) [10], rough c-means [6, 8, 12, 23, 24] and rough–fuzzy c-means (RFCM) [21, 23] algorithms, respectively. While membership in FCM enables efficient handling of overlapping partitions, the roughest [17, 19] deal with uncertainty, vagueness and incompleteness of data in terms of upper and lower approximations.

Rough C-means

Rough c-means algorithm was introduced by Lingras, which describes a cluster by its centroid and its lower and upper approximations. In rough c- means, an object can belong completely in one cluster or can be in the uncertainty region or boundary of two clusters. The lower and upper approximations are weighted differently. In each iteration step of the algorithm, the distance of objects from the cluster centroids are computed and if the difference between the two lowest distances is less than a specified threshold value the element is placed in the boundary of the two clusters. Otherwise, the element is placed in the cluster for which the distance is the minimum.

Fuzzy C-means

Developed by Bezdek, the fuzzy C-means algorithm is a powerful method to classify fuzzy data by using the concept of objective function. This approach which minimizes the objective function is expressed in the form of an iterative algorithm makes it possible to reach at an optimal solution, where the solution space is of infinite cardinality. In fuzzy c-means data may belong to one or more than one clusters. It brings in the concept of having membership values. Each object will have a membership in every cluster; which represents the degree to which the element belongs to the cluster. So, here also the clusters are not disjoint. The multiple membership of data models uncertainty of elements belonging to clusters.

Rough-Fuzzy C-Means

It combines the concepts of rough set theory and fuzzy set theory. It has been established that the rough membership function is more general than the fuzzy membership function. However, this generalized membership function has some costs to pay as it does not provide a formula to find the membership values for union and intersection of rough sets. However, in fuzzy set theory we have definite formulae for the computation of the membership values. Thus the hybrid algorithms takes care of both the features by providing membership values to elements as well as modeling vagueness in data through the boundary concept. The concepts of lower and upper approximations in rough set deals with uncertainty and, vagueness whereas the concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters.

According to rough set theory if $x_j \in \underline{B}Ui$ then object x_j is contained completely in cluster U_i and if then object x_j belongs to cluster U_i and also belongs to another cluster. Hence according to fuzzy set theory the objects in boundary approximation should have different degree of membership on the clusters. So in RFCM the membership values of objects in lower approximation are μ_{ij} = 1 while for those in boundary region are determined by the membership values.

THE IIOAB JOURNAL ISSN: 0976-3104

1. Assign initial means v_i, i=1, 2, 3,…., c. Choose values for fuzzifier m_1 and threshold ε and δ. Set iteration counter t=1.
2. Compute membership μ_{ij} by equation (1) for c clusters and n objects.
3. If μ_{ij} and μ_{ik} be the two highest membership value of x_j and $(\mu_{ij} - \mu_{kj}) \leq \delta$, then $x_j \in \overline{A}(\beta_i)$ and $x_j \in \overline{A}(\beta_k)$. Furthermore, x_j is not part of any lower bound.
4. Otherwise, $x_j \in \underline{A}(\beta_i)$. In addition, by properties of rough sets, $x_j \in \overline{A}(\beta_i)$.
5. Modify μ_{ij} considering lower and boundary regions for c clusters and n objects.
6. Compute new centroid as per equation (1).
7. Repeat steps 2 to 7, by incrementing t, until $|\mu_{ij}(t-1) - \mu_{ij}(t)| > \varepsilon$

$$v_i = \begin{cases} w \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$$

$$C_1 = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{1}{n_i} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j \quad where \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$$

$|\underline{A}(\beta_i)|$ represents the cardinality of $\underline{A}(\beta_i)$. 0 < w < w < 1

Tuned soft clustering

In particular the recent promising developments in the fusion of soft cluster algorithms show the need for approaches that holistically address uncertainty. Hence, soft clustering will continue in the attention of researchers and most likely attract yet more practitioners in the ground of data mining in support of their real life applications. The objective of this paper is to analyze Georg Peters [6] cluster algorithm rigorously and point out potential for further development. Based on the analyze we have presented a tuned rough fuzzy cluster algorithm and apply it to synthetic and real time market data.

Tuned rough C-means [6]

Lingras et.a[12] discussed rough clustering algorithm. Georg Peters evaluated Lingras et al. rough cluster algorithm and recommended some alternative solutions. This led to the new refined rough k-means algorithm.

Georg Peters cluster rough cluster algorithm goes as follows:

(a) Initialization: Randomly assign each data object to exactly one lower approximation. Hence, the data object will also belong to the upper approximation of the same cluster.

(b) Calculation of the new means. The means are calculated as follows:

$$\overline{m}_k = \omega_l \frac{1}{\underline{X}_n \in \underline{C}_k} \sum \frac{\vec{X}_n}{|\underline{C}_k|} + \omega_u \frac{1}{\underline{X}_n \in \underline{C}_k} \sum \frac{\vec{X}_n}{|\overline{C}_k|}$$

With $\omega_l + \omega_u = 1$

Now, the lower approximation of each cluster always has at least one member. Therefore $|\underline{C}_k| \neq \varphi, \forall k$ and by definition $|\overline{C}_k| \neq \varphi, \forall k$

(c) (i) Assign the data objects to the approximations. Assign the data object that represents a cluster to its lower and upper approximation.

1. Find the minimal distance between cluster k and all data objects n and assign data object l to lower and upper

THE IIOAB JOURNAL ISSN: 0976-3104

approximation of cluster h:

$$d(\vec{X}_l, \vec{m}_h) = \min_{n,k} d(\vec{X}_n, \vec{m}_k) \Rightarrow \vec{X}_l \in \underline{C_k} \wedge \vec{X}_l \in \overline{C_k}$$

2. Exclude $\vec{X}_l$ and $\vec{m}_h$. If clusters are left – so far, in the above step (a) no data object has been assigned to them – go back to Step (a). Otherwise continue with Step (ii).

(ii) For each remaining data point $\vec{X}_m'$ (m=1, 2, ….., M, with M=N-K) determine its closest mean $\vec{m}_h$:

$$d_{m,h}^{\min} = d(\vec{X}_m', \vec{m}_h) = \min_{k=1,\ldots,K} d(\vec{X}_m', \vec{m}_k)$$

Assign $\vec{X}_m'$ to the upper approximation of cluster h.

(iii) Determine the mean $\vec{m}_t$ that are also close to $\vec{X}_m'$. Take the relative distance as defined above where ζ is a given relative threshold

$$T' = \left\{ t : \frac{d(\vec{X}_m', \vec{m}_k)}{d(\vec{X}_m', \vec{m}_h)} \leq \zeta \wedge h \neq k \right\}$$

If T'≠ ϕ ($\vec{x}_m'$ is also close to at least one other mean $\vec{m}_t$ besides $\vec{m}_h$).

Then $\vec{X}_m' \in \overline{C_t}, \forall t \in T'$.

Else $\vec{X}_m' \in \underline{C_h}$

(d) Check convergence for the algorithm. If the algorithm has not converged continue with step 2 else stop.

George peters refined rough c-means algorithm by replacing boundary into upper approximation in mean computation

Tuned rough fuzzy C-means

The algorithm as presented by Lingas et al. is numerical instable since there are data constellations where lower approximation is empty in some cases. The clusters will be weak if there is no representative the proposed algorithm ensures that each lower approximation has at least one member. It is implemented by assigning the data point that is closest to a mean to the lower approximation of the cluster. Otherwise the cluster seems to be weak since it has no sure representative. We have used relative distance represented by George peters instead of Lingras' et al. absolute distance measure to determine the set T. Rough C-means is one of the most extensively used partitioned based clustering algorithms and it is extremely sensitive to the initial placement of the cluster centers. Numerous initialization methods have been proposed [15] to deal with this problem. Here, we also addressed the solution for selection of cluster centers.

The tuned rough fuzzy c-means as follows:

Algorithm: Tuned Rough Fuzzy C-means

1 Assign initial means vi, i=1, 2, 3,…., c. Choose values for fuzzifier m_1 and threshold ε and δ. Set iteration counter t=1.

2 Compute membership μij by equation (2) for c clusters and n objects.

THE IIOAB JOURNAL ISSN: 0976-3104

3 If μij and μik be the two highest membership value of xj and (μij /μkj) ≤ δ, then xj $\in\overline{\underline{A}}(\beta_i)$ and xj $\in\overline{\underline{A}}(\beta_k)$. Furthermore, xj is not part of any lower bound.

4 Otherwise, xj $\in\underline{A}(\beta_i)$. In addition, by properties of rough sets, xj $\in\overline{\underline{A}}(\beta_i)$.

5 Modify μij considering lower and boundary regions for c clusters and n objects.

6 Compute new centroid as per equation (2).

7 Repeat steps 2 to 7, by incrementing t, until |μij(t-1)- μij(t)| > ε

$$v_i = \begin{cases} w \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$$

$$C_1 = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{\sum_{x_j \in \underline{A}(\beta_i)} x_j + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j}{n_i + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}} \in \quad where \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$$

$|\underline{A}(\beta_i)|$ represents the cardinality of $\underline{A}(\beta_i)$. $0 < \tilde{w} < w < 1$

Selection of initial centroid for rough fuzzy C-means algorithm

Step 1: From n objects calculate a point whose attribute values are average of n objects attribute values. Hence, first initial centroid is average on n- objects.
Step 2: Select next initial centroids from n-objects in such a way that the Euclidean distance of that object is maximum from other selected initial centroids.
Step 3: Repeat step 2 until we get k initial centroids.

From these steps the initial centroids are derived and tuned rough fuzzy c-means algorithm is tested for the dynamic centroids and random centroids.

RESULTS

Experimental analysis

The traditional soft clustering algorithms such as rough c-means(RCM), Fuzzy C-means (FCM), Rough-Fuzzy C-means(RFCM), Rough-Intuitionistic-fuzzy C-means (RIFCM) and proposed tuned rough fuzzy c-means (TRFCM) algorithms are implemented using Java.UCI Machine Learning Repository, Wholesale customers Data Set [29] is used to evaluate the performance of the above said algorithms. The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories. The centroid formulae for the algorithms are given in the Table 1. The traditional and tuned soft clustering algorithms are tested with random centroid selection and proposed centroid computation method and it's shown in Figure- 1.

Table: 1. Comparisons of Centroid formulae of the various soft clustering algorithms

Algorithm	Formula for Centroid calculation
RCM	$v_i = \begin{cases} w \times A + w \times B & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ A & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ B & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$ $A = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad B = \frac{1}{|B(\beta_i)|} \sum_{x_j \in B(\beta_i)} x_j$,

THE IIOAB JOURNAL
ISSN: 0976-3104

FCM	$v_i = \frac{1}{n_i} \sum_{j=1}^{n} (\mu_{ij})^{m_1} x_j ;\ where : n_i \cdot = \sum_{j=1}^{n} (\mu_{ij})^{m_1}$
RFCM	$v_i = \begin{cases} \tilde{w} \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$ $where \quad C_1 = \frac{1}{\lvert \underline{A}(\beta_i) \rvert} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{1}{n_i} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$,
RIFCM	$V_i = \begin{cases} w_{low} \frac{\sum_{x_k \in \underline{BU}_i} x_k}{\lvert \underline{BU}_i \rvert} + w_{up} \frac{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m x_k}{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m} & if \lvert \underline{BU}_i \rvert \neq \varphi \quad and \lvert BN(U_i) \rvert \neq \varphi \\ \frac{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m x_k}{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m} & f \lvert \underline{BU}_i \rvert = \varphi \quad and \lvert BN(U_i) \rvert \neq \varphi \\ \frac{\sum_{x_k \in \underline{BU}_i} x_k}{\lvert \underline{BU}_i \rvert} & ELSE \end{cases}$
Tuned RCM	$\bar{m}_k = \omega_l \sum_{\underline{X}_n \in \underline{C}_k} \frac{\vec{X}_n}{\lvert \underline{C}_k \rvert} + \omega_u \sum_{\underline{X}_n \in \overline{C}_k} \frac{\vec{X}_n}{\lvert \overline{C}_k \rvert}, \omega_l + \omega_u = 1$
Tuned RFCM	$v_i = \begin{cases} \tilde{w} \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$ $where\ C_1 = \frac{1}{\lvert \underline{A}(\beta_i) \rvert} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{\sum_{x_j \in \underline{A}(\beta_i)} x_j + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j}{n_i + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}} \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$,

THE IIOAB JOURNAL
ISSN: 0976-3104

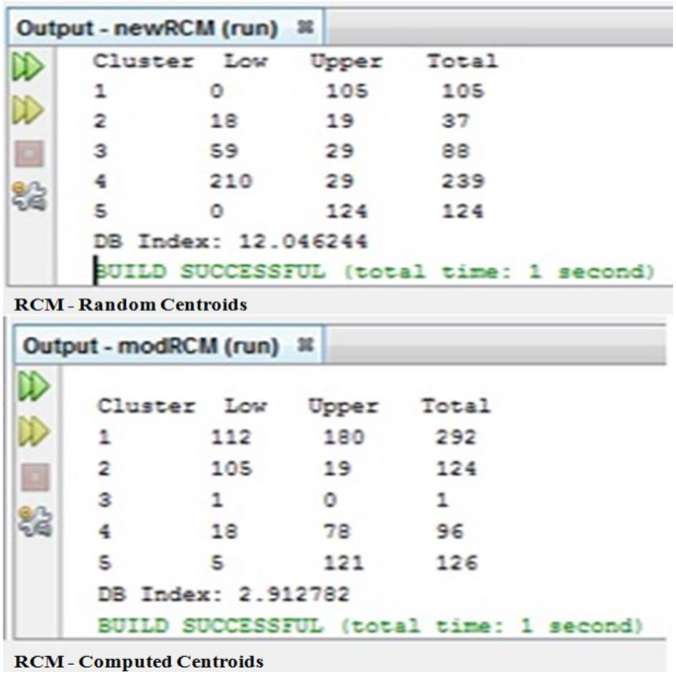


Fig: 1. a

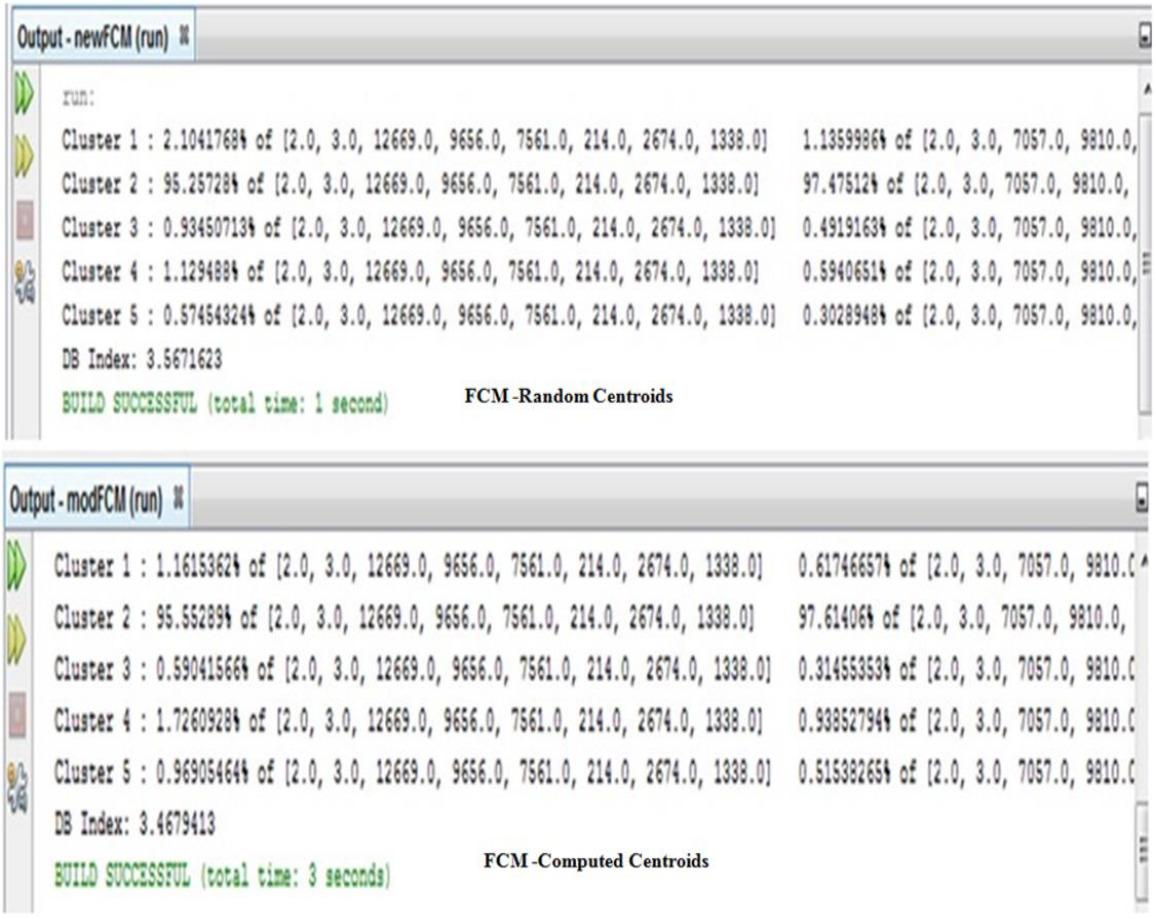


Fig: 1. b

THE IIOAB JOURNAL
ISSN: 0976-3104

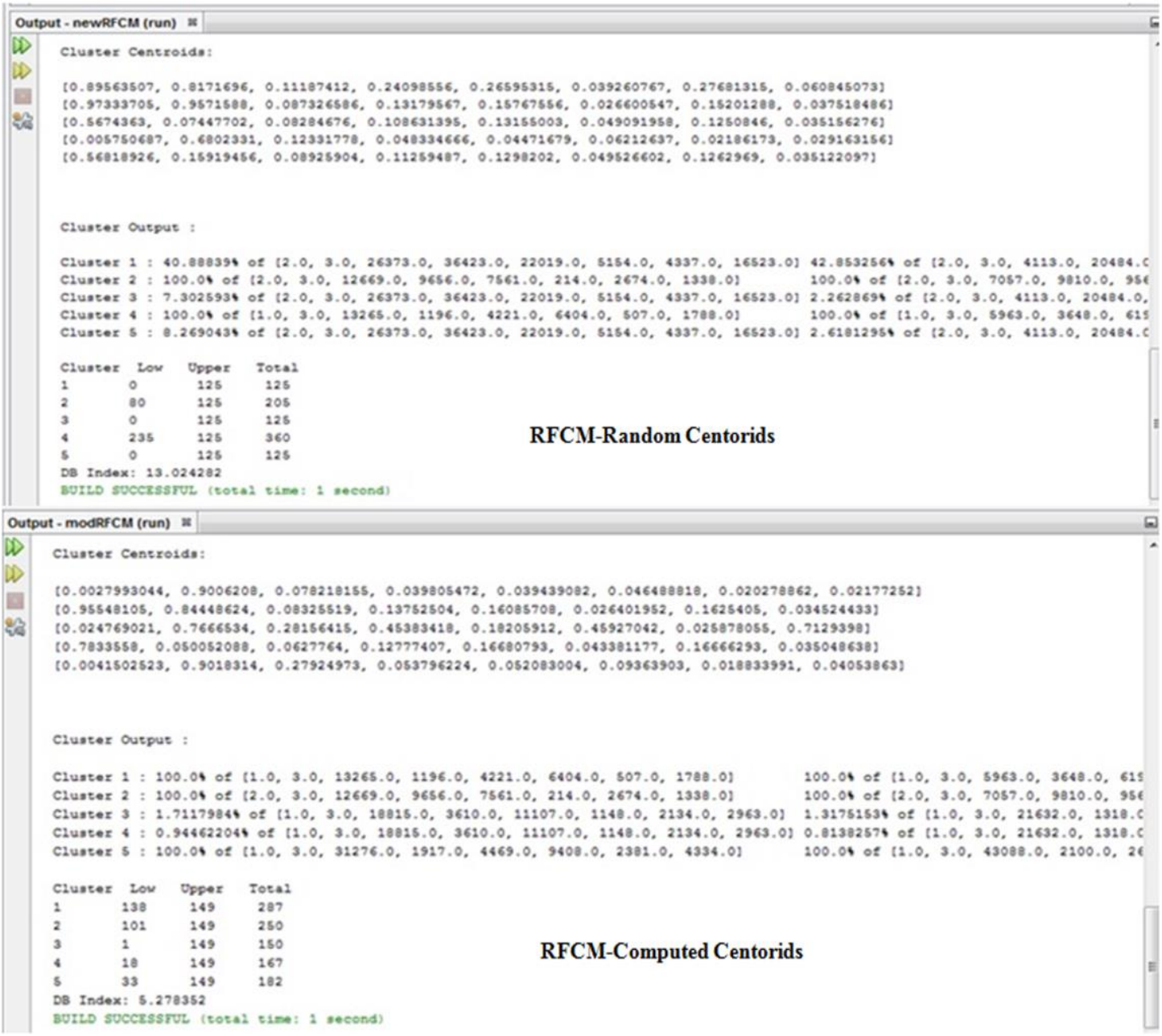


Fig: 1. c

THE IIOAB JOURNAL
ISSN: 0976-3104

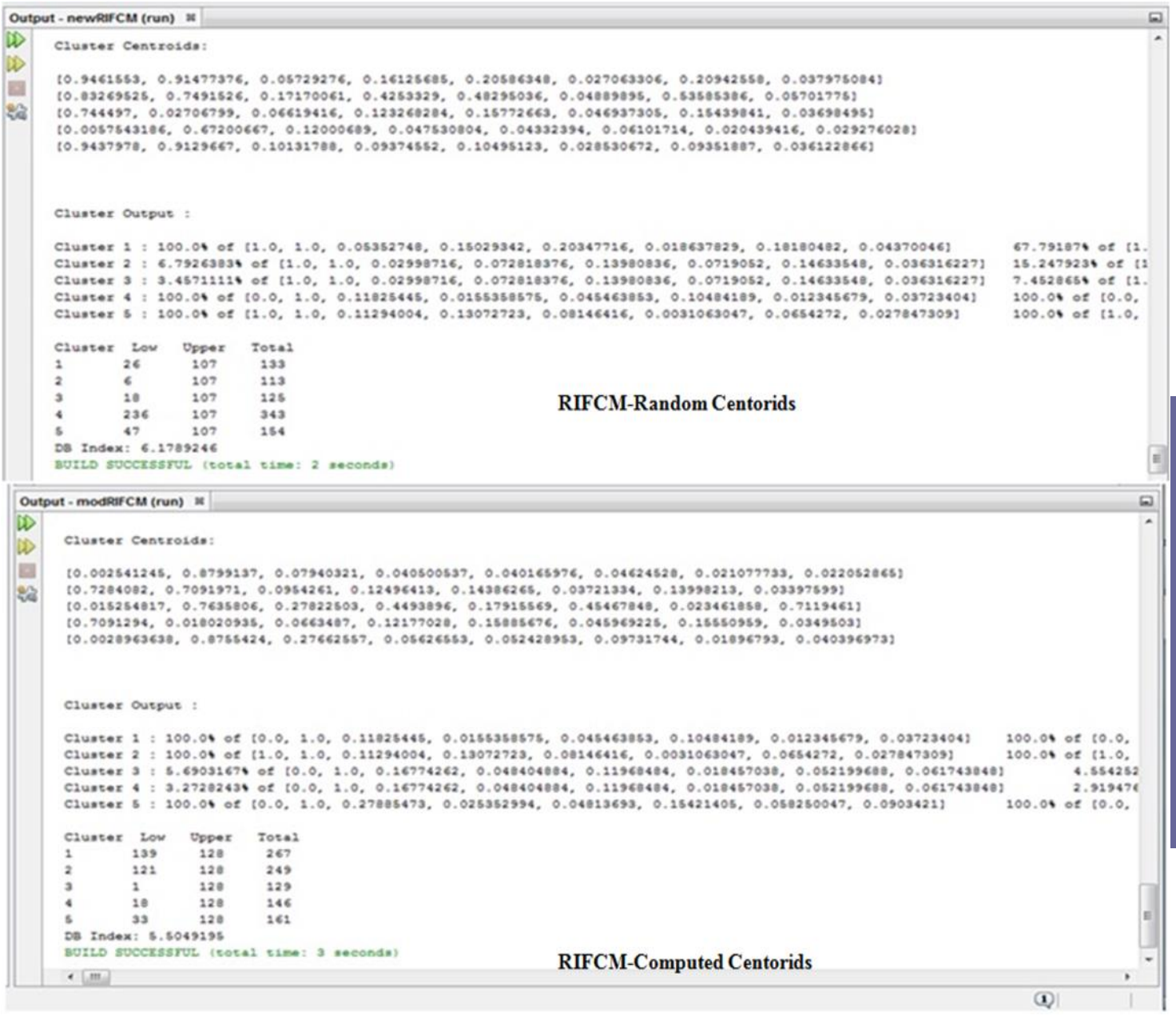
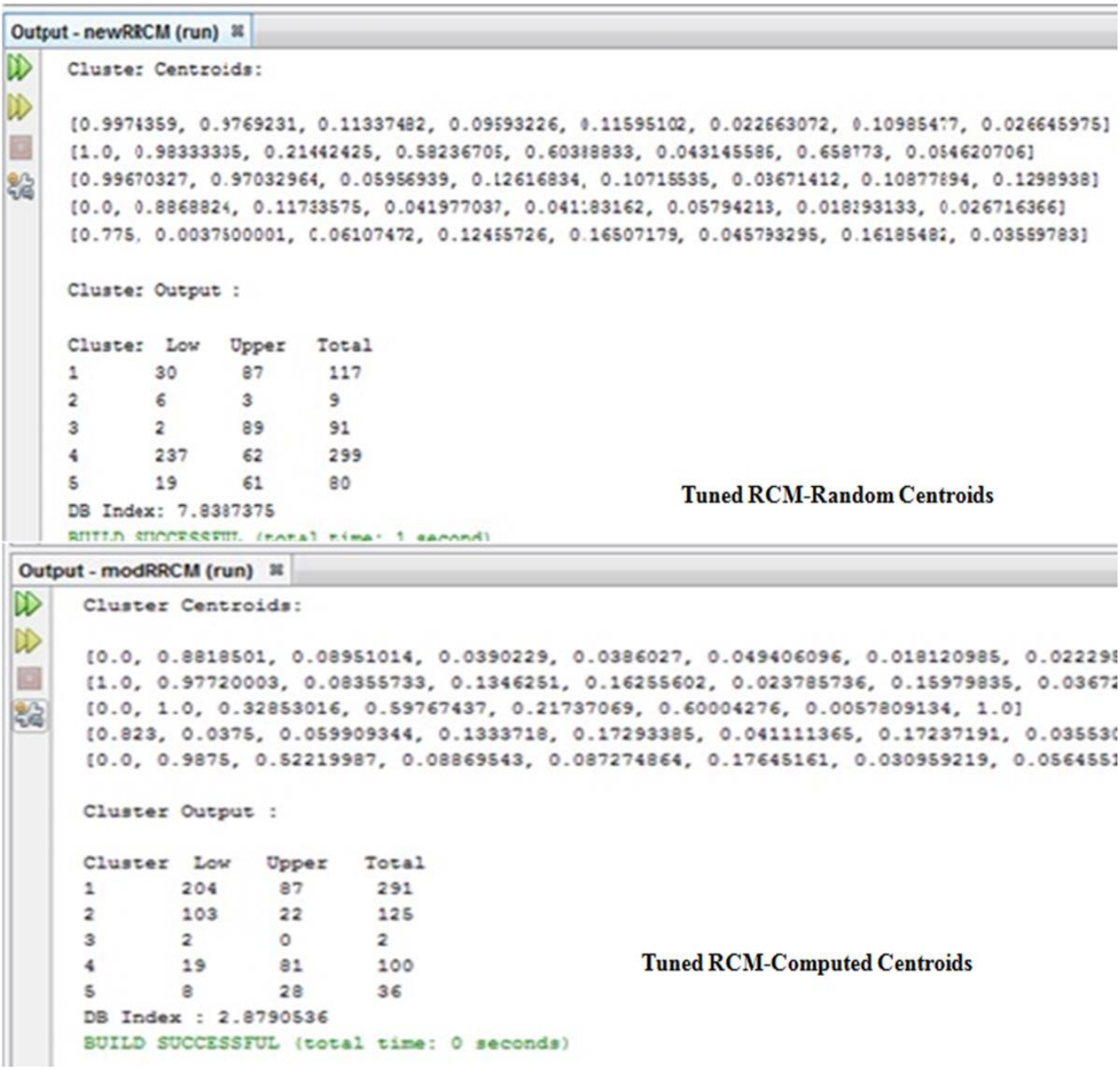


Fig: 1. d

THE IIOAB JOURNAL
ISSN: 0976-3104



Output - newRRCM (run)

```
Cluster Centroids:

[0.9974359, 0.9769231, 0.11337482, 0.09593226, 0.11595102, 0.022663072, 0.10985477, 0.026645975]
[1.0, 0.98333335, 0.21442425, 0.58236706, 0.60388833, 0.043145586, 0.658173, 0.054620706]
[0.99670327, 0.97032964, 0.05956939, 0.12616834, 0.10715535, 0.03671412, 0.10877894, 0.1298938]
[0.0, 0.8868824, 0.11733575, 0.041977037, 0.041183162, 0.05794213, 0.018193133, 0.026716366]
[0.775, 0.0037500001, 0.06107472, 0.12455726, 0.16507179, 0.045793295, 0.16185482, 0.035597783]

Cluster Output :

Cluster  Low   Upper   Total
1        30     87      117
2        6      3       9
3        2      89      91
4        237    62      299
5        19     61      80
DB Index: 7.8387375
BUILD SUCCESSFUL (total time: 1 second)
```

Tuned RCM-Random Centroids

Output - modRRCM (run)

```
Cluster Centroids:

[0.0, 0.8818501, 0.08951014, 0.0390229, 0.0386027, 0.049406096, 0.018120985, 0.022295
[1.0, 0.97720003, 0.08355733, 0.1346251, 0.16255602, 0.023785736, 0.15979835, 0.03672
[0.0, 1.0, 0.32853016, 0.59767437, 0.21737069, 0.60004276, 0.0057809134, 1.0]
[0.823, 0.0375, 0.059909344, 0.1333718, 0.17293385, 0.041111365, 0.17237191, 0.035530
[0.0, 0.9875, 0.52219987, 0.08869543, 0.087274864, 0.17645161, 0.030959219, 0.056455

Cluster Output :

Cluster  Low   Upper   Total
1        204    87      291
2        103    22      125
3        2      0       2
4        19     81      100
5        8      28      36
DB Index : 2.8790536
BUILD SUCCESSFUL (total time: 0 seconds)
```

Tuned RCM-Computed Centroids

Fig: 1. e

THE IIOAB JOURNAL
ISSN: 0976-3104

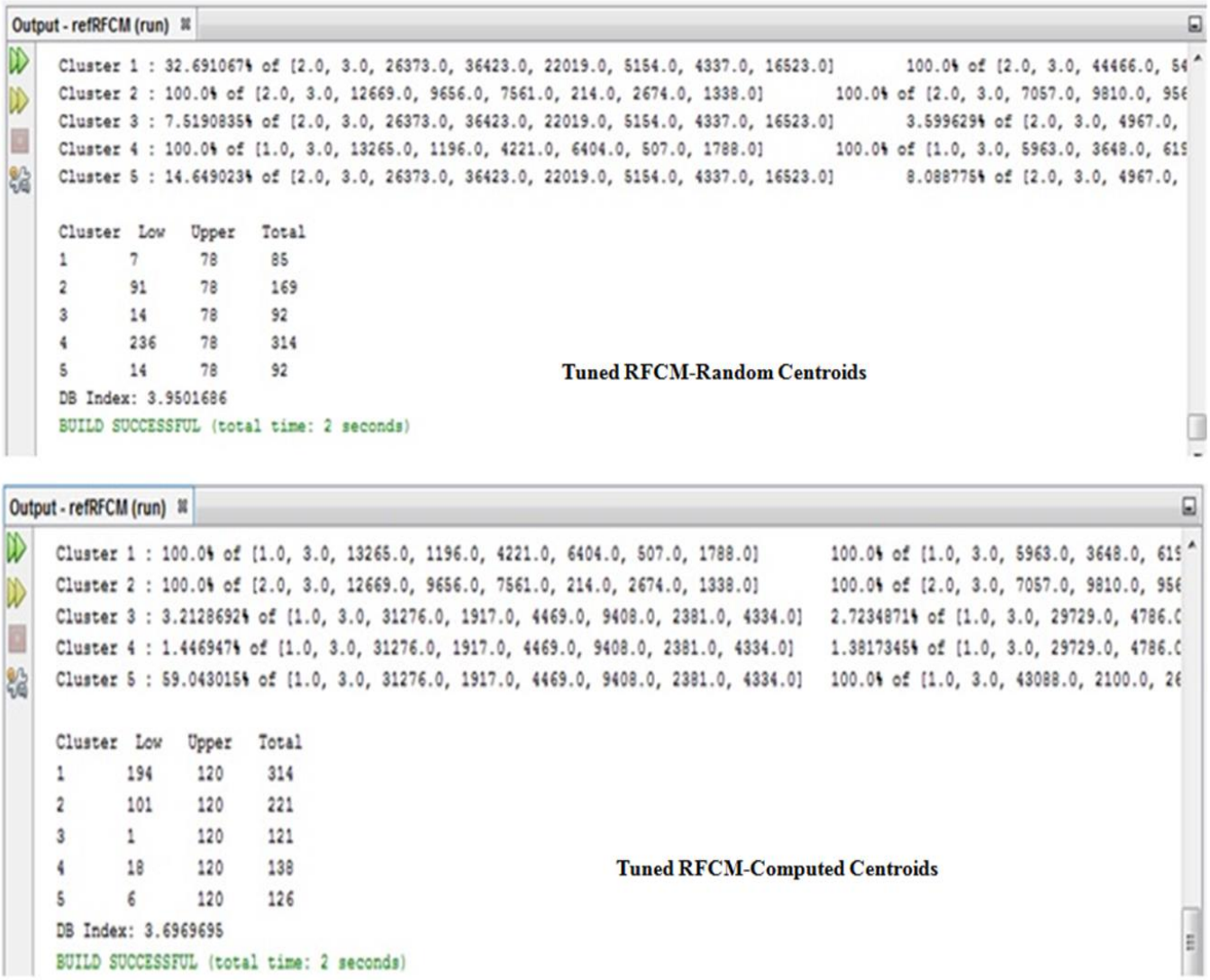


Fig: 1. f

Fig: 1. a-f. The Cluster formation comparisons using random centroid with proposed computation centroid for traditional soft clustering algorithms (RCM,FCM,RFCM,RIFCM) and Tuned hybrid soft clustering algorithms(TRFCM).

The clustering algorithms described are partitive, requiring pre-specification of the number of clusters. The results are dependent on the choice of c.

DISCUSSION

There exist validity indexes to evaluate the goodness of clustering, corresponding to a given value of c. In this paper, we compute the optimal number of clusters c0 in terms of the DB and D cluster validity indexes. The DB is a function of the ratio of the sum of within-cluster distance to between-cluster separation.

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{|ck|}\}$ be a set of patterns lying in a cluster U_k.. Then, the average distance between objects within the cluster U_k is expressed as:

$$S(U_k)=\frac{\sum_{i,i'}\|x_i - x_i'\|}{|c_k|(|c_k|-1)} \quad \text{where } \mathbf{x}_i, \mathbf{x}_{i'} \in U_k, \text{ and } i \neq i'$$

The between-cluster separation is defined as:

THE IIOAB JOURNAL ISSN: 0976-3104

$$d(U_k, U_l) = \frac{\sum_{i,j} \|x_i - x_j\|}{|c_k||c_l|}$$

Where $\mathbf{x}_i \in U_k$, $\mathbf{x}_j \in U_l$, such that $k = l$. The optimal clustering, for $c = c_0$, minimizes

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

for $1 \leq i, j \leq c$. Thereby, the within-cluster distance $S(U_i)$ is minimized while the between-cluster separation $d(U_i, U_j)$ gets maximized. Like DB index, the D index is designed to identify sets of clusters that are compact and separated. Here, we maximize for $1 \leq i, j \leq c$. The inter-cluster separation is maximized, while minimizing intra-cluster distances. Note that the denominator of DB is analogous to the numerator of D.

$$D = \min_j \left\{ \min_{i \neq j} \left\{ \frac{d(U_i, U_j)}{\max_k S(U_k)} \right\} \right\}$$

The computation of the initial centroids of each cluster instead of random allocation generates a lower DB index resulting in clusters with greater accuracy. The traditional and tuned soft clustering algorithms are compared using DB index with default initial centroids and computed initial centroids. The results are shown in **Table– 2** for 5 cluster and **Table– 3** for **Table–4** clusters.

Table: 2. Comparisons of Clustering Algorithms using DB index with different centroids for 5 cluster

Algorithm	No. of Clusters	Del	Epsilon	DB index with default initial centroids	DB index with computed initial centroids
RCM	5	0.3	-	12.046244	2.912782
FCM	5	-	0.05	3.5671623	3.4679413
RFCM	5	0.2	0.05	6.4094977	5.577581
RIFCM	5	0.2	0.05	6.1789246	5.5049195
Tuned RCM	5	1.5	0.05	14.211797	2.0191371
Tuned RFCM	5	1.4	0.05	4.375386	2.335935

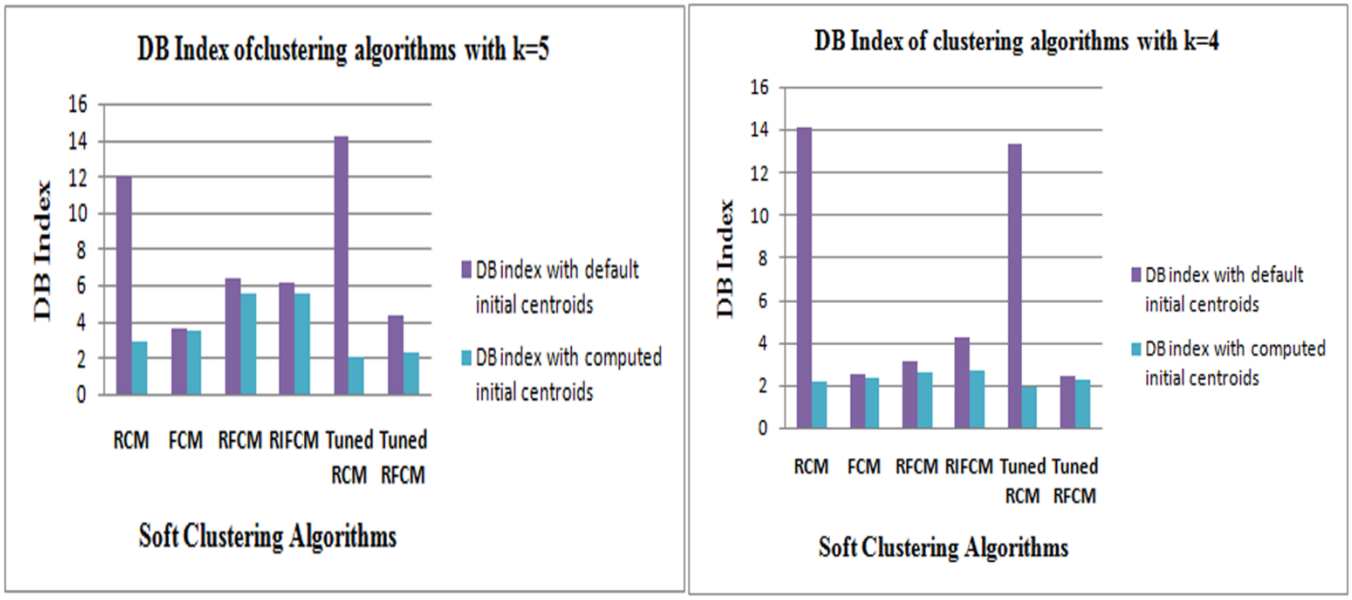
Table: 3. Comparisons of Clustering Algorithms using DB index with different centroids for 4 clusters

Algorithm	No. of Clusters	Del	Epsilon	DB index with default initial centroids	DB index with computed initial centroids
RCM	4	0.2	-	14.110096	2.1260233
FCM	4	-	0.05	2.4907343	2.3366437
RFCM	4	0.2	0.05	3.1227834	2.5758417

THE IIOAB JOURNAL
ISSN: 0976-3104

RIFCM	4	0.2	0.05	4.22623	2.7078934
Tuned RCM	4	1.4	0.05	13.310548	1.8863555
Tuned RFCM	4	1.4	0.05	2.4247031	2.2060814

Upon analyzing the output produced by each algorithm in terms of DB index, it can be concluded that the efficiency of the algorithm is greatly affected by the parameters that are used for conclusion.



(a) **(b)**

Fig: 2. Performance of Clustering Algorithms using DB index with different centroids (a) for 5 cluster (b) for 4 Cluster

The performances of the various soft clustering algorithms are represented in **Figure– 2 (a) and (b)** respectively. The various soft clustering approaches are validated with number of cluster 4 and 5. All the approaches are tested for random centriod and dynamic centriod computation. The Result shows that, the proposed tuned RFCM algorithm performs very well than other soft clustering approaches with respect to number of cluster and dynamic centriod computation.

CONCLUSION

Data Clustering is one of the vital research domains with a number of issues. Much of the work done in hard clustering algorithms and a few work carried out in traditional soft clustering algorithms such as rough c-means and fuzzy c-means. In this paper, a tuned hybrid soft clustering algorithm termed as tuned rough fuzzy c-means algorithm is presented. The selection of initial centroid is one of the issues in c-means algorithm, which is resolved by dynamic computation in the proposed algorithm. UCI Machine Learning Repository, Wholesale customers Data Set has been used to compare and validate the performance of the proposed algorithm with traditional soft clustering approaches.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

This work is part of PhD Research work. It is not supported by any agency.

THE IIOAB JOURNAL
ISSN: 0976-3104

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] AK Jain, MN Murty, PJ Flynn. [1999] Data clustering: a review, *ACM Computing Surveys* 31 (3) 264–323.

[2] Anna Maria Radzikowska, Etienne E. Kerre.[2002] A comparative study of fuzzy rough sets, *Fuzzy Sets and Systems*, 126(2): 137–155, ISSN 0165-0114.

[3] Selman Bozkir, Ebru Akcapinar Sezer.[2013] FUAT – A fuzzy clustering analysis tool, *Expert Systems with Applications*, 40(3): 15 842–849.

[4] D Dubois, H Prade.[1990] Rough fuzzy sets and fuzzy rough sets, *Internat J General Systems* 17 (2): 3191–209.

[5] Fan Li, Mao Ye, Xudong Chen. [2014] An extension to Rough c-means clustering based on decision-theoretic Rough Sets model, *International Journal of Approximate Reasoning*, 55(1): 116–129.

[6] G.Peters.[2006] Some refinements of rough k-means clustering, *Pattern Recognition* 39 :1481–1491.

[7] Georg Peters, Fernando Crespo, Pawan Lingras, Richard Weber.[2013] Soft clustering – Fuzzy and rough approaches and their extensions and derivatives, *International Journal of Approximate Reasoning*, 54(2):307–322.

[8] Georg Peters, Richard Weber, René Nowatzke,[2012] Dynamic rough clustering and its applications, Applied Soft Computing, 12 (10): 3193–3207.

[9] Ganesh Krishnasamy, Anand J Kulkarni, Raveendran Paramesran. [2014]A hybrid approach for data clustering based on modified cohort intelligence and K-means, *Expert Systems with Applications*, 41(13): 6009–6016.

[10] JC Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[11] Lin Zhu, Longbing Cao, Jie Yang, Jingsheng Lei.[2014] *Evolving soft subspace clustering, Applied Soft Computing*, 14(Part B) :210–228.

[12] Lingras C.[2004] West, Interval set clustering of web users with rough k-means, *Journal of Intelligent Information Systems* 23 (1): 5–16.

[13] Liyong Zhang ,Witold Pedrycz , Wei Lu , Xiaodong Liu , Li Zhang. [2014]" An interval weighed fuzzy c-means clustering by genetically guided alternating optimization", *Expert Systems with Applications* 41: 5960–5971.

[14] LA Zadeh.[(1994)] Fuzzy logic, neural networks, and soft computing, *Communica- tions of the ACM* 37:77–84.

[15] M Emre Celebi, Hassan A Kingravi, Patricio A Vela.[1994] A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications*, 40(1): 200–210.

[16] Matteo Brunelli, József Mezei.[2013] How different are ranking methods for fuzzy numbers? A numerical study, International *Journal of Approximate Reasoning*, 54(5): 627–639.

[17] BK Tripathy, GK Panda and A Mitra. "Covering Based Rough Equality of sets and Comparison of Knowledge", in: Proceedings of the Inter. *Conf. in Mathematics and Computer Science* (ICMCS 2009), 5-6 Jan. 09', Chennai, INDIA,2:438–443.

[18] Pawlak, Z. [1982]"Rough Sets", *Int Jour Inf Comp Sc*, 11:341–356.

[19] Pawlak Z. [1991] Rough Sets: Theoretical Aspects of Reasoning about Data, *Kluwer Academic Publishers.*

[20] Prakash Kumar, BK Tripathy.[2009] MMeR: an algorithm for clustering heterogeneous data using rough set theory,*International Journal of Rapid Manufacturing*, 1(2):189–207.

[21] P Maji, SK Pal. [2007] RFCM: a hybrid clustering algorithm using rough and fuzzy sets, *Fundamenta Informaticae* 79:1–22.

[22] Pierpaolo D'Urso, Riccardo Massari.[2013]Fuzzy clustering of human activity patterns, *Fuzzy Sets and Systems*, 215: 29–54.

[23] S Mitra, H. Banka, W Pedrycz. [2006] Rough–fuzzycollaborativeclustering,IEEE Transactionson Systems, Man, and Cybernetics—Part B36:795–805.

[24] S Mitra, T Acharya. [2003] Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley, New York

[25] Sotirios P Chatzis.[2011] A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional, Expert Systems with Applications, 38(7): 8684-8689.

[26] Tripathy BK, Tripathy HK. [2009]Covering Based Rough equivalence of Sets and Comparison of Knowledge, Proceedings of the IACSIT Spring Conference 2009, Singapore, 303–307.

[27] Tripathy BK, Mitra A and Ojha J. [2008] On Rough Equalities and Rough Equivalences of Sets, RSCTC 2008-Akron, U.S.A., Springer-Verlag Berlin Heidelberg LNAI 5306, 92–102.

[28] Tripathy BK, Tripathy HK. [2009]] Covering Based Rough Equivalence of Sets and Comparison of Knowledge. Computer Science and Information Technology - Spring Conference, 2009.IACSITSC '09. *International Association of* ,303–307, 17–20 April 2009

[29] https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

[30] BKTripathy and K.Govindarajulu. [2014] On Covering Based Pessimistic Multi Granular Rough Sets,2014 Sixth International Conference on Computational Intelligence and Communication networks,978-1-4799-6929-6 (14): 708–713.

[31] BK Tripathy, K. Govindarajulu. [2015] Some more properties of covering based multigranular rough sets, INDIA 2015,Kalayni University, J.K.Mandal et al(Eds),Information system design and applications,Advances in *Intelligent Systems and Computing*,339:555–564.

ABOUT AUTHORS

THE IIOAB JOURNAL
ISSN: 0976-3104

Prof. Prabhavathy Panneer *is working as Assistant Professor (Senior) in School of Information Technology and Engineering,VIT University,Vellore. Her research area includes computational intelligence, data mining, database. She has published 8 journal paper in her research filed. She is life member of CSI and IEEE. She is also part of various school activity committees. She has published number of papers in international conferences.*

Dr. B. K Tripathy *is a senior professor in the school of computing sciences and engineering, VIT University, at Vellore, India. He has been awarded with Gold Medals both at graduate and post graduate levels of Berhampur University, India. Also, he has been awarded with the best post graduate of the Berhampur University. He has received national scholarship, UGC fellowship, SERC visiting fellowship and DOE (Govt. of India) scholarship at various levels of his career. He has published more than 240 technical papers in various international journals, conferences, and Springer book chapters. He has produced 18 PhD's under his supervision. He is associated with many professional bodies like IEEE, ACM,IRSS, WSEAS, AISTC, ISTP, CSI, AMS, and IMS. His name also appeared in the editorial board of several international journals like CTA, ITTA, AMMS, IJCTE, AISS, AIT,and IJPS. His research interest includes fuzzy sets and systems, rough sets and knowledge engineering, data clustering, social network analysis, soft computing and granular computing.*