Defence Science Journal, Vol. 66, No. 2, March 2016, pp. 113-121, DOI : 10.14429/dsj.66.9463

Outlier Detection Method on UCI Repository Dataset by Entropy Based Rough K-means

P. Ashok* and G.M Kadhar Nawaz

*Research scholar, Bharathiar University, Coimbatore - 641 046, India
*E- mail: ashokcutee@gmail.com

ABSTRACT

Rough set theory is used to handle uncertainty and incomplete information by applying two sets, lower and upper approximation. In this paper, the clustering process is improved by adapting the preliminary centroid selection method on rough K-means (RKM) algorithm. The entropy based rough K-means (ERKM) method is developed by adapting entropy based preliminary centroids selection on RKM and executed and also validated by cluster validity indexes. An example shows that the ERKM performs effectively by selection of entropy based preliminary centroid. In addition, Outlier detection is an important task in data mining and very much different from the rest of the objects in the cluster. Entropy based rough outlier factor (EROF) method is used to detect outlier effectively for yeast dataset. An example shows that EROF detects outlier effectively on protein localisation sites and ERKM clustering algorithm performed effectively. Further, experimental readings show that the ERKM and EROF method outperformed the other methods.

Keywords: Clustering process, entropy, rough set, outlier detection, validity index, data mining

1. INTRODUCTION

1.1 Rough Clustering Analysis

Cluster analysis[1,12] is the data analysis tool in data mining. It is used to grouping the objects into classes or clusters, so that objects within a cluster have high similarity and dissimilarity to objects in other clusters. The main resolution of clustering is to reduce the size and complexity of the dataset. The rough set[9,10] based clustering method is a mathematical tool to handle uncertainty and incomplete information by applying two accurate sets, the lower and upper approximation. The lower approximation is the set of objects definitely belonging to the imprecise concept. The upper approximation is the set of objects probably belonging to the imprecise concept.

Figure 1 show that the lower and upper approximation of the rough set is represented as *positive region and negative region*. The properties of the rough k-means clustering 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 is also member of the upper approximation of the same cluster.
- *Property 3*: A data object that does not belong to any lower approximation is member of at least two upper approximations.

In this study, entropy rough K-Means method is developed by adapting entropy measure based preliminary centroid selection method with rough k-means algorithm and detection of outliers and also validated the clustering algorithms using validity indexes are described with an example in the following section.

Received 16 December 2015, revised 28 January 2016, online published 22 March 2016

2. CLUSTERING PROCESS BY ENTROPY MEASURE

2.1 Methodology for Finding Preliminary Centroids

The preliminary centroids play a vital role in clustering process. Sometimes the algorithm produces bad clustering results due to selecting wrong centroids by randomly in the given data set. The entropy measure is used to select the preliminary centroids for rough clustering process is described with an example in the following section.

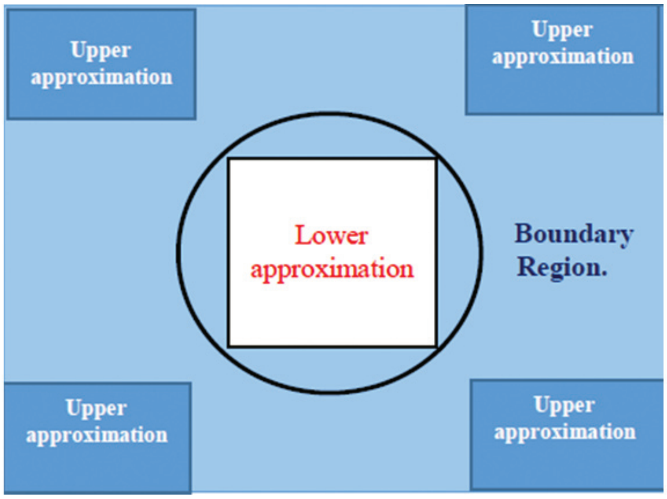


Figure 1. Regions of rough set.

2.2 Entropy based Preliminary Centroid Selection Method

Step 1: The objects having minimum entropy value are assigned as initial objects of lower approximation for each cluster and measured by the following steps

$$Epy_{ij} = -\sum_{j\varepsilon x}^{j\neq i}\Big(\big(S_{ij}\log_2 S_{ij}\big)+\big(1-S_{ij}\big)\log_2(1-S_{ij})\Big) \qquad (1)$$

S_{ij} represents the similarity value between the two objects and is calculated as follows.

$$s_{ij} = e^{-\alpha Dist_{ij}} \qquad (2)$$

Step 2: Dist$_{ij}$ represents the distance between the two objects and is calculated by using Eqn. (3)

$$Dist_{ij} = \sqrt{\sum_{m=1}^{n}(X_{im}-X_{jm})^2} \qquad (3)$$

The similarity value between any two points lies in the range of 0.0 to 1.0

$$\alpha = -\frac{\ln 0.5}{\overline{Dist}} \qquad (4)$$

α represents geometric constant and is determined that the ln (0.5) is calculated as -0.693, $\overline{Dist}$ represents the average distance of all the objects and is determined as follows

$$\overline{Dist} = \frac{1}{N}\sum_{k=1}^{N}\sum_{j>i}^{N} Dist_{ij} \qquad (5)$$

Step 3: The total entropy (*TE*) value of an object is calculated as follows.

$$TE_i = \sum_{j\in x}^{j\neq i} Epy_{ij} \qquad (6)$$

2.3 Example of Entropy based Centroid Selection Method

Table 1 is used as sample dataset for calculation of entropy measure and the distance between the two objects is calculated as follows. The value of *i* should be greater than *j*

$$Dist_{ij} = \sqrt{(x_1-x_1)^2+(x_2-x_2)^2} \qquad (7)$$

Table 1. Sample data set

Object No	Objects (*x*)	
	Attribute1 (X_1)	Attribute2 (X_2)
1	0	3
2	1	3
3	3	1
4	5	0
5	6	0
6	4	2
7	7	5
8	8	6
9	2	6
10	9	3

x and *y* represent an object (0 3) and calculate the distance between object1 and object2 as follows

$$d_{12} = \sqrt{(0-1)^2+(3-3)^2}$$
$$= 1.00$$

$$d_{13} = \sqrt{(0-3)^2+(3-1)^2}$$
$$= 3.60$$

Similarly, we can calculate the remaining object distances are listed in the matrix below

$$Dist_{ij} = \begin{bmatrix} 1.00 & 3.60 & 5.83 & 6.7 & 4.12 & 7.28 & 8.54 & 3.60 & 9.00 \\ & 2.82 & 5.00 & 5.83 & 3.16 & 6.32 & 7.61 & 3.16 & 8.00 \\ & & 2.23 & 3.16 & 1.41 & 5.66 & 7.07 & 5.09 & 6.32 \\ & & & 1.00 & 2.24 & 5.39 & 6.71 & 6.70 & 5.00 \\ & & & & 2.83 & 5.01 & 6.32 & 7.21 & 4.24 \\ & & & & & 4.24 & 5.65 & 4.47 & 5.1 \\ & & & & & & 1.41 & 5.01 & 2.83 \\ & & & & & & & 6.00 & 3.16 \\ & & & & & & & & 7.61 \end{bmatrix}$$

The mean distance is calculated by average of all object distances in the matrix

Mean distance = 220.9195 / 45
= 4.9093

The value of α can be calculated using the Eqn. (4)

α = - (ln 0.5) / 4.9093
= - (-0.6931) / 4.9093
= 0.1412

The similarity measure between two objects is calculated by using the Eqn. (4) as follows

$S_{12} = e^{(-0.1412*1.00)} = 0.8683$
$S_{13} = e^{(-0.1412*3.60)} = 0.6015$
$S_{14} = e^{(-0.1412*5.83)} = 0.4390$

Likewise, we can calculate the similarity values of remaining objects and are listed in the matrix below

$$S_{ij} = \begin{bmatrix} 0.86 & 0.60 & 0.43 & 0.38 & 0.55 & 0.35 & 0.29 & 0.60 & 0.28 \\ 0.86 & 0.67 & 0.49 & 0.43 & 0.63 & 0.40 & 0.34 & 0.64 & 0.32 \\ 0.60 & 0.67 & 0.73 & 0.64 & 0.82 & 0.45 & 0.37 & 0.49 & 0.41 \\ 0.44 & 0.49 & 0.73 & 0.87 & 0.73 & 0.47 & 0.39 & 0.39 & 0.49 \\ 0.39 & 0.44 & 0.64 & 0.87 & 0.67 & 0.49 & 0.41 & 0.36 & 0.55 \\ 0.56 & 0.64 & 0.82 & 0.73 & 0.67 & 0.55 & 0.45 & 0.53 & 0.49 \\ 0.36 & 0.41 & 0.45 & 0.47 & 0.49 & 0.55 & 0.82 & 0.49 & 0.67 \\ 0.30 & 0.34 & 0.37 & 0.39 & 0.41 & 0.45 & 0.82 & 0.43 & 0.64 \\ 0.60 & 0.64 & 0.49 & 0.39 & 0.36 & 0.53 & 0.49 & 0.43 & 0.34 \\ 0.29 & 0.32 & 0.41 & 0.49 & 0.55 & 0.49 & 0.68 & 0.64 & 0.34 \end{bmatrix}$$

The Entropy value between object 1 and object2 can be calculated by using the Eqn (1)

Epy_{12} = - (S_{12}* log$_2$ * S_{12}) + (1- S_{12})* (log$_2$ * 1- S_{12})
= - (0.86 *log2 (0.86) + 0.14 *log2 (0.14))
= - (-0.1871 + (-0.3971))
= 0.5842

Similarly, we can calculate the entropy value of remaining objects and are listed in the matrix below.

$$Epy_{ij} = \begin{bmatrix} 0.58 & 0.97 & 0.99 & 0.96 & 0.99 & 0.94 & 0.88 & 0.97 & 0.86 \\ 0.58 & 0.91 & 1.0 & 0.99 & 0.94 & 0.98 & 0.92 & 0.94 & 0.91 \\ 0.97 & 0.91 & 0.84 & 0.94 & 0.68 & 0.99 & 0.95 & 1.00 & 0.98 \\ 0.99 & 1.00 & 0.84 & 0.56 & 0.84 & 0.99 & 0.96 & 0.96 & 1.00 \\ 0.96 & 0.99 & 0.94 & 0.56 & 0.91 & 1.00 & 0.98 & 0.94 & 0.99 \\ 0.99 & 0.95 & 0.68 & 0.84 & 0.91 & 0.99 & 0.99 & 0.99 & 1.00 \\ 0.94 & 0.98 & 0.99 & 1.00 & 1.00 & 0.99 & 0.68 & 0.99 & 0.91 \\ 0.88 & 0.93 & 0.95 & 0.96 & 0.98 & 0.99 & 0.68 & 0.98 & 0.94 \\ 0.97 & 0.94 & 1.0 & 0.96 & 0.94 & 0.99 & 1.00 & 0.98 & 0.92 \\ 0.86 & 0.91 & 0.98 & 0.99 & 0.99 & 1.00 & 0.914 & 0.94 & 0.92 \end{bmatrix}$$

The Total entropy of an object is calculated as follows

TE_i = 0.58+0.97+0.99+0.96+0.99+0.94+0.88+0.97+0.86
= 8.1401

Similarly, we can calculate the total entropy (*TE*) values of remaining objects and are listed in the matrix below

$$TE_i = \begin{bmatrix} 8.1401 \\ 8.1608 \\ 8.2700 \\ 8.1598 \\ 8.2840 \\ 8.3540 \\ 8.4951 \\ 8.2986 \\ 8.7275 \\ 8.5156 \end{bmatrix}$$

Based on the maximum *TE* values of the objects are arranged in the order of 1,4,2,3,5,8,6,7,10,9 and assigned in the lower approximation of consecutive clusters and then the preliminary centroids for clustering process from Eqn. (8) are determined.

2.3.1 *Entropy based Rough K-Means Clustering*

Step 1: Allocate objects as preliminary objects in lower approximation of clusters by using Eqn. (6)

Step 2: Calculation of the centroids m_k are calculated as follows:

$$m_k = \left\{ \begin{array}{c} W_l \sum_{X_{k\varepsilon C_k}} \frac{Xn}{|C_k|} + W_B \sum_{X} \frac{X_n}{\left|C_k^B\right|} \; for C_k^B \neq \varphi \\ W_l \sum_{X_{k\varepsilon}C_k} \frac{X_n}{\left|C_k\right|} .otherwise \end{array} \right\} \quad (8)$$

w_l and w_b represent the lower approximation and boundary area of the cluster. The $\left|\underline{C_k}\right|$ and $\left|C_{Bk}\right| = \left|\underline{C_k} - \overline{C_k}\right|$ represent the numbers of objects in lower approximation and boundary area.

Step 3: Assign the objects to the lower and upper approximations.

(i) Object X_n, determines its closest mean m_h

$$d_{n,h}^{\min} = \min_{k=1...k} d(X_n, m_k) \quad (9)$$

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

(ii) Determine the mean m_t that is also close to X_n. Which is not farther away from X_n than $d(X_n, m_h)$. Here *T* is a given threshold:

$$T = \{t : d(X_n, m_k) - d(X_n, m_h) \leq \varepsilon \cap h \neq k\} \quad (10)$$

If $T = \varnothing$ (X_n is also close to at least one other mean m_t besides m_h) Then $X_{n \in}$ C_t, $\forall\, t \in T$.

Else $X_{n \in} Ch$.

Step 4: If the algorithms do not meet convergence criteria, continue with step otherwise stop the process.

2.4 Example of Entropy Rough K-Means Clustering Method

The sample dataset for rough k-means is listed in the Table 1.

Algorithm: Rough K-Means clustering algorithm

Step 1: Select the number of clusters *n* = 3

Assign initial objects to lower approximation of each cluster by using entropy measure

$$C_{lower} = \begin{bmatrix} 1 & - & - \\ - & - & 3 \\ 1 & - & - \\ - & 2 & - \\ - & 2 & - \\ 1 & - & - \\ - & 2 & - \\ - & - & 3 \\ 1 & - & - \\ - & - & 3 \end{bmatrix}$$

C_{lower} represents lower approximation of the cluster. The value 1, 2, 3 indicated in the matrix represents the objects placed in lower approximation of the cluster1, cluster2 and cluster3, respectively and are mentioned below

cluster1 = {obj1, obj3, obj6, obj9}
cluster2 = {obj4, obj5, obj7}
cluster3 = {obj2, obj8, obj10}

Step 2: Calculate the preliminary centroid (c_{ij} or mn_{ij}) of each clusters by using the following step

Centroid = (sum of column wise objects values in a cluster) / no of objects in a cluster

Determination of Centroid 1 for cluster 1:

C(1,1) = (*x*(1,1)+ *x*(3,1)+ *x*(6,1)+ *x*(9,1)) / 4
= (0+3+4+2)/4 = 2.25

C(1,2) = *x*(1,2)+ *x*(3,2)+ *x*(6,2)+ *x*(9,2) / 4
= (3+1+2+6) / 4 = 3.00

Determination of Centroid 2 for cluster 2:

C(2,1) = (*x*(4,1)+ *x*(5,1)+ *x*(7,1))/ 3
= (5+6+7)/3 = 6.00

C(2,2) = (*x*(4,2)+ *x*(5,2)+ *x*(7,2)) / 3
= (0+0+5) / 3 = 1.66

Determination of Centroid 3 for cluster 3:

C(3,1) = (*x*(2,1)+ *x*(8,1)+ *x*(10,1)) / 3

= (1+8+9)/3 = 6.00

$C(3,2) = (x(2,2) + x(8,2) + x(10,2))\ /\ 3$

= (3+6+3) /3 = 4.00

The calculated preliminary centroids values for all the clusters and are listed below

$$c_{ij} = \begin{bmatrix} 2.25 & 3.00 \\ 6.00 & 1.66 \\ 6.00 & 4.00 \end{bmatrix}$$

Step 3: Calculation of distance between objects and centroids.

Distance between object1 and centroid1 is calculated as follows

$$d(x_1, c_1) = d_{11} = \sqrt{(0-2.25)^2 + (3-3.00)^2}$$
$$= 2.25$$
$$d_{12} = \sqrt{(0-6.00)^2 + (3-1.66)^2}$$
$$= 6.14$$
$$d_{13} = \sqrt{(0-6.00)^2 + (3-4.00)^2}$$
$$= 6.08$$

Similarly, the distance between remaining objects (obj2 to obj10) with three centroids are calculated and listed in the matrix below

$$d_{ij} = \begin{bmatrix} 2.25 & 6.14 & 6.08 \\ 1.25 & 5.17 & 5.09 \\ 2.13 & 3.07 & 4.24 \\ 4.06 & 1.93 & 4.12 \\ 4.80 & 1.66 & 4.00 \\ 2.01 & 2.02 & 2.82 \\ 5.15 & 3.48 & 1.41 \\ 6.48 & 4.77 & 2.81 \\ 3.01 & 5.90 & 4.47 \\ 6.75 & 3.28 & 3.16 \end{bmatrix}$$

Step 4: Assignment of objects in appropriate clusters

The distance between object1 and centroid1 is minimum. So object1 is placed in cluster 1 and then the object placed in lower or upper approximation is identified by using the following steps.

In each row from above matrix, the d_{ij} is divided by by minimum distance value of that row and the object whose difference is less than threshold is found and the object only in the upper approximation of more than one clusters is assigned.

Assignment of objects in lower and upper approximation

```
If (diff > threshold)
    (threshold = 0.4)
    Objects in lower and upper approximation of the
    same cluster can be placed.
Else
    Objects only in upper approximation of more than
    one clusters can be placed.
```

If ($diff = (d_{12}/d_{11})$)

$= 6.14/2.25$

$= 2.272 >$ threshold

If ($diff = (d_{13}/d_{11})$) > threshold

$= 6.08/2.25$

$= 2.70 >$ threshold

The *diff* value is greater than threshold value for object1. Hence the object1 is placed in lower approximation and upper approximation of cluster1.

By property 3, if the data object does not belong (*diff* < threshold) to any lower approximation, it might be the member of at least two upper approximations.

Similarly, the assignment of remaining objects (obj2 to obj10) into the lower and upper approximations of the appropriate clusters can be determined and the assignment of objects in the lower approximation of specific clusters are listed in the matrix below

$$c_{Lower} = \begin{bmatrix} 1 & - & - \\ 1 & - & - \\ 1 & - & - \\ - & 2 & - \\ - & 2 & - \\ 1 & - & - \\ - & - & 3 \\ - & - & 3 \\ 1 & - & - \\ - & - & 3 \end{bmatrix}$$

By rough property2: The data object in the lower approximation of a cluster is also member of the upper approximation of the same cluster.

$$c_{Upper} = \begin{bmatrix} 1 & - & - \\ 1 & - & - \\ 1 & - & - \\ - & 2 & - \\ - & 2 & - \\ 1 & - & - \\ - & - & 3 \\ - & - & 3 \\ 1 & - & - \\ - & - & 3 \end{bmatrix}$$

C_{upper} represents upper approximation of the clusters and finally the objects are placed in the appropriate clusters.

cluster1 = {obj1, obj2, obj3, obj6, obj9}

cluster2 = {obj4, obj5}

cluster3 = {obj7, obj8, obj10}

Step 5: Checking the convergence criteria of the clustering algorithm

Target = (new centroid – old centroid)

If convergence criterion (Target < 0.01) is met, the algorithm has be stopped

Else go to step2.

3. OUTLIER DETECTION TECHNIQUES

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

3.1 Detecting outliers on Cellular Localisation Sites of Proteins

Protein is a macro nutrient composed of amino acids that is essential for the proper growth and function of the human body. While the body can construct several amino acids required for protein production, a set of fundamental amino acids needs to be obtained from sources. In the yeast dataset, the defective proteins (abnormal proteins) are considered as outliers, which are identified by the entropy rough outlier factor (EROF) method on ERKM clustering algorithm.

3.1.1 Detection of outlier by Entropy Rough Outlier Factor

The outlier is identified by entropy based rough outlier factor (*EROF*), which indicates the degree of outlierness for every object in the yeast dataset. The *EROF* based outlier detection method is defined as follows

The entropy value of an object is calculated as follows

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

P_i represents the distance between the object and centroid.

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

$E_i^{\max}$ and $E_i^{\min}$ represent maximum and minimum entropy value of i^{th} object. $|C_i|$ denotes the cardinality of cluster c_i.

For any object $x \in c_i$, $EROF_i$ is entropy based rough outlier factor of cluster c_i and it is calculated for each cluster separately, the entropy value of each objects in the clusters is compared by EROF of that cluster. If $E_i < EROF_i$ then object$_i$ is consider as ER-based outlier. The detected outlier objects are considered as abnormal protein or defective protein in cellular localisation sites of protein. The outlier detection algorithm is defined as follows

3.1.2 Outlier Detection Algorithm

Input:

$|U|$ = total number of objects,

C = the number of clusters

Output:

Outliers and Objects in appropriate clusters

Steps

```
For every $x_i \in U$
    For j = 1 to n
        Calculate the entropy value $E_i$ of each object by using Eqn.(11)
    End
For i =1to c
    Calculate $EROF_i$ of the cluster by using Eqn. (12)
End
For i = 1 to c
    For j = 1 to n
        If $EROF_i > E_j$, then $object_j$ is called as outlier
    End
End
End
```

3.1.3 Numerical Example of Outlier Detection Method

Table 1 is used as sample dataset for outlier detection method. The calculated centroid values are

$$c_{ij} = \begin{bmatrix} 2.25 & 3.00 \\ 6.00 & 1.66 \\ 6.00 & 4.00 \end{bmatrix}$$

The entropy value of an objects in the clusters is calculated by using Eqn (11).

cluster1 = {obj1, obj2, obj3, obj6, obj9}

The calculated distance between objects and centroid1 in cluster1 is {2.25, 1.25, 2.13, 2.01 and 3.01}.

The entropy value between object1 and centroid in cluster1 (E_{ij}) is calculated as follows

E_{11} = 2.25*log 2.25 = 0.7924
E_{12} = 1.25*log 1.25 = 0.1211
E_{13} = 2.13*log2.13 = 0.6994
E_{14} = 2.01*log2.01 = 0.6094
E_{15} = 3.01*log3.01 = 1.4404

The EROF of cluster1 is calculated by using Eqn. (12) as follows

EROF$_1$ = (1.4404-0.1211)* 1-(5/10)
= 0.6595

The entropy value of objects 2 and 4 are less than *EROF* value of cluster1 and is detected as outliers.

Objects in cluster2 = {obj4, obj5}. The entropy values of objects in the cluster2 are calculated as follows

E_{21} = 1.93*log1.93 = 0.5511
E_{22} = 1.66*log1.66 = 0.365

The EROF value of the cluster2 is calculated as follows

EROF$_2$ = (0.1854)*(0.8)
= 0.1483

Hence, there is no outlier in cluster 2.

Objects in cluster3 = {obj7, obj8 and obj10}. The entropy value for 3 objects in the cluster 3 is calculated as follows

E_{31} = 1.41*log1.41 = 0.2103
E_{32} = 2.81*log2.81 = 1.2621
E_{33} = 3.16*log3.16 = 1.5790

The EROF value of cluster 3 is calculated as follows

EROF$_3$= (1.3687)*(0.7)
= 0.9580

Hence, the entropy value of an object 7 is less than EROF value of cluster3 and is identified as outlier. Likewise the outliers are detected in the protein localisation sites using EROF method.

4. CLUSTER VALIDATION TECHNIQUES

Clustering validity measure[3] is used to evaluate the quality of clustering results. It is also used to find the optimal number of clusters. The rough clustering methods are validated by two cluster validity measures namely

- Rand index
- Adjusted Rand index

4.1 Rand Index

The rand index (*RI*)[11] proposed by Rand is a popular cluster validity measure used for cluster validation and computed as follows

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

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

4.2 Adjusted Rand Index

The adjusted Rand index 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.

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

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

4.3 Numerical Example of Cluster Validity Measure

The rand index of the clustering process needs the contingency table and it can be developed during the clustering process. Table 2 represents the contingency table of the clustering process.

Table 2. Contingency table for comparing partitions *u* and *v*

		V				
	Clusters	***v1***	***v2***	**…**	***Vc***	**Total**
Class	u_1	t_{11}	t_{12}	…	t_{1c}	$t_{1.}$
	u_2	t_{21}	t_{22}	…	t_{2c}	$t_{2.}$
	.	.	.	.	.	.
	.	.	.	.	.	.
	u_r	t_{R1}	t_{R2}	…	t_{RC}	t_{R}.
Total		$t_{.1}$	$t_{.2}$	…	$t_{.C}$	$t_{..}= n$

t_{RC}, represents the number of objects that were classified in the *r*th subset of partition *R* and in the *c*th subset of partition *C*. From the total number of possible combinations of pairs $\binom{n}{2}$ from a given set, the results in four different types of pairs can be represented in the rand index as follows

$$n = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

The Rand index can be calculated by substituting the values from Table 3 in the Eqn. (13). The value of a, b, c and d can be calculated as follows

$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1) \tag{16}$$

For *I* =1 to 2 and *j*=1 to 2

$$a = \frac{1}{2}\begin{pmatrix} n(1,1)*(n(1,1)-1)+n(1,2)*(n(1,2)-1)+ \\ n(2,1)*(n(2,1)-1)+n(2,2)*(n(2,2)-1) \end{pmatrix}$$

= (4*(4-1) + 1*(1 – 1) + 3*(3-1) + 2*(2-1))/2

a = 10

Table 3. Example for contingency table

Segment	**Clusters**		
	***C*₁**	***C*₂**	**Total**
Segment1	4	1	5
Segment2	2	3	5
Total	6	4	*N* =10

The value of *b* can be calculated as

$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{i.}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \tag{17}$$

$$\sum_{j=1}^{k'} n_{i.}^{2} = \left\{(n(1,1)+n(1,2))^2+(n(2,1)+n(2,2))^2\right\}$$

= (4+1)^2 + (2+3)^2

= 50

$$\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{.j}^{2} = \left\{n(1,1)^2+n(1,2)^2+n(2,1)^2+n(2,2)^2\right\}$$

= 4^2 +1^2+2^2+3^2

= 30

b = (50-30) / 2

b = 10

The value of c can be calculated as

$$c = \frac{1}{2}\left(\sum_{i=1}^{K} n_{i.}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \tag{18}$$

$$n_{.j}^2 = \left\{(n(1,1)+n(2,1))^2+(n(1,2)+n(2,2))^2\right\}$$

= (4+2)^2+ (1+3)^2

= 52

C = (52-30)/ 2

= 11

The value of d can be calculated as

$$d=\frac{1}{2}\left(N^2+\sum_{i=1}^{K}\sum_{j=1}^{K'}n_{ij}^2-\left(\sum_{i=1}^{K}n_{i.}^2+\sum_{j=1}^{K'}n_{.j}^2\right)\right) \qquad (19)$$

d =1/2((10^2 +30) – (50+52))

d = 14

By using the value of a, b, c and d the rand index is calculated as follows

RI = (10+14) / (10+10+11+14)

= 0.588

RI = 0.588

4.4 Numerical Example of Adjusted Rand index

The Adjusted rand index can be calculated by using the Eqn. (14-15). The value of a = 10, b = 10, c = 11, d = 14 are calculated by rand index method. The Adjusted rand index is calculated as follows

ARI = 2((10*14) - (10*11)) / (10+10) * (10+14) + (10+11) * (11+14)

= 60 / (480+525)

= 0.0597

ARI =0.0597

5. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental result analysis is carried out by considering different data sets from UCI data repository and the rough clustering algorithms are validated through cluster validity measure.

5.1 Datasets

The dataset of the clustering process can be downloaded from the UCI repository www.ucirepository/dataset/ (Table 4.).

Table 4. UCI reposistory dataset

Data set	No of objects	No of features	No of clusters	No of objects in per cluster
Yeast	1484	8	10	463-429-244-163-51-44-37-30-20-05
Iris	150	4	3	50-50-50
E-coli	336	3	8	143-77-52-35-20-05-02-02
Glass	214	9	6	70-76-17-13-9-29
Wine	178	13	3	59-71-48
Diabetes	768	8	2	500-268

5.2 Accuracy of Clustering Method by Selection of Preliminary Centroid Method

The entropy based preliminary centroid method was discussed and ERKM clustering algorithms was developed (section.2.5) The rough K-Means clustering algorithm is executed with entropy based preliminary centroid selection method. The accuracy of clustering method is determined by comparing the clustering results obtained by the experiment with the clusters already available in the UCI dataset. The evaluated accuracy rate of the clustering process is depicted in the Table 5.

Table 5 show that ERKM algorithm achieves better clustering results because of the highest accuracy rate obtained by selection of better preliminary cluster centroids for clustering process.

Table 5. Accuracy of the clustering algorithms

Dataset	Clusters	Selection of preliminary centroids by	
		Random method (Accuracy rate %)	Entropy method (Accuracy rate %)
Iris	3	67	93
Yeast	10	70	92
Ecoli	8	69	96
Glass	6	55	85
Wine	3	70	90
Diabetes	2	59	92

5.3 Detecting Outlier on Cellular Localisation Sites of Protein

The outliers are identified by the entropy based rough outlier factor (EROF) method and it was discussed early section. It is able to identify the outliers on protein localisation sites (yeast dataset) very well. The ERKM algorithm is executed with EROF outlier detection method on yeast dataset. The yeast dataset is defined as follows

It has been used to find localisation site of protein. The data set contains 1484 records (objects). It has eight features (attributes) are mcg, gvh, alm, mit, erl, pox, vac, nuc. Proteins are classified into various clusters are cytosolic or cytoskeletal, nuclear, mitochondrial, membrane protein without N-terminal signal, membrane protein with uncleaved signal, membrane protein with cleaved signal, extracellular, vacuolar, peroxisomal, Endoplasmic reticulum lumen. The objects in the yeast dataset are segregated by a number of clusters with outliers by ERKM with EROF method and are represented in the Table. 6.

The solid balls in the Fig. 2 represent data objects in the appropriate clusters and the red solid balls represent outliers (defective protein) identified by EROF method. In the protein localisation sites, the defective proteins are considered as outliers, which are identified by the EROF methods with ERKM clustering method. The defective proteins are may decrease the performance of clustering process. After removal of outliers in the protein localisation sites improves the performance of clustering process.

Likewise the EROF outlier detection method is executed on some other UCI repository dataset are iris, wine, Ecoli and Diabetes. The dataset wise outliers are detected and are denoted in the Table 7. The solid balls in Fig. 3 (a)-(d) represent data objects in the appropriate clusters and the red solid balls represent outliers in the appropriate clusters identified by EROF method.

5.4 Cluster Validity Measure

5.4.1 Rand index Analysis

The RKM and ERKM clustering algorithms are validated

Table 6. Cluster wise outlier detection on yeast dataset

Outlier detection on yeast dataset by EROF			
Cluster name	**Objects in cluster**	**Cluster wise outlier**	**Outliers in a cluster (%)**
Cytosolic (CYT)	463	10	4.63
Nuclear (NUC)	429	14	3.20
Mitochondrial (MIT)	224	10	2.24
Membrane 3 (ME3)	163	12	1.66
Membrane 2 (ME2)	51	5	2.33
Membrane 1 (ME1)	44	6	8.00
Extracellular (EXC)	37	4	7.36
Vacuolar (VAC)	30	4	7.50
Peroxisomal (POX)	20	2	10.0
Endoplasmic reticulum lumen (ERL)	5	1	20.0
Total	1484	68	4.50

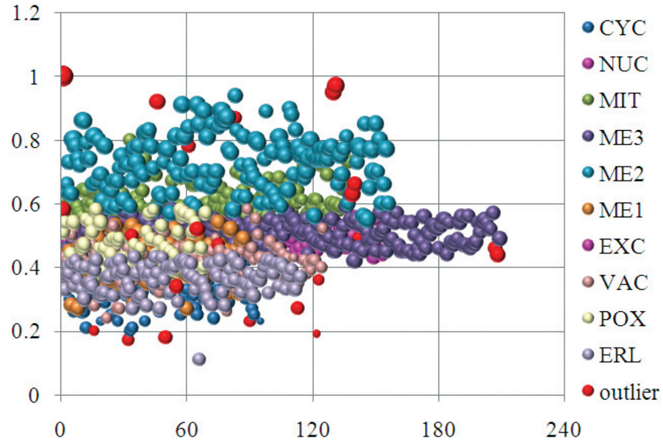


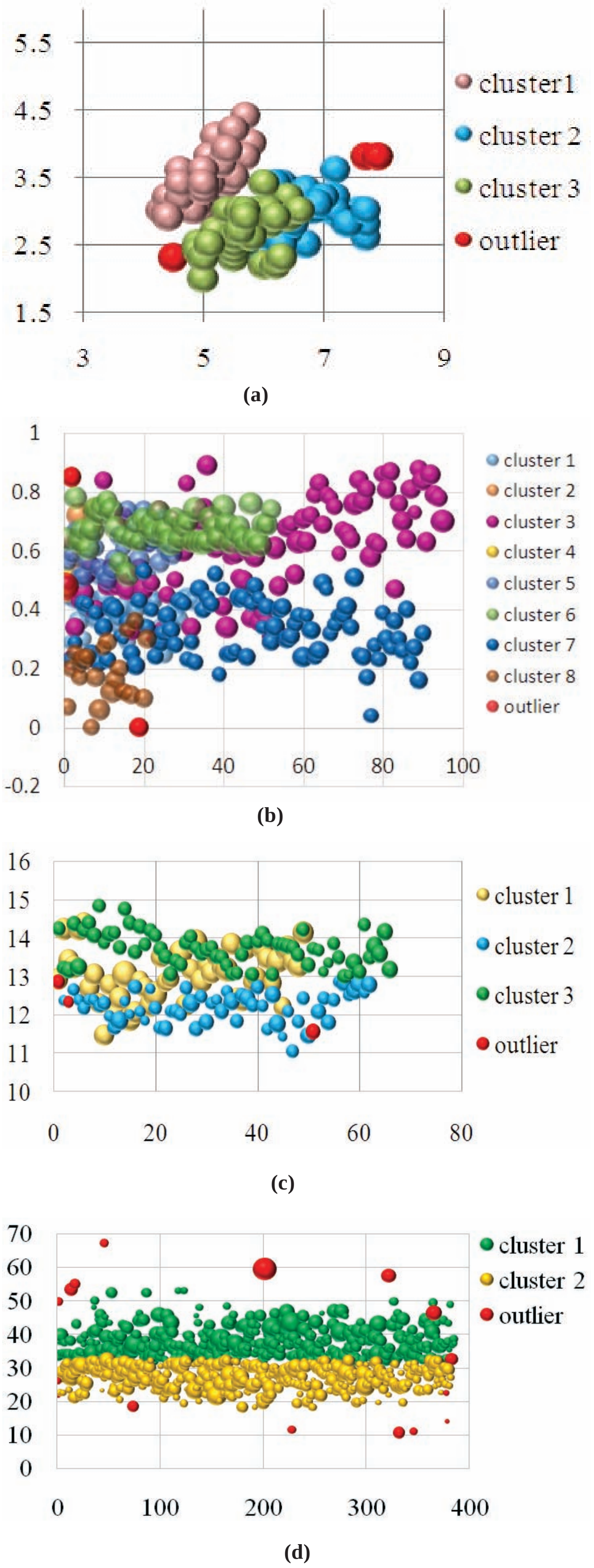
Figure 2. Detecting outliers on protein localisation sites.

Table 7. Outlier detection on UCI repository dataset

EROF based outlier detection Method				
Datasets	**Clusters**	**No of objects**	**Outliers**	**Outliers in per cluster (%)**
Yeast	10	1484	68	6.80
Iris	3	150	3	1.00
E-coli	8	336	8	1.00
Glass	6	214	10	1.66
Wine	3	178	7	2.33
diabetes	2	768	16	8.00

and compared by rand index and adjusted rand index with yeast dataset. They are executed by changing the cluster value from 3 to 30 and the obtained rand index values are listed in the Table 8.

It show that the rough clustering algorithms performed very well and obtained different rand index and adjusted rand index values. But the ERKM algorithm obtained higher rand



(a)

(b)

(c)

(d)

Figure 3. Detecting outliers on dataset (a) Iris, (b) Ecoli (c) wine and (d) diabetes.

Table 8. Rand index analysis

Clusters	Rand index		Adjusted rand index	
	RKM	ERKM	RKM	ERKM
3	0.5685	0.709	0.281	0.338
6	0.6827	0.662	0.342	0.354
9	0.6811	0.698	0.354	0.411
12	0.6713	0.685	0.302	0.341
15	0.6714	0.706	0.287	0.331
18	0.6832	0.681	0.301	0.303
21	0.6814	0.678	0.295	0.281
24	0.6818	0.715	0.292	0.303
27	0.6839	0.678	0.296	0.276
30	0.6806	0.699	0.279	0.311

index and adjusted rand index values than RKM from the most of the obtained results.

6. CONCLUSION

In this study, the proposed method entropy rough K-Means clustering algorithm performed effectively and obtained higher clustering accuracy rate than other methods when executed with UCI repository dataset. The RKM and ERKM clustering algorithms are executed with UCI datasets and evaluated by rand and adjusted rand validity indexes. The experimental results shows that the ERKM clustering algorithm delivers better results than RKM clustering algorithms and increases the performance of clustering method. The EROF based outlier detection method can detect outliers' effectively on protein localisation sites and improves the quality of the rough clustering method after removal of outlier in the dataset. The rough clustering is hybrid with fuzzy clustering method to improve the clustering process and detecting outliers is our future work.

REFERENCES

1. Ashok, P.; Nawaz, Kadhar G.M.; Elayaraja, E. & Vadivel, V. Improved performance of unsupervised method by renovated K-means. *Int. J. Adv. Study Comput. Sci. Eng.*, 2013, **2**(1), 41-47.
2. Ashok, P.; Nawaz, Kadhar G.M. & Elayaraja, E. Outliers detection on protein localization sites by partitional clustering methods. *In* proceedings of Pattern Recognition, Informatics and Medical Engineering (PRIME), Salem, Tamilnadu, Feb 2013.
3. Davies D.L. & Bouldin, D.W. A Cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1979, **1**(2), 224-227.
   doi: 10.1109/TPAMI.1979.4766909
4. Georg, Peters. Some refinements of rough k-means clustering. *Pattern Recognition*, 2006, **39**(8), 1481-1491.
   doi: 10.1016/j.patcog.2006.02.002
5. Voges, Kevin E. Research techniques derived from rough sets theory, rough classification and rough clustering. University of Canterbury, Christchurch, 2005, 437-444.
6. Lingras, P. & West, C. Interval set clustering of web users with rough K-means. *J. Intell. Info. Sys.*, 2004, **23**(23), 5–16.
   doi: 10.1023/B:JIIS.0000029668.88665.1a
7. Lingras, P. Rough set clustering for web mining. *In* Proceedings of IEEE International Conference on Fuzzy Systems. Honolulu, HI, May 2002, 5-16.
   doi: 10.1109/fuzz.2002.1006647
8. Pawlak Z. Rough sets. *Int. J. Comput. Info. Sci.*, 1982 **11**(5), 341-356.
9. Pawlak Z. Rough sets-theoretical aspects of reasoning about data. Kluwer Academic Publisher, Kluwer, Dordrecht, Netherlands, 1991, 229-243.
10. Pawlak, Z. Concurrent Versus sequential - the rough sets perspective. *Bulletin EATCS*, 1992, **48**, 178-190.
11. Rand W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, 1971, **66**(336), 846-850.
    doi: 10.1080/01621459.1971.10482356
12. Sarmah, Sauravjoyti & Bhattacharyya, Dhruba K. An effective technique for clustering incremental gene expression data. *Int. J. Comput. Sci. Iss.*, 2010, 7(3), 31-41.
13. Thangadurai, K. & Uma, M. A study on rough clustering. *Global J. Comput. Sci. Technol.*, 2010, **10**(5), 55-58.

CONTRIBUTORS

Mr P. Ashok received his MSc (Computer Science) and MPhil (Computer Science) from Periyar University, Salem, India in 2008 and 2009, respctively. Currently pursuing his PhD (Computer Science) from Bharathiar University. His research area of interests includes : Data mining, rough set, fuzzy logic and bioinformatics.
In the current study, he has contributed in the designing of the clustering algorithms. The coding and the implementation of the ideas were also carried out by him.

Dr G.M. Kadhar Nawaz, received his PhD (Computer Science) from Periyar University, Salem and MCA from Madras University, Chennai. Currently working as a Director, Department of Computer Applications, Bharathiar University. He presented and published over 30 papers in National, International conferences and Journals. His area of interest is N/W security, stegnography and cryptography. His research area of interests includes : Data mining, networking fuzzy logic and image processing.
In the current study, he has done data analysis. The experimental results were analyzed, interpreted and evaluated by him.