International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:7, 2015

# Upgraded Rough Clustering and Outlier Detection Method on Yeast Dataset by Entropy Rough K-Means Method

P. Ashok, G. M. Kadhar Nawaz

***Abstract*—**Rough set theory is used to handle uncertainty and incomplete information by applying two accurate sets, Lower approximation and Upper approximation. In this paper, the rough clustering algorithms are improved by adopting the Similarity, Dissimilarity–Similarity and Entropy based initial centroids selection method on three different clustering algorithms namely Entropy based Rough K-Means (ERKM), Similarity based Rough K-Means (SRKM) and Dissimilarity-Similarity based Rough K-Means (DSRKM) were developed and executed by yeast dataset. The rough clustering algorithms are validated by cluster validity indexes namely Rand and Adjusted Rand indexes. An experimental result shows that the ERKM clustering algorithm perform effectively and delivers better results than other clustering methods. Outlier detection is an important task in data mining and very much different from the rest of the objects in the clusters. Entropy based Rough Outlier Factor (EROF) method is seemly to detect outlier effectively for yeast dataset. In rough K-Means method, by tuning the epsilon (ε) value from 0.8 to 1.08 can detect outliers on boundary region and the RKM algorithm delivers better results, when choosing the value of epsilon (ε) in the specified range. An experimental result shows that the EROF method on clustering algorithm performed very well and suitable for detecting outlier effectively for all datasets. Further, experimental readings show that the ERKM clustering method outperformed the other methods.

***Keywords*—**Clustering, Entropy, Outlier, Rough K-Means, validity index.

## I. Introduction

CLUSTER analysis [10], [1] is one of the crucial data analysis tools in data mining. The K-Means clustering algorithm is a very simple and fast efficient one. Clustering is the method of grouping the data into classes or clusters. So that objects within a cluster have high similarity in evaluation to one another, but are very dissimilar to objects in other clusters. Clustering can be implemented on Nominal, Ordinal and Ratio-Scaled variables. The main resolution of clustering is to reduce the size and complexity of the dataset. In rough clustering [5], [6] each cluster has two approximations, lower and upper approximations. The lower approximation is a subset of the upper approximation. The members of the lower approximation [13] belong positively to the cluster, therefore they cannot belong to any other cluster. The data objects in the upper approximation may belong to the cluster. Since their membership is uncertain, they must be a member of an upper approximation of at least another cluster.

This paper is organized as follows. Section II presents an overview of Rough set theory, Section III describes upgraded rough clustering algorithms, Section IV expresses outlier detection methods, Section V states cluster validity techniques, Section VI presents Experimental analysis and discussion and Section VII presents conclusion and future work.

Ashok p is with the Bharathiar University, Coimbatore, Tamilnadu, India. Mobile: +91 9944050320; Email: ashokcutee@gmail.com

Kadhar Nawaz G.M is a Director of M.C.A Department, Sona college of Technology, Salem, Tamilnadu, India. E-mail: nawazse@yahoo.co.in

## II. Overview of Rough Set Theory

### A. Rough Set

The concept of rough set [8], [9] has recently appeared as another major mathematical tool for managing uncertainty that arises from granularity in the area of discourse that is, from the indiscernibility between objects in a set. The aim is to estimate a rough set [4] concept in the domain of discourse by a couple of exact concepts. These exact concepts are unwavering by an indiscernibility relation on the domain, which in turn, may be induced by a given set of attributes belonging to the objects of the domain. Fig. 1 represents the area of the rough set that the region in the corners of the rectangle represents upper approximation, solid circle represents lower approximation and the area between lower and upper approximation represents boundary area. The lower approximation is the set of objects definitely belonging to the imprecise concept, whereas the upper approximation is the set of objects probably belonging to the imprecise concept. The smaller value of threshold, the more likely is the object to lie within the rough boundary [7] (between upper and lower approximations) of a cluster. A larger value of threshold states that more objects are permitted to belong to any of the lower approximation.

### B. Rough Clustering

A rough clustering is defined in a similar manner of rough set theory and it has lower and upper approximations. The lower approximation of a rough cluster contains objects that only belong to that cluster. The upper approximation of a rough cluster contains objects in the cluster which are also members of other clusters.

Rough properties of the clustering algorithm are:

- Property 1: A data object can be a member of one lower approximation at most.
- Property 2: A data object that is a member of the lower approximation of a cluster and it is also member of the

upper approximation of the same cluster.

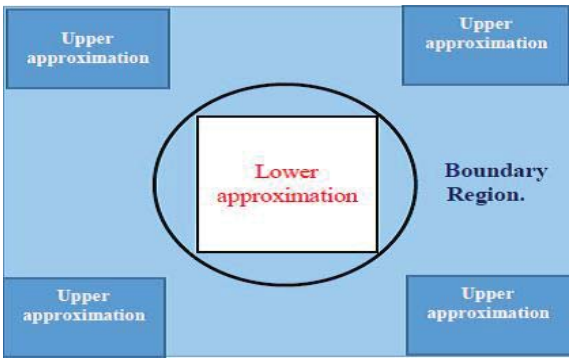


Fig. 1 Regions of Rough Set

## III. Upgraded Rough K-Means Clustering Techniques

The Rough K-Means clustering algorithms are upgraded by assignment of preliminary cluster centroid for the rough clustering method.

The preliminary centroids play an important role in the clustering process, sometimes the algorithm produce bad clustering results due to selecting wrong centroids by randomly in the given data set. To avoid this problem and improve the clustering process, we need a new methodology to select perfect preliminary centroids for the clustering process. The three different types of preliminary centroid selection methods namely similarity measure, entropy measure and similarity to dissimilarity based preliminary centroid selection methods are described in the following section.

### *A. Similarity Based Rough K-Means (SRKM)*

The preliminary centroid of clustering process is identified by the method of similarity measure between objects. The object having highest similarity is selected as preliminary centroid for clustering process. The similarity value between objects are calculated and the preliminary centroids of the Rough K-Means algorithm are selected. The Similarity Rough K-Means (SRKM) clustering algorithm is developed by implementing the similarity measure based preliminary centroid selection method on Rough K-Means clustering algorithm and it is described in the algorithm given below.

*Algorithm: SRKM*

*Input:* Rough clustering algorithm requires the addition of the concept of lower and upper approximations.

Lower = 0.7 Upper = 0.3 Epsilon ($\varepsilon$) = 1.0

*Output:* Set of objects in lower and upper approximation**.**

Step 1: Assignment of each data object to exactly one lower approximation based on the object having maximum similarity value. By definition (Property 2) the data object also belongs to the upper approximation of the same cluster. The similarity value of an object can be measured by the following steps:

$$s_{ij} = e^{-\alpha Dist_{ij}} \tag{1}$$

Here the **Dist$_{ij}$** is represents the distance between the two objects which can be calculated by the expression below.

$$Dist_{ij} = \sqrt{\sum_{m=1}^{n} (X_{im} - X_{jm})^2} \tag{2}$$

Here α is a geometric constant. The similarity value between any two points lies in the range of 0.0 to 1.0. The value of α is determined that the similarity value $S_{ij}$ is set equal to 0.5, when the distance between two data points (i.e., $Dist_{ij}$) becomes equal to the mean distance D, which is represented as:

$$\alpha = -\frac{\ln 0.5}{\overline{Dist}} \tag{3}$$

Here $\overline{Dist}$ represents the average distance of all the objects and the value can be determined from:

$$\overline{Dist} = \frac{1}{N} \sum_{k=1}^{N} \sum_{j>i}^{N} Dist_{ij} \tag{4}$$

$$TS_i = \sum_{j \in x}^{j \neq i} S_{ij} \tag{5}$$

Here $TS_i$ represents Total Similarity value between objects. An object having the maximum Total Similarity has been assigned as the preliminary centroid for consecutive clusters.

Step 2: Calculation of new means as preliminary centroids. The means $m_k$ are calculated as:

$$m_k = \begin{cases} W_l \sum_{X_{kx} C_k} \frac{Xn}{|C_k|} + W_B \sum_{X} \frac{X_n}{|C_k^B|} \; for C_k^B \neq \phi \\ W_l \sum_{X_{kx} C_k} \frac{X_n}{|C_k|} otherwise \end{cases} \tag{6}$$

Here the parameters wl and wb define the significance of the lower approximation and boundary area of the rough cluster. The expression |Ck| point out the numbers of data objects in lower approximation of the cluster and $|C_{Bk}| = |\underline{C_k} - \overline{C_K}|$ is the number of data objects in the boundary area.

Step 3: Assign the data objects to the approximations.

(i) For a given data object $X_n$ determine its closest mean $m_h$

$$d_{n,h}^{\min} = \min_{k=1..k} d(X_n, m_k) \tag{7}$$

Assign $X_n$ to the upper approximation of the cluster h: $X_n \in$ Ch.

(ii) Determine the means $m_t$ that are also close to $X_n$ they are not farther away from $X_n$ than $d(X_n, m_h)$where is a given threshold:

$$T = \{t : d(X_n, m_k) - d(X_n, m_h) \leq \varepsilon \cap h \neq k\} \quad (8)$$

If T= Ø ($X_n$ is also close to at least one other mean $m_t$ besides $m_h$) Then $X_n \in Ct$, $\forall$ t $\in$ T.

Else $X_n \in Ch$.

Step 4: If the algorithm does not meet the convergence criteria, continue Step 2. Otherwise stop the process.

*B. Entropy Based Rough K-Means (ERKM)*

The preliminary centroid for clustering process is identify by the method of entropy measure between objects. The object having minimum entropy value is selected as preliminary centroid for clustering process. The entropy measure between objects are calculated and the preliminary centroids for the Rough k-means algorithm are selected for the clustering process. The Entropy Rough K-Means (ERKM) clustering algorithm is developed by implementing the entropy based preliminary centroid selection method on Rough K-Means clustering algorithm and it is described in the algorithm given below.

*Algorithm: ERKM*

*Input:* Entropy based Rough clustering necessitates the values of lower and upper approximations.

Lower = 0.7 Upper = 0.3 Epsilon ($\varepsilon$) = 1.0

*Output:* Group of objects in lower and upper approximation.

Step 1: Distribute each data object to exactly one lower approximation based on minimum entropy values of each object and the data object also belongs to the upper approximation of the same cluster. The entropy value of an object can be determined by the following steps.

$$E_i = -\sum_{j \alpha}^{j \neq i} \left( \left( S_{ij} \log_2 S_{ij} \right) + \left( 1 - S_{ij} \right) \log_2 (1 - S_{ij}) \right) \quad (9)$$

$E_i$ represents entropy value of $i^{th}$ object and $S_{ij}$ represents the similarity value between two objects i and j. The objects having minimum entropy values are assigned to subsequent clusters successively.

Step 2: Calculation of the new means as preliminary centroids. The means are calculated from (6).

Step 3: Assign the data objects to the approximations.

i. For a given data object $X_n$ determine its closest mean $m_h$ from (7)
ii. Determine the means $m_t$ that are also close to $X_n$ they are not farther away from $X_n$ than $d(X_n, m_h)$where T is a given threshold, the threshold value can determined from (8).

Step 4: If the algorithm does not meet the convergence criteria, continue Step 2. Otherwise the algorithm can be stopped.

*C. Dissimilarity-Similarity Rough K-Means (DSRKM)*

The preliminary centroid for clustering process is identify by the method of dissimilarity – similarity measure between objects. The object having minimum dissimilarity – similarity value is selected as preliminary centroid for clustering process. The Dissimilarity-Similarity Rough K-Means (DSRKM) clustering algorithm is developed by implementing the Dissimilarity- Similarity measure based preliminary centroid selection method on Rough K-Means clustering algorithm and it is described in the algorithm given below.

*Algorithm: DSRKM*

*Input:* Rough clustering needs the initial value of lower and upper approximations.

Lower = 0.7 Upper = 0.3 Epsilon ($\varepsilon$) = 1.0

*Output:* cluster of objects in lower and upper approximation.

Step 1: The objects are assigned to each clusters sequentially depends upon the objects having minimum proportion of dissimilarity to similarity values, which can be calculated from:

$$DS_i = \sum_{j \in n}^{j=i} \frac{\left(1 - S_{ij}\right)}{S_{ij}} \quad (10)$$

Here $DS_i$ represents the ratio of dissimilarity to similarity value of an object.

Step 2: Calculate the mean value as preliminary centroid each for cluster from (6)

Step 3: Assign each object to the lower approximation, when the distance between object and centroid is below the threshold value determined by (8), otherwise the object assigned in upper approximation only.

Step 4: if the objects assigned to the lower approximation and also assigned in upper approximation if the distance value between the object and centroid is below the threshold value.

Step 5: if the clustering algorithm does not meet the convergence criteria, continue step 2. Otherwise the algorithm can be stopped.

## IV. Outlier Detection Techniques

The objects that do not obey with the general behaviour of the object, such data objects which are grossly different from or inconsistent with the remaining set of object are called outliers [2]. The outlier detection and analysis is an interesting data mining task referred to as outlier mining or outlier analysis.

*A. Detecting Outlier by Entropy based Rough Outlier factor (EROF)*

The Entropy based Rough Outlier Factor (EROF), which can indicate the degree of outlierness for every object of the universe in an information system. Let *IS* = (*U*, *A*) be an information system, the Entropy based Rough Outlier Factor

*EROF*($x$) of object $x$ in *IS* is defined as:

$$E_i = P_i * \log P_i \tag{11}$$

Here $P_i$ is the distance between the object and centroid.

$$EROF_i = \left( \frac{(E_i^{\max} - E_i^{\min})}{2} * \left( 1 - \frac{|C_i|}{n} \right) \right) \tag{12}$$

Here $\boldsymbol{E_i^{max}}$ represents the object having maximum entropy value in the i[th] cluster and $\boldsymbol{E_i^{min}}$ represents object having minimum entropy value in the i[th] cluster. $|ci|$ denotes the cardinality of cluster $c_i$. For any object $x \in c_i$, Where $EROF_i$ is the relative entropy based rough outlier factor of the cluster $c_i$, the EROF is calculated for each cluster separately and the entropy value for each objects on that clusters is compared by the EROF of that cluster, if $E_i < EROF_i$ then object$_i$ is called an outlier.

### *B. Outlier Detection Algorithm*

$|U| = n$, c is the number of clusters and EROF is the entropy Rough Outlier Factor.

*Steps:*

1) *For every* $x_i \in U$
2) *For* $j = 1$ *to* $n$
3) *Calculate the entropy of* $E_{xj}$ *each object in U*
4) *End*
5) *For* $i=1$*to* $c$
6) *Calculate* $EROF_i$
7) *End*
8) *For* $I = 1$ *to* $c$
9) *For* $j = 1$ *to* $n$
10) *If* $EROF_i > E_j$, *then object*$_j$ *as outlier*
11) *End*
12) *End*
13) *End*

### *C. Detecting Outlier on Boundary Region of Rough Clustering Method by Tuning Parameter Epsilon*

In rough clustering, the region between lower approximation and upper approximation is called as boundary region. If the objects do not belong to the lower and upper approximation of the rough clustering, then it belongs to the boundary region of the rough clustering. The objects in the boundary region are detected as outliers. The outliers are identified by tuning the parameter epsilon ($\varepsilon_i$) on rough clustering method. The epsilon ($\varepsilon_i$) is one of the important factor for Rough K-Means clustering algorithm. The Rough K-Means clustering algorithm delivers better clustering results depends upon the selection of epsilon value. The rough clustering algorithm is executed by adjusting the value of epsilon ($\varepsilon_i$) between 0.8 and 1.16 without changing the cluster value 10. The outlier in the boundary region are minimized or maximized depends upon the value of epsilon ($\varepsilon_i$). The objects in the lower approximation may be empty due to wrong selection of epsilon value. By tuning the epsilon value from 1.0 to 1.08 can avoid the empty of lower approximation in rough clustering method. Hence the parameter epsilon ($\varepsilon_i$) plays the vital role in rough clustering method.

## V. Cluster Validation Techniques

Clustering validity [3] is a concept of evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, the clustering validity index may also be used to find the optimal number of clusters. When using clustering algorithms one must use performance measures to evaluate the clustering results. There are some well-known performance measures with their inherent advantages and drawbacks. To validate the rough clustering method by using two cluster validity index namely:

- Rand index
- Adjusted Rand index

### *A. Rand Index*

The Rand Index (RI) [11] proposed by Rand is a popular validity index and probably most used for cluster validation. Rand Index can be easily computed by:

$$RI = \frac{a+d}{a+b+c+d} \tag{13}$$

***a*** - objects in a pair are placed in the same group in U and in the same group in V. ***b*** - Objects in a pair are placed in the same group in U and in different groups in V. ***c*** - Objects in a pair are placed in the same group in V and in different groups in U. ***d*** - Objects in a pair are placed in different groups in U and in different groups in V.

The measures ***a*** and ***d*** be taken as agreements, ***b*** and ***c*** as disagreements. The Rand index value lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

### *B. Adjusted Rand Index*

The Adjusted Rand index [12] proposed by Rand (1971) is the corrected and improved version of the Rand index. Though the Rand Index may only yield a value between 0 and +1, the Adjusted Rand Index can yield negative values if the index is less than the expected index.

$$ARI = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{14}$$

(Or)

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]} \tag{15}$$

Here a, b, c and d are calculated same as defined in the rand index.

## VI. Experimental Results and Discussion

### A. Dataset

The experimental analysis is carried out in this chapter by considering different data sets from UCI data repository and the clustering algorithms are validated through Rand index and Adjusted Rand index [11].

1. *Yeast:* The yeast dataset contains 1400 samples with 8 attributes. The dataset is collected from the location which is given at: *http://archive.ics.uci.edu/ml/machine-learning-databases /yeast/yeast.data*
2. *Iris:* The iris dataset contains the information about the iris flower. The data set contains 150 samples with four attributes. The dataset is collected from the location which is given at: *http://archive.ics.uci.edu/ml/machine-learning-database/ iris/ iris.data*
3. *Ecoli:* The Ecoli dataset contains protein localization sites having 336 samples with 3 attributes. The dataset is collected from the location which is given at: *http://archive.ics.uci.edu/ml/machine-learning-databases /ecoli/ecoli.data.*
4. *Glass:* The Glass dataset contains the information about glass and its types having 214 samples with 9 attributes. The dataset is collected from the location which is given at: *http://archive.ics.uci.edu/ml/machine-learning-databases /glass/glass.data.*
5. *Diabetes:* The diabetes dataset contains information about diabetes patients having 768 samples with 8 attributes. The dataset is collected from the location which is given at: *http://archive.ics.uci.edu/ml/machine-learning-databases /diabetes/diabetes.data.*

### B. Cluster Validity Measure Techniques

1. Analysis of Rough K-Means Algorithms by Rand Validity Index

The Rand index is one of the cluster validation index, which is used to validate the clustering process. The three clustering algorithms (ERKM, SRKM and DSRKM) are executed with yeast dataset by changing the cluster value from 3 to 30. They are validated and compared by Rand index and the obtained Rand index values are listed in Table I.

TABLE I
Rand Index Analysis

| Clusters | Rand Index | | |
|---|---|---|---|
| | ERKM | SRKM | DSRKM |
| 3 | 0.709 | 0.616 | 0.5685 |
| 6 | 0.662 | 0.654 | 0.6827 |
| 9 | 0.698 | 0.664 | 0.6811 |
| 12 | 0.685 | 0.669 | 0.6713 |
| 15 | 0.706 | 0.679 | 0.6714 |
| 18 | 0.681 | 0.679 | 0.6832 |
| 21 | 0.678 | 0.678 | 0.6814 |
| 24 | 0.715 | 0.671 | 0.6818 |
| 27 | 0.678 | 0.684 | 0.6839 |
| 30 | 0.699 | 0.675 | 0.6806 |

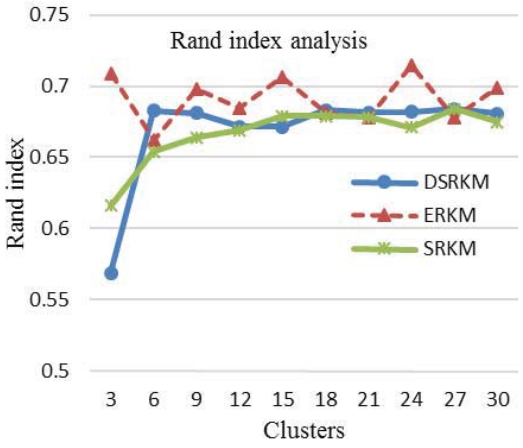


Fig. 2 Rand index chart for Rough clustering algorithms

Fig. 2 shows that the three rough clustering algorithms are performed effectively with yeast dataset and obtained different Rand index values for different cluster values, but the Entropy Rough K-Means (ERKM) algorithm indicated by dotted line obtained higher Rand index value than SRKM and RKM for the most of the obtained results. Hence Entropy Rough K-Means clustering method is most suitable for clustering yeast dataset.

2. Analysis of Rough K-Means Algorithms by Adjusted Rand Validity Index

The three rough clustering algorithms can be validated by another powerful cluster validity index called Adjusted Rand index, which can be used to evaluate the clustering process. The three rough clustering algorithms are executed with the dataset of Yeast and changing cluster values from 3 to 30. The obtained Adjusted Rand index value of the rough clustering methods are depicted in Table II.

TABLE II
Adjusted Rand Index Analysis

| Clusters | Adjusted Rand Index | | |
|---|---|---|---|
| | ERKM | SRKM | DSRKM |
| 3 | 0.338 | 0.319 | 0.281 |
| 6 | 0.354 | 0.337 | 0.342 |
| 9 | 0.411 | 0.311 | 0.354 |
| 12 | 0.341 | 0.297 | 0.302 |
| 15 | 0.331 | 0.311 | 0.287 |
| 18 | 0.303 | 0.308 | 0.301 |
| 21 | 0.281 | 0.299 | 0.295 |
| 24 | 0.303 | 0.291 | 0.292 |
| 27 | 0.276 | 0.311 | 0.296 |
| 30 | 0.311 | 0.276 | 0.279 |

Fig. 3 shows that the dotted line represents Entropy Rough K-Means (ERKM) clustering algorithm is obtained higher Adjusted Rand index value than other clustering algorithms for most of the cluster values. Similarly, the ERKM algorithm is executed by iris, *E. coli* and some other dataset, it delivers better clustering results for most of the dataset. The Similarity

Rough K-Means algorithm also deliver better results for some other cluster values. Hence the ERKM clustering algorithm performs very well, when compared with other two clustering algorithms for yeast dataset.

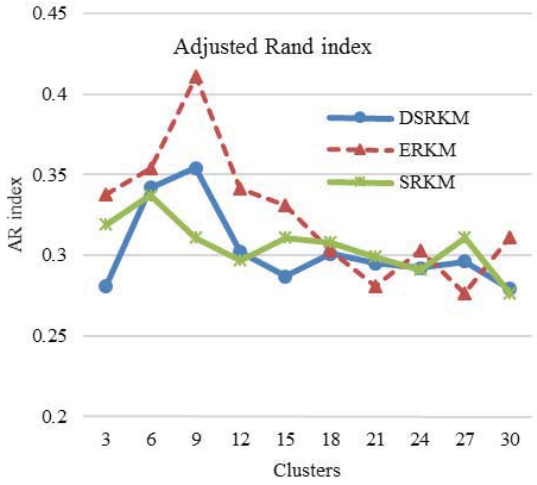


Fig. 3 Adjusted Rand index chart for Rough Clustering

*C. Analysis of Outlier Detection Technique*

1. Detecting Outlier by Entropy Rough Outlier Factor Method

The Entropy based outlier detection algorithm is used to find the outlier in clusters. The EROF method is discussed in the section IV and the outlier detection algorithm is able to identify the outliers very meritoriously. The rough clustering algorithm is executed with EROF method for different dataset. The obtained experimental results are listed in Table III.

TABLE III
DETECTING OUTLIERS ANALYSIS

| Dataset | Total No. of Objects | Detection Outliers by EROF | | | Total Outliers | % of Outliers in Dataset |
|---|---|---|---|---|---|---|
| | | Cluster1 | Cluster2 | Cluster 3 | | |
| **Yeast** | 1484 | 29 | 23 | 16 | 68 | 4.582 |
| **Iris** | 150 | 06 | 04 | 0 | 10 | 6.666 |
| ***E. coli*** | 336 | 05 | 11 | 03 | 19 | 5.654 |
| **Glass** | 214 | 03 | 0 | 05 | 08 | 3.738 |
| **Wine** | 178 | 01 | 03 | 03 | 07 | 3.932 |
| **Diabetes** | 768 | 29 | 23 | 16 | 68 | 4.581 |

The experimental results show that the EROF method detect 68 outliers out of 1484 objects of 4.582% outliers in yeast dataset. Similarly, it can detect 6.66% of outliers in Iris dataset, 5.654% of outliers in *E. coli* dataset, 3.738% of outliers in Glass dataset, 3.932% of outliers in Wine dataset and 4.581% of outliers in Diabetes dataset. The removal of outliers on each clusters improves the performance of the rough clustering process.

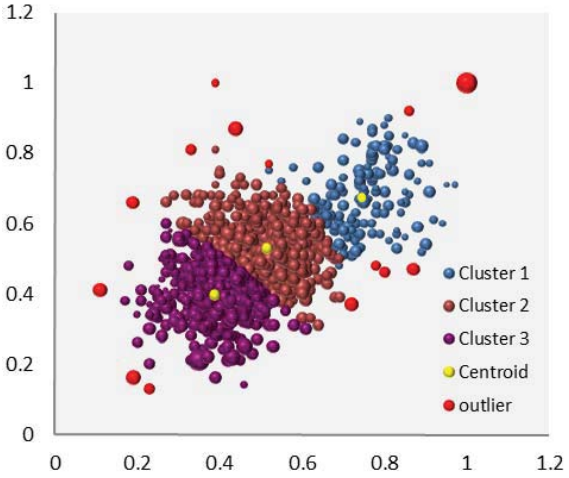


Fig. 4 Outlier Detection chart for ERKM method

Fig. 4 shows that outlier detection on yeast dataset by EROF method, the solid balls in three clusters are represent as objects, the blue colour solid balls are represent objects in cluster 1, brown colour solid balls are represent objects in cluster 2, violet colour solid balls are represent objects in cluster 3, yellow colour solid balls are represent centroids of the different clusters and finally the red colour solid balls are represent as outliers that can be identified by the Entropy based Rough K-Means clustering with outlier detection method. Hence the Entropy based outlier detection method delivers better results when compared with other methods.

2. Detecting Outlier on Boundary Region of Rough Clustering Algorithm

The three kinds of Rough K-Means algorithm are executed using yeast data set by tuning the parameter epsilon(ε) between 0.8 to 1.16 and initialize the number of clusters value is 10 for all the execution, the obtained results are listed in Table IV.

TABLE IV
OUTLIER DETECTION ON BOUNDARY REGION

| S. No | Dataset | Clusters | Epsilon (£) | Outlier in Boundary Region |
|---|---|---|---|---|
| | | | | ERKM |
| **1** | Yeast | 10 | 0.8 | 0 |
| **2** | | | 1.0 | 1 |
| **3** | | | 1.02 | 3 |
| **4** | | | 1.04 | 3 |
| **5** | | | 1.06 | 12 |
| **6** | | | 1.08 | 30 |
| **7** | | | 1.10 | 110 |
| **8** | | | 1.12 | 252 |
| **9** | | | 1.14 | 338 |
| **10** | | | 1.16 | 409 |

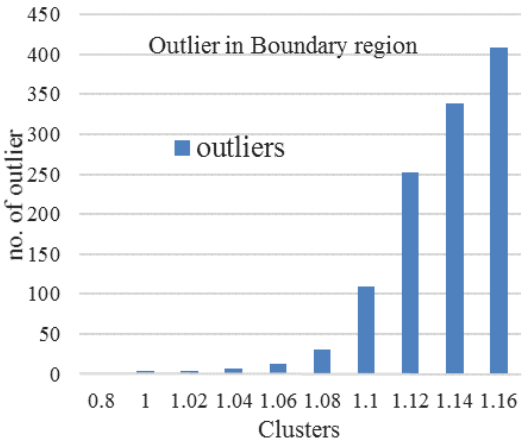


Fig. 5 Boundary elements chart for Rough K-Means clustering

Fig. 5 shows that the Entropy Rough K-Means clustering algorithm detecting outlier effectively and delivers better results, when selecting the value of epsilon ($\varepsilon_i$) between 1.0 and 1.08 for 10 clusters. The number of outliers detected in boundary region is minimum and avoid more number of objects assigned in boundary region. By increasing the epsilon ($\varepsilon_i$) value between 1.10 and 1.16, the Rough K-Means algorithm did not deliver better results and decrease the performance of the clustering method, because more number of outlier are detected in the boundary region. Hence the selection of epsilon ($\varepsilon_i$) value between 1.0 and 1.08 increases the performance of rough clustering algorithms. Tuning of epsilon value helps to avoid the lower approximation of a cluster becomes empty.

### *D.Performance of Rough Clustering Methods*

The performance of the clustering algorithms can be determined by the time taken for the clustering algorithm meet convergence criteria. The execution time of the Entropy-Rough K-Means, Similarity-Rough K-Means and DS-Rough K-Means clustering algorithms are calculated with various cluster values and listed in Table V.

TABLE V
CLUSTERING PERFORMANCE ANALYSIS

| S. No | Clusters | Execution Time (in Seconds) | | |
|---|---|---|---|---|
| | | ERKM | SRKM | DSRKM |
| **1** | 3 | 0.771 | 0.644 | 0.732 |
| **2** | 6 | 1.568 | 1.283 | 1.508 |
| **3** | 9 | 3.557 | 1.898 | 2.368 |
| **4** | 12 | 2.405 | 3.273 | 4.576 |
| **5** | 15 | 4.044 | 2.645 | 3.832 |
| **6** | 18 | 3.918 | 7.629 | 10.813 |
| **7** | 21 | 7.614 | 6.965 | 5.941 |
| **8** | 24 | 7.927 | 6.215 | 8.471 |
| **9** | 27 | 10.53 | 12.23 | 14.175 |
| **10** | 30 | 14.06 | 6.096 | 7.3333 |

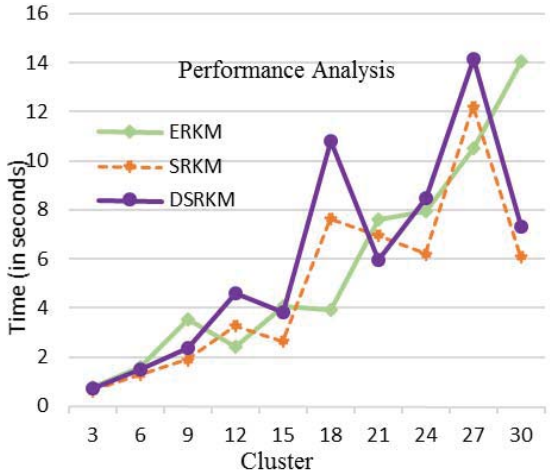


Fig. 6 Execution speed chart of clustering algorithms

Fig. 6 shows that the three rough clustering algorithms (ERKM, SRKM and DSRKM) perform very well when executed with different cluster values, but the SRKM clustering algorithm delivers better results and the algorithm meet the convergence criteria by minimum number of iteration, because it avoids more computational time due to having simple expression and it also expected to be computationally faster than ERKM and DSRKM methods.

## VII. CONCLUSION

In this paper, Rough K-Means clustering algorithms and outlier detection methods are studied very well and implemented on Matlab 2011 tool. The experimental results of the clustering algorithms, the following conclusions were reached.

The various kinds of preliminary centroids selection methods for clustering process were presented and compared. The experimental results show that the Entropy Rough K-Means clustering method obtained higher clustering accuracy rate than other methods. The rough clustering algorithms are executed by different dataset and validated. The ERKM clustering method perform very well and obtained higher rand index and adjusted rand index values for most of the datasets like yeast dataset.

The outliers are detected by the various methods are entropy measure, lower approximation analysis and boundary region of the rough clustering method. In that the EROF method can detect outliers effectively and improve the quality of the Rough K-Means clustering method after removal of outlier in the dataset. The parameter epsilon ($\varepsilon_i$) of the clustering method changes the behaviour of the Rough K-Means clustering process. Its performance can be improved and the objects (outlier) in the boundary region are minimized, when the Epsilon ($\varepsilon_i$) value established between 1.0 and 1.08. Entropy Rough K-Means clustering algorithm can deliver better results than SRKM and DSRKM clustering algorithms in terms of clustering quality, and its computational time is

larger than other methods.

To enrich the clustering process, the rough clustering method is hybrid with fuzzy clustering method and develop Rough-Fuzzy Clustering algorithm to detect outlier and then evaluated by different cluster validity measures is our future work.

## REFERENCES

[1] P. Ashok, G.M Kadhar Nawaz, E. Elayaraja, "Outliers detection on protein localization sites by Partitional clustering methods", In *Proc. International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME)*, Salem, Feb 2013, pp. 447 – 453

[2] P. Ashok, G.M Kadhar Nawaz , V. Vadivel , "Improved Performance of Unsupervised Method by Renovated K-Means", *IJASCSE*, Vol 2, no 1, pp. 41-47, 2013.

[3] D.L Davies, D.W Bouldin, "A cluster separation measure". *IEEE Trans.Pattern Anal. Machine Intell*, vol. 1, no 4, pp. 224-227, 2000

[4] Georg Peters, "Some refinements of rough *k*-means clustering", *Pattern Recognition,* vol. 39, pp. 1481 – 1491, *2006*

[5] Kevin E. Voges, "Research Techniques Derived from Rough Sets Theory: Rough Classification and Rough Clustering", *4th European Conference on Research Methodology for Business and Management Studies*, April 2005, pp. 437- 444.

[6] P. Lingras, "*Rough* Set Clustering for Web Mining", *In Proc. IEEE International Conference on Fuzzy Systems*. May 2002, pp. 5-16.

[7] P. Lingras, C. West, "Interval set clustering of web users with rough K-means", *J. Intell. Inform. Syst*., vol. 23, pp. 5–16, 2004.

[8] Z. Pawlak, *Rough Sets-Theoretical Aspects of reasoning about Data*, Kluwer Academic Publisher, Dordrecht, 1991, pp. 229-243.

[9] Z. Pawlak, "Concurrent Versus Sequential, *The Rough Sets Perspective"*, *Bulletin of the EATCS*, vol. 48, pp. 178-190, 1992.

[10] Z. Pawlak, "*Rough sets"*, *International Journal of Computer and Information Sciences,* vol 11, no 5, pp: 341-356, 1982.

[11] W.M Rand, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*, vol. 66, no 336, pp. 846-850, 1971.

[12] Sauravjoyti Sarmah and Dhruba K. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data", *International Journal of Computer Science Issues,* vol. 7, no 3, pp. 31-40, 2010.

[13] K. Thangadurai, "A Study on Rough Clustering". *Global Journal of Computer Science and Technology*, vol. 10, no 5, pp. 55-58, 2010.

**P. Ashok** was born in 1984 at Attur, Tamilnadu, India. He received his Master of Science in Computer Science from Periyar University, Salem, India in 2008. He obtained his M.Phil (Computer Science) Degree from Periyar University, Salem, India in 2009. Currently he pursuing his Ph.D. Degree in Computer Science from Bharathiar University. His research area of interests includes Data Mining, Rough Set, Fuzzy Logic and Bioinformatics.

He has been working as an Assistant Professor in Muthayammal College of arts and Science College, Rasipuram, Tamilnadu, India, since 2011. He presented 2 papers and published in IEEE international conferences and published 3 papers in international journals.

**Dr. G.M. Kadhar Nawaz** was born at Salem, Tamilnadu, India, He did master of computer application from Anna University, Chennai, India. He received Ph.D degree from Periyar University, Tamilnadu, India. Director, Department of Computer Applications.

He has been working as a Director, department of Computer Application, Sona College of technology, Salem, Tamilnadu, India. He presented and published over 30 papers in National, International conferences & Journals. He guiding 12 Research scholars in Anna University, Bharathiar University, Annai Therasa University. He visited Jordan, Singapore, Srilanka to present his research papers in International Conferences. His research area of interests includes Data Mining, Networking Fuzzy Logic and Image Processing.