HYBRID CLUSTERING ALGORITHM USING POSSIBILISTIC ROUGH C-MEANS

*ANURADHA J, ANVESHA SINHA AND TRIPATHY B K

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India

ABSTRACT

A cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to objects in another cluster. This clustering mechanism ensures high intra-class similarity but low inter-class similarity which can be achieved by the c-means architecture. Though much clustering algorithm has evolved since Hard CMeans, Fuzzy C Means (FCM) and Rough C Means (RCM) are widely used in applications for their superiority in handling vague and uncertain data. The probabilistic approach to clustering techniques can handle noisy data. In this paper we propose a hybrid clustering model on Possibilistic Rough C Means that can handle uncertainty even in presence of noisy data. The theory of Rough sets which has emerged as one of the most efficient tools of the decade can generate a pair of set with lower and upper approximations of the objects determined to establish the cluster. Possibilistic theory to cluster generates the typicality or possibility value of the objects that belongs to the rough cluster. This model will be more efficient and fool proof than each of the individual models. Experimental analysis reveals that the proposed algorithm enables to streamline the clustering process and have more precise clustering mechanisms.

KEYWORDS: Clustering, Fuzzy C Means, Rough C-Means, Possibilistic C-Means.

ANURADHA J
School of Computing Science and Engineering,
VIT University, Vellore – 632014, Tamil Nadu, India
1 INTRODUCTION

Clustering is one of the most important techniques that has made a tremendous development in computing history. Clustering, is the process of grouping of data in such a way that the objects in a cluster are as similar as possible and those in different clusters are as dissimilar as possible. Figure 1.1 shows the formation of clusters having two groups viz. cluster 1 and cluster 2. The quality of the cluster is determined by this nature of the cluster partition. The degree of similarity with in cluster and the dissimilarity across the clusters together determine the validity of the cluster. The degree of similarity is measured in terms of intra cluster distance and the degree of dissimilarity is measured by intra cluster distance. C - Means algorithm is one such clustering mechanism which ensures high intra-class similarity but low inter-class similarity.

There exist different approaches for formation of clusters. It is broadly classified as divisive and agglomerative clustering. In divisive clustering techniques, starting with the Hard C-means, a number of algorithms have been suggested and reviewed in literature with each algorithm trying to overcome the disadvantages of the previous algorithms. The fuzzy C-means algorithm, rough C-means algorithm, possibilistic C-means algorithm are more efficient and the enhanced versions like dynamic, adaptable and hybrid models are competitive each other in its performance showing the superiority in different dimensions by overcoming the disadvantage of each other. However, it should also be noted that a collaborative approach has also been used to derive algorithms of better efficiency. Two or more algorithms are combined or collaborated together to create an algorithm which is more efficient by getting the benefits from more than one technique). For example-the rough-fuzzy collaborative clustering (Mitra et al, 2006) is a collaboration between the rough and fuzzy C-means algorithms and seeks to overcome both of their disadvantages. The latest in line has been the possibilistic fuzzy c-means algorithms which has been an amalgamation of the possibilistic C-means algorithm and the Fuzzy C-means algorithm. The above mentioned algorithms are sensitive to noisy data and may form coincidental clusters. In this direction we propose a Possibilistic Rough C-Means (PRCM) algorithm in this paper that overcomes the drawback of previous algorithms. It applies the concept of typicality to rough technique in the formation of clusters. The superiority of the proposed algorithm is measured using various cluster validity indices. The results are analyzed considering different parameter setting in cluster formation.

2 LITERATURE SURVEY

In the past few decades, clustering algorithms have evolved tremendously. However, the concept of high intra-cluster similarity and low inter-cluster similarity remains the same. Hard C – Means which is the foremost of all the clustering algorithms, is not much useful in real life situations. It is because of the fact that it uses crisp values i.e. 0 or 1 for representing belongingness of objects to the cluster. The cluster boundary is rigid and the object may belong to at most one cluster. This constraint pulls down the performance of
HCM on real life data sets. Hence HCM is not much suitable for real life applications. The Fuzzy C Means (FCM) algorithm introduced in\(^8\) forms overlapping clusters based on membership value of the objects belonging to all the clusters. There were many improvements over the basic FCM algorithm that is more flexible and more adaptable which enhances its performance\(^7,9\). FCM algorithm is widely used in image processing, pattern recognition, data mining and in various applications of science and engineering. In this algorithm, membership of objects to each of the clusters is calculated. The membership of an object \(k\) belonging to \(i\)th is given by (2.1)

\[
\mu_{ik} = \left[ \sum_{j=1}^{c} \left( \frac{d_{ik}}{d_{jk}} \right)^{2/m-1} \right]^{-1} \tag{2.1}
\]

Clearly, membership of each object to a cluster depends on the distance of the object with the cluster as well as the distance of the object from all the other clusters. It wouldn’t be wrong to say that these memberships do not actually represent the belongingness of the object to a cluster; rather it represents the degree of sharing. The memberships are relative quantities because they depend on the distance of the object to all the other clusters and therefore on the total number of clusters. FCM requires the number of clusters to be specified initially and forms exactly that many number of clusters even though clusters may be identical. This leads to the problem of low cluster validity and formation of coincident clusters\(^11\). Another problem that is of significant concern in case of FCM is the presence of noisy data\(^10\). It fails to identify outliers in the data which can be explained with the help of the following figure.

![Figure 2.1 Presence of noisy data](image)

From the figure 2.1, it is quite obvious that A and B are outliers and differently separated from each cluster. However, FCM assigns a membership value of 0.5 to both A and B\(^10\). It is so as for the calculation of membership of B, instead of taking the smaller distance from the two clusters, distances from both the clusters are taken. FCM, therefore, has a probabilistic constraint i.e. the sum of memberships of an object to all the clusters should be 1.

\[
\sum_{i=1}^{c} \mu_{ik} = 1 \tag{2.2}
\]

To overcome the drawbacks of FCM, the probabilistic constraint was removed by introducing the concept of typicality or possibility. The only constraint in this approach is that the typicality values must lie in the interval [0, 1]. Typicality values are
absolute and denote degrees of belongingness of an object to a cluster. This algorithm is called the Possibilistic C-Means (PCM). Moreover, it has been referred to as a "mode-seeking" algorithm because even if the number of clusters is unknown, good clusters (dense regions) can be formed with the proper estimation of parameters. This ensures that the clusters formed are valid and not identical. Another clustering algorithm which has gained attention recently is the Rough C- Means (RCM) introduced by Lingras and west (2004). RCM is popular for their capability in handling uncertain objects and it does not require any prior knowledge about the data. It has been presented as a collaborative clustering technique along with FCM, jointly referred as RFCM. Rough sets are used extensively in areas of data mining. Rough C- Means uses the concept of rough sets to address the in discernibility relation between objects. It classifies the data into two distinct groups for each cluster – namely the lower approximation and the boundary region. Thus, only those objects belonging completely to a cluster can be separated from those which are ambiguous. The boundaries of clusters overlaps which makes sure that the clustering does not become crisp and hence further improvements in the cluster formation is possible. We can further diversify this technique to Rough C-means (RCM) and probabilistic C- means (PCM) so as to establish the procurement of the C-means value. The theory of Rough sets which has emerged as one of the most important tools of the decade can be applied to the further classify the C-mean concept such that the lower approximations and boundary region of the object set are determined to consequently establish the cluster. Possibilistic C-means is also a method of establishing C –means wherein a number of possibilities are generated, these possibilities are further evaluated through concepts of typicality to determine the C-value. We attempted to create a hybrid of these two prominent models to be termed as "Possibilistic Rough C-means" (PRCM) such that it incorporates the best of both models. This model is to be established as more efficient than each of the individual models. It will also enable to streamline the clustering process and have more precise clustering mechanisms.

3 THE PROPOSED ALGORITHM
Possibilistic Rough C-Means (PRCM) algorithm is a combination of PCM and RCM. It applies the concept of typicality to rough sets for the formation of new clusters. Basically, we focus on clustering objects on the basis of typicality which then is assigned to either the lower or boundary region of a cluster. This is done to ensure that all the objects that are atypical (decided on the basis of a threshold value) have chances to be assigned to any of the clusters rather than to both. The algorithm of PRCM is given below.

**PRCM Algorithm**

1. Assign the initial centroids. These are chosen randomly from the dataset.
2. Calculate the distances of all the other data points from the centroids.
3. Assign each data object $x_i$ to the lower approximation $\Delta X_i$ or boundary regions $\overline{\Delta X_i}$ - $\Delta X_i$, $\overline{\Delta X_j}$ - $\Delta X_j$ of cluster pairs $X_i$ and $X_j$ by computing the difference in its distances $d_{ik}$ and $d_{jk}$ from the cluster centroid pairs $v_i$ and $v_j$, respectively as follows.
4. Let $d_{ik}$ be first minimum and $d_{jk}$ be the next minimum.
   If $|d_{ik} - d_{jk}|$ is less than $\theta$ (threshold value), then
   $$x_i \in \overline{\Delta X_i} - \Delta X_i \text{ and } x_i \in \overline{\Delta X_j} - \Delta X_j \text{ and } x_i \text{ cannot be a member of lower approximation of any cluster.}$$
   else
   $$x_i \in \overline{\Delta X_i}$$
5. For each data point belonging to the cluster, a typicality value is calculated by
\[
t_{ik} = \frac{1}{(1 + \frac{d_{ik}^2}{\gamma_i})}
\]  
where \(\gamma_i\) is calculated by

\[
\gamma_i = \frac{\sum_k d_{ik}}{|AX_i - AX_i|}
\]  
(For the objects in boundary region)

\[
\gamma_i = \frac{\sum_k d_{ik}}{|AX_i|}
\]  
(For the objects in lower approximation)  
(3.2)

6. Compute the new centre for each cluster \(X_i\), applying (3.4), (3.5) and (3.6).

\[
v_i = \frac{\sum \mathbf{x}_k \in [AX_i - AX_i]^{x_k} \cdot t_{ik}}{|AX_i - AX_i|}, \quad \text{if } |AX_i - AX_i| \neq 0 \quad \land \quad |AX_i| = 0
\]  
(3.4)

\[
v_i = \frac{\sum \mathbf{x}_k \in AX_i^{x_k}}{|AX_i|}, \quad \text{if } |AX_i - AX_i| = 0 \quad \land \quad |AX_i| \neq 0
\]  
(3.5)

\[
v_i = w_{low} \frac{\sum \mathbf{x}_k \in AX_i^{x_k}}{|AX_i|} + w_{up} \frac{\sum \mathbf{x}_k \in AX_i - AX_i^{x_k} \cdot t_{ik}}{|AX_i - AX_i|}, \quad \text{if } |AX_i - AX_i| \neq 0 \quad \land \quad |AX_i| \neq 0
\]  
(3.6)

7. Repeat Steps 2 – 6 until there is no change in centroids.

The objective function for the proposed algorithm is given by (3.7)

\[
J_{PRCM} = \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{i=1}^{c} (\tau_{ki} \mu_{li} (\overline{wd}_{ki} + \overline{wd}_{li}) + (\tau_{ki} \tau_{li} + \mu_{ki} \mu_{li}) D_{kl})
\]  
(3.7)

where \(d_{kl} = \mathbf{x}_k - \mathbf{v}_l\), \(D_{kl} = ||\mathbf{x}_k - \mathbf{x}_l||\), \(\tau_{ki}\) is the typical belongingness of an object \(\mathbf{x}_k\) to lower approximation of \(AX_i\) and \(\mu_{ki}\) is the typical belongingness of an object \(\mathbf{x}_k\) to the boundary of \(AX_i - AX_i\).

The typicality of the object to a cluster helps to identify its proper cluster and it works well even in presence of noisy data. The experimental analysis of the proposed PRCM is conducted on various data sets with different cluster validity measures.

4 TESTING AND VALIDATION
The validity of the cluster formed by PRCM is measured against various metrics and it is tested on various data sets given in table 1.

4.1 DB INDEX
Davies Bouldin (DB) index and Dunn (D) index given by Bezdek 8 is for validating the formation of the clusters. DB is the ratio of the sum of within-cluster distances to between-cluster distance. This factor should be low because the objects within a cluster should be close to each other and far to those belonging to other clusters. It is calculated using (4.1.1).
\[ DB_{pr} = \frac{1}{c} \sum_{i=1}^{c} \max_{j \neq i} \left\{ \frac{S_{pr}(X_i) + S_{pr}(X_j)}{d(X_i, X_j)} \right\} \quad (4.1.1) \]

\[ d(X_i, X_j) = \frac{\sum_{i,j} \|x_i - x_j\|}{\|c_i\|c_j\|} \quad (4.1.2) \]

\[ S_{pr}(X_i) = \begin{cases} \sum_{x_k \in (\overline{B}X_i - BX_i)} \|x_k - v_i\|^2 * t_{ik} / |BX_i - BX_i| \quad ; \quad if \quad |BX_i - BX_i| \neq 0 \land |BX_i| = 0 \\ \sum_{x_k \in BX_i} \|x_k - v_i\|^2 * t_{ik} / |BX_i| \quad ; \quad if \quad |BX_i - BX_i| = 0 \land |BX_i| \neq 0 \\ \sum_{x_k \in BX_i} \|x_k - v_i\|^2 * t_{ik} / |BX_i| \quad + \quad \sum_{x_k \in (\overline{B}X_i - BX_i)} \|x_k - v_i\|^2 * t_{ik} / |BX_i - BX_i| \quad ; \quad otherwise \end{cases} \quad (4.1.3) \]

where \(d(X_i, X_j)\) is the distance between the clusters \(X_i\) and \(X_j\) calculated using (4.1.2), \(S_{pr}(X_i)\) is the average distance between objects within the clusters \(X_i\) and \(X_j\) and \(t_{ik}\) is given by equation (4.1.3) and \(v_i\) is the centroid of the cluster. The parameters \(w_{low}\) and \(w_{up}\) can be adjusted based on the importance of the lower approximation and the boundary subject to the constraint \(w_{low} + w_{up} = 1\). Also, these parameters are not taken into consideration when either of the lower approximation or the boundary region is empty\(^\text{13}\).

### 4.2 D INDEX

D index is an indication of the distance between the clusters, i.e. how compact and separated the clusters are from each other. This factor should be high. D is given by using (4.2.1).

\[ D_{pr} = \min_j \left\{ \min_{i \neq j} \left\{ \frac{d(X_i, X_j)}{\max_k S_{pr}(X_k)} \right\} \right\} \quad (4.2.1) \]

where \(S_{pr}(U_k)\) is given by equation (4.1.3).

### 4.3 QUANTITATIVE ANALYSIS

Various qualitative measures are used to evaluate the performance of the rough fuzzy algorithm\(^\text{15}\). We present below the definitions of these indices. Before presenting these indices we introduce some notations.

Let \(\underline{A}(X_i)\) and \(\overline{A}(X_i)\) be the lower and upper approximations of cluster \(X_i\), and \(B(X_i) = \overline{A}(X_i) - \underline{A}(X_i)\) denote the boundary region of cluster \(X_i\). The parameters \(\omega\) and \(\overline{\omega}\) represent \(w_{low}\) and \(w_{up}\). It is the weighting that correspond to the relative importance of lower and boundary region. Also, \(\overline{\omega} = 1 - \omega\).

**Definition 4.3.1 (α index)**

It is given by the expression

This article can be downloaded from www_ijpbs_net
\[ \alpha = \frac{1}{c} \sum_{i=1}^{c} \frac{\omega A_i}{\omega A_i + \omega B_i} \], where
\[ A_i = \sum_{x_j \in A(X_i)} x_j = |A(X_i)| \quad \text{and} \quad B_i = \sum_{x_j \in B(X_i)} x_j \]

The \( \alpha \) index represents the average accuracy of \( c \) number of clusters. It is the average of the ratio of the number of objects in lower approximation to that in upper approximation of each cluster. In effect, it captures the average degree of completeness of knowledge about all clusters. A good clustering procedure should make all objects as similar to their centroids as possible. The index increases with increase in similarity within a cluster. Therefore, for a given data set and \( c \) value, the higher the similarity value of objects within the clusters gives a higher index value. The value of \( \alpha \) increases with the value of \( c \). In an extreme case when the number of clusters is maximum, i.e., \( c = n \), the total number of objects in the data set, the value of \( \alpha = 1 \), when \( A(X_i) = A(X_i) \forall i \), that is, all the clusters \( \{X_i\} \) are exact or definable. Whereas if \( A(X_i) = B(X_i) \, \forall i \), the value of \( \alpha = 0 \). Thus, \( 0 \leq \alpha \leq 1 \).

**Definition 4.3.2: (\( \rho \) index)**
It represents the average roughness of \( c \) number of clusters and is obtained by subtracting the average accuracy \( \alpha \) from 1.

\[ \rho = 1 - \alpha = 1 - \frac{1}{c} \sum_{i=1}^{c} \frac{\omega A_i}{\omega A_i + \omega B_i} \]

**Definition 4.3.3: (\( \alpha^* \) index)**
The measure represents the accuracy of approximation

\[ \alpha^* = -\frac{\sum_{i=1}^{c} \omega A_i}{\sum_{i=1}^{c} \{\omega A_i + \omega B_i\}} \]

**Definition 4.3.4: (\( \gamma \) index)**
It represents the approximation of the clustering algorithm obtained by finding the ratio of the total number of objects in lower approximation to that of cardinality of objects in the universe.

\[ \gamma = \frac{\sum_{i=1}^{c} |A(X_i)|}{|U|} \]

### 4.5 RESULTS AND DISCUSSION

For testing purposes we use eight data sets from UCI repository with varied dimensions. The data sets and their dimensions are given in table 4.5.1. The datasets have been chosen from a wide domain such as medicine, astrology, biotechnology, organisms etc. They have been taken from a wide domain so that the experimentation may be done across data from different fields and to check whether our results hold true across domains. Attributes of a dataset refers to the dimensions of the dataset. For example, for LIVER data set, it has 5 attributes based on which class labels are recorded for the data set. Number of instances refers to the total number of entries in the dataset.
ADHD data set is about Attention Deficit Hyperactive Disorder, collected by us under the supervision of a pediatric doctor from the schools in our place. The PRCM when executed on ADHD data set, has generated a good cluster whose performance is tabulated in Table 4.5.2. The table shows the comparison of PRCM and RCM clusters formation measured various indices explained in section 4. From the table, it is clear that PRCM shows superiority over RCM. The DB value is minimized to 0.6898 from 0.744. Similarly the D value is maximized from 1.52 to 1.59. The other indices like $\alpha$, $\rho$, $\alpha^*$ and $\gamma$ shows minor difference between the two algorithm and it is clear from the figure 4.5.1.

**Table 4.5.2**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Instances</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>120</td>
<td>11</td>
</tr>
<tr>
<td>Liver</td>
<td>583</td>
<td>5</td>
</tr>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
</tr>
<tr>
<td>Cancer</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>Libra Movements</td>
<td>360</td>
<td>91</td>
</tr>
<tr>
<td>E coli</td>
<td>336</td>
<td>7</td>
</tr>
<tr>
<td>Abalone</td>
<td>4177</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 4.5.2 illustrates the comparison of PRCM and RCM on ADHD data set.

![Comparison of PRCM and RCM on ADHD data set](image)

**Figure 4.5.1**

Comparison of PRCM and RCM on ADHD data set

Further to study the superiority of the proposed algorithm, it is tested on seven real data set considered from UCI repository. Clusters formed using Rough C-means (RCM) and Possibilistic Rough C-Means (PRCM) are compared. The cluster formation procedure is iterated until there is no more change in the cluster objects. The performance of clustering...
by both the algorithms are compared against various measures describe in section 4. The results are summarized in tables 4.5.3, 4.5.4 and 4.5.5. The algorithm uses some parameters like $w_{low}$, $w_{up}$, and $\theta$. These parameters are to be initialized at the beginning of the algorithm. These parameters play a major role in cluster formation where $w_{low}$ and $w_{up}$ values represents the importance or weightage given to lower and boundary region respectively. The value of $\theta$ determines the degree of similarity between the objects falling in the same cluster. It determines the object to be in lower or boundary region. The parameter values should be fixed at most care as they highly influence the cluster formation. These parameters are fixed based on literature and by trial. By literature $w_{up} = 1 - w_{low}$, $w_{low}$ should be between 0.5 and 1 and $\theta$ lying between 0 and 0.5\textsuperscript{13}. For our experiments we have taken $w_{low}$ as 0.9, $w_{up}$ as 0.1 and $\theta$ as 0.3.

### Table 4.5.3

**Comparison of PRCM and RCM based on number of objects and iterations**

<table>
<thead>
<tr>
<th>Data Set</th>
<th># LOWER APPROX</th>
<th># BOUNDARY</th>
<th># Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCM</td>
<td>PRCM</td>
<td>RCM</td>
</tr>
<tr>
<td>Liver</td>
<td>304</td>
<td>312</td>
<td>556</td>
</tr>
<tr>
<td>Iris</td>
<td>139</td>
<td>135</td>
<td>20</td>
</tr>
<tr>
<td>Wine</td>
<td>170</td>
<td>159</td>
<td>14</td>
</tr>
<tr>
<td>Cancer</td>
<td>0</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>Libra Movements</td>
<td>0</td>
<td>360</td>
<td>720</td>
</tr>
<tr>
<td>E coli</td>
<td>290</td>
<td>279</td>
<td>92</td>
</tr>
<tr>
<td>Abalone</td>
<td>3743</td>
<td>3731</td>
<td>866</td>
</tr>
</tbody>
</table>

From the table 4.5.3, one can understand that PRCM takes more number of iteration for convergence. On some cases PRCM has more objects in lower approximation than RCM and vice versa. The DB and D values are on seven data set are summarized in table 4.5.4. From this table it is evident that PRCM shows superior performance than RCM by minimizing DB value, at the same time maximizes D value.

### Table 4.5.4

**Comparison of the two algorithms based DB index and D index**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DB Index</th>
<th>D Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCM</td>
<td>PRCM</td>
</tr>
<tr>
<td>Liver</td>
<td>0.413818</td>
<td>0.232295</td>
</tr>
<tr>
<td>Iris</td>
<td>0.479248</td>
<td>0.082232</td>
</tr>
<tr>
<td>Wine</td>
<td>0.621923</td>
<td>0.448224</td>
</tr>
<tr>
<td>Cancer</td>
<td>11.82712</td>
<td>2.413243</td>
</tr>
<tr>
<td>Libra Movements</td>
<td>14.6831</td>
<td>2.199986</td>
</tr>
<tr>
<td>E coli</td>
<td>0.35195</td>
<td>0.197267</td>
</tr>
<tr>
<td>Abalone</td>
<td>0.177104</td>
<td>0.118933</td>
</tr>
</tbody>
</table>

### Table 4.5.5

**Comparison of PRCM and RCM based on $\alpha$, $\rho$, $\alpha^*$ and $\gamma$**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Alpha Index ($\alpha$)</th>
<th>P Index ($\rho$)</th>
<th>Alpha* Index ($\alpha^*$)</th>
<th>Gamma Index ($\gamma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCM</td>
<td>PRCM</td>
<td>RCM</td>
<td>PRCM</td>
</tr>
<tr>
<td>Liver</td>
<td>0.752823</td>
<td>0.791332</td>
<td>0.247177</td>
<td>0.208668</td>
</tr>
<tr>
<td>Iris</td>
<td>0.765172</td>
<td>0.936256</td>
<td>0.234828</td>
<td>0.082232</td>
</tr>
<tr>
<td>Wine</td>
<td>0.987987</td>
<td>0.967337</td>
<td>0.012013</td>
<td>0.032663</td>
</tr>
<tr>
<td>Cancer</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.28E-08</td>
</tr>
<tr>
<td>Libra Movements</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.87E-08</td>
</tr>
<tr>
<td>E coli</td>
<td>0.952812</td>
<td>0.944534</td>
<td>0.047188</td>
<td>0.053466</td>
</tr>
<tr>
<td>Abalone</td>
<td>0.976125</td>
<td>0.962947</td>
<td>0.023875</td>
<td>0.017053</td>
</tr>
</tbody>
</table>
On studying table 4.5.4, we note that while comparing the DB indices of both the algorithms, the values obtained for the data sets when the Rough C-means algorithm is used is fairly higher. For the Libra movements and cancer data sets, the DB index values are considerably higher than its Possibilistic rough C-means counterpart. All of the DB indices for the PRCM algorithm are comfortably lesser than the RCM algorithm, indicating that PRCM forms clusters having lower within-cluster distance. Obviously, PRCM performed better. When the D indices of table 4.5.4 are compared, the D indices for the Rough C-means algorithm are considerably lower than that of the Possibilistic Rough C-means hence again reassuring that the Possibilistic Rough C-means algorithm has fared better. Although the number of iterations required to cluster are higher in for the hybrid algorithm but, when this attribute is compared to the performance metric of the other parameters, this factor can be ignored.

![Figure 4.5.2](image1)

**Figure 4.5.2**
*Comparison of DB index*

![Figure 4.5.3](image2)

**Figure 4.5.3**
*Comparison of D index*

From Table 4.5.5, we notice that the $\alpha$ and $\alpha^*$ indices for the PRCM table is higher for some and lower for some values as compared to RCM table indicating that PRCM does not form quantitatively good but qualitatively good clusters. Since $\rho$ Index is again a quantitative measurement of clusters, it is higher for some and lower for some values when PRCM and RCM results are compared. Again, Gamma Index is a measure of the number of objects in the lower approximation to the cardinality of the universe. Since number of objects in lower approximation for PRCM is lower in some and higher in some when compared to RCM, it gives mixed results. Furthermore, we varied the values of $w_{\text{low}}$ and $w_{\text{up}}$ for both the algorithms and observed their respective performances for DB and D indices. The graph is generated for both the algorithms on these indices with $w_{\text{low}}$ values as 0.9, 0.8 and 0.5 for various data sets given in table 4.5.1. The figures 4.5.4 show the comparison of DB Index and the figure 4.5.5 shows the comparison on D Index.
On careful examination of figure 5.5.4, we conclude that DB of PRCM is lower than RCM irrespective of $w_{low}$ values. Similarly, we compare D values of RCM and PRCM for each dataset in figure 4.5.5. Here again, the value of D is higher for PRCM for all $w_{low}$ values. It should be noted that the value of $w_{up}$ changes with respect to $w_{low}$ and need not be considered separately.

From the figures 4.5.4 and 4.5.5, it is very clear that the performance of PRCM is superior to RCM in both DB and D index. The performance of PRCM is better when $w_{low}$ and $w_{up}$ values are 0.9 and 0.1.

4.6 CONCLUSION

In this part of work, we proposed a clustering algorithm called Possibilistic Rough C Means (PRCM) that is capable of forming good clusters even in presence of noisy data. The typicality or possibility value resolves the dilemma in placing the noisy data to the cluster. After analyzing the results and comparisons, we conclude that our proposed PRCM algorithm forms more cohesive clusters. DB index for all the datasets is lower than that of Rough C-Means. On the other hand, PRCM also separates the clusters from each other more efficiently than the
conventional RCM algorithm. D Index for PRCM is higher than RCM for all the data sets, indicating that the quality of clusters formed by PRCM is better than RCM. Also, the performance of the algorithm remains same even for varied values of $w_{\text{low}}$ and $w_{\text{up}}$.

REFERENCES