Comparative Study of Clustering Techniques in Iris Data Sets

V. Leela, K. Sakthi priya and R. Manikandan

1Department of Computing, SASTRA University, Thanjavur-613401, India
2Department of ICT, SASTRA University, Thanjavur-613401, India

Abstract: Clustering is one among the prominent area in the research field of data mining. Data clustering plays the major research at pattern recognition, information retrieval and bioinformatics. Clustering process is an unsupervised learning techniques where it generates a group of object based on their similarity in such a way that the objects belonging to other groups are similar and those belonging to other are dissimilar. This paper analysis the four different data types clustering techniques like K-Means, Fuzzy c mean, Mountain clustering and Subtractive clustering in Iris flower data set. The accuracy, run time, time complexity are compared among them and then newly improved Y-means algorithm are proposed in order to improve the obtained clustering result using Matlab tool. The result shows that improved Y-means algorithm yields better result when compared to other clustering techniques with less computation time.

Key words: K-means • Y-means • Fuzzy c-means • Mountain clustering

INTRODUCTION

Data clustering is an active approach for exploring the common metric similarities within the data and making them as groups. Every Clustering techniques, nomimates the clustered data set into numerous groups such that the common features among the group is maximum when compared to other groups. The process of data clustering is simple with few sets of data but when its a large amount of data, it has to be categorized into small number of groups to analyze the performance. The data sets of higher complexity with inherent properties and representation in 2D or 3D needs the help of soft computing and they are solved through the process known as “Data clustering methods”. The widely approached process for all clustering the data sets is to estimate the clusters Centre that will signify each cluster. The Figure 1 represents the data clustering process involved with raw data sets.

In the data mining concept cluster Centre plays a major role to group the data. In the group of data, Cluster Centre represents the input vector representing to which group this cluster belongs by estimating the similarity distance and determining which data cluster is nearest. Data clustering algorithms are widely employed to represent and analyzing the data, but also used in the model construction and compression techniques. In our paper, we represented five off-line clustering techniques by using iris data sets K-means, Improved Y-means, Fuzzy C-means Clustering, Subtractive Clustering and Mountain Clustering. These clustering techniques are implemented and analyzed using the iris flower data set which includes three species provided by Fisher's Iris data set.

Existing Data Clustering Techniques and Drawbacks

K-means: K-Means is one of the simplest methods among all partitioning based data clustering methods. It segments a iris flower data set into k clusters, where k is the number of user desired clusters and defined by the user. Every cluster are assigned with centroids.[1] All the data sets are set in a cluster having centroid nearest to that data object. In each iteration centroids change their location step by step. This process is continued until centroid are stable at place.
\[ u_{ij} = \begin{cases} \frac{1}{|g_i|} \sum_{x_k \in g_i} x_j \end{cases} \quad \text{(2)} \]

Strengths:
- Relatively efficient scalable process for huge sum of data sets
- Easy to understand and implement.

Weaknesses:
- Process begins only after the mean of a cluster is initialized.
- User defined clusters constant ie k.
- Not recommended for variant size or non-convex shaped clusters.
- Hard to handle data with noise and outliers.
- May break at local optimum.
- Result and total run time directly proportional to initial partition

Fuzzy c mean: Fuzzy C-Means clustering (FCM), is the data clustering algorithm in which each data set belongs to a cluster to a degree assigned by a membership. This techniques works iteratively until no further clustering is possible. Fcm undergoes following techniques for centroid calculations

Step 1: Assign the process randomly in matrix U within 1 by estimating in Equation (3)are satisfied.

\[ \sum_{i=1}^{k} u_{ij} = 1, \forall j = 1, \ldots, n. \quad \text{(3)} \]

Step 2: Estimate cluster points in centers assigned in fuzzy using Equation (4).

\[ c_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}} \quad \text{(4)} \]

Step 3: Calculate the cluster centroids function and check the values, if it is below the tolerance or threshold value then terminate the function.

\[ j(u, \text{cl}, \ldots, c) = \sum_{i=1}^{k} j_i = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}d_{ij}^2 \quad \text{(5)} \]

Step 4: Update new matrix function by \( u \) and estimate the cluster from step 2.

\[ u_{ij} = \frac{1}{\sum_{i=1}^{n} d_{ij}^m} \quad k=0,1,\ldots, n. \quad \text{(6)} \]

Strengths:
- Depends on initial clustering.
- The weighting exponent \( m \) is analyzed.

Weakness:
- Higher the weight exponent it lowers the accuracy.
- Increase the computation time for high complexity data sets.

Mountain Clustering: The effective clustering techniques like mountain clustering is a successive way known as mountain function to find cluster centers based on the density measures. This method find the approximate suitable cluster centers and can be used as a processed data for other process. It includes the following process for clustering the given set of data sets

Step 1: Initialize on the available data space with grid values and then the formed conjunction of the lines leads to the initial cluster centers.

Step 2: Constructs a matrices evaluating for data density measure.

\[ m(v) = \sum_{i=1}^{n} \exp \left( -\frac{|v - x|^2}{2\sigma} \right) \quad \text{(7)} \]
The constant s determines the height and smoothness of the resultant mountain function.

**Step 3:** Choose the centers of the cluster by eliminating its mountain function. The initial cluster center is fixed with the greatest density.

**Step 4:** Revisiting the mountain function as new by subtracting a scaled Gaussian function and it results for further points of cluster Centre with maximum value.

**Step 5:** This process repeats iteratively until the cluster centers are attained sufficiently.

**Strengths:**
- Estimates the approximate cluster centers k.
- Efficient way for clustering the density based data sets.

**Weakness:**
- Time complexity
- Not suitable for higher two or three dimensions problems.
- Computation is exponentially proportional to dimensions.

**Subtractive Clustering:** The drawback of the previous mountain clustering in the computation process where it increases due to the rise of dimensions. Subtractive clustering diminish the demerits by using data points as the member for cluster centers.

**Step 1:** Initialize the data sets as random value \{1,2..n\}.

**Step 2:** Calculate the all data points for its density functions by using

\[
p_i = \sum_{j=1}^{n} e^{\frac{[x_i-x_j]^2}{2s^2}}
\]  

**Step 3:** The calculated data function points is revisited and are valued till n with x function

\[
p_i = p_i - p_k e^{-\frac{x_i}{s}}, i = 1..n
\]  

**Step 4:** Let assume the points with high density and equal \(P\). If the value is less then assign x as new cluster center and back to Step 2, else go to Step 5.

**Step 5:** Publishes clustering results.

**Strengths:**
- Data functions are used as the members for cluster centers.
- Reduces the time complexity
- Results are fixed and has no random cluster value.

**Weakness:**
- Accuracy is less
- Cautious about choosing the neighbour radius.

**Proposed Methods:**

**Improved Y -means Clustering:** In order to overcome the drawbacks of K means algorithm, Y-means clustering established. The basic Y-mean clustering undergoes the following process. As the initial process it allows the self defined cluster value k, then it undergoes the following process as like K means in order to detect the centroids, then it checks for degeneracy, if its found it removes the empty cluster else undergoes splitting and merging action. The number of clusters is adjusted by removing empty clusters as well as merging and linking existing clusters. Due to the fact that the final number of clusters is independent of initial k. But it has the time complexity and iteration steps when the data sets are high. In order to overcome the drawbacks of basic Y-mean algorithm, Improved Y-Mean clustering techniques are employed in Iris data sets and process are followed.

**Step 1:** Initialize the size of the cluster Ci by \((n/k)=S\), where n is the no of data points and k is the cluster size.

**Step 2:** Create the array Bk and include all the Ak(data array) to Bk until the value is equal to Ci Cluster.

**Step 3:** Compute the arithmetic mean function M and compare the initial cluster value with the new data points. ie i<k<j

\[
M = \frac{1}{K} \sum_{i=1}^{K} \sum_{x_i \in C_i} D \_x \_C_i
\]  

**Step 4:** Compute the distance D of all A to M of Initial Clusters Cj If Akof Dk and M is less than or equal to other distances then remain in same cluster else undergoes merging and splitting techniques with centroids by the gaussian distribution is given by
Table 1: Traditional Y-mean algorithm

<table>
<thead>
<tr>
<th>Run</th>
<th>No of Clusters</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 2: Proposed Y-mean algorithm

<table>
<thead>
<tr>
<th>Run</th>
<th>No of Clusters</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>17</td>
</tr>
</tbody>
</table>

\[ f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \] (11)

**Step 5:** Recompute M and move to BK until there exit no change in clusters.

The experimental results between traditional Y-means and proposed Y-means algorithm are compared between iris data sets. The results shows that the iteration value and time complexity is reduced with the data sets. The initial clusters values are based on the explore and find mechanism. Initially two smallest elements are searched and those elements are then removed from the input array and moved to the new sub set of array list. In the proposed Y-mean clustering the searching mechanism for centroid fixed value is not included due to which it reduces the iteration and time complexity. Traditional Y-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs. Table 1 shows different results for same data set. Proposed Y-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Table 2 shows same number of iterations and cluster size.

**RESULTS**

The Clustering the functional data sets received particular attention from researchers in the last decade. The above section involves the discussion of various implementation of clustering techniques steps along with their merits and demerits. The Iris data, which is created by R.A. Fisher is a well-known dataset for classification of various data clustering techniques [4]. The iris dataset
includes 3 classes of 50 instances in each and each class refers to a type of Iris flower. The three classes are 1. Iris Setosa, 2. Iris Versicolor and 3. Iris Virginica

The Matlab is efficient tool used in order to cluster the given iris data sets in order the evaluate clustering techniques. Each algorithm has different techniques to cluster the iris data sets. The Figure 2,3,4,5,6,7 shows the results of K-means, improved Y-means, Fuzzy c means, Subtractive and Mountain clustering with their respective iteration number using Matlab R2010 tool. The red, green and blue colour indicates the various species of the iris flower. The Time complexity along with their accuracy level are evaluated to categorize the performance and results are listed in table 3 as comparison between various clustering techniques [5]. The graph has been drawn as Figure 8 in order to show the run time performance. The results obtained are able to conclude that improved Y mean is performed well when compared with other algorithm in the time complexity and accuracy.

### Table 3: Comparison between various data clustering techniques

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No of Clusters</th>
<th>Iteration</th>
<th>Run Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-MEANS</td>
<td>4</td>
<td>6</td>
<td>1.3</td>
<td>82.00%</td>
</tr>
<tr>
<td>Improved Y-MEANS</td>
<td>2</td>
<td>4</td>
<td>0.9</td>
<td>85.00%</td>
</tr>
<tr>
<td>FC MEANS</td>
<td>5</td>
<td>5</td>
<td>2.2</td>
<td>81.00%</td>
</tr>
<tr>
<td>MOUNTAIN</td>
<td>3</td>
<td>6</td>
<td>118.0</td>
<td>77.4%</td>
</tr>
<tr>
<td>SUBTRACTIVE</td>
<td>4</td>
<td>5</td>
<td>3.6</td>
<td>75.00%</td>
</tr>
</tbody>
</table>

CONCLUSION

Thus the most efficient five types of data clustering techniques have been analyzed using iris flower data sets. The estimation process are done for the accuracy of each algorithm and also the time complexity function are evaluated for performance measures.

The mountain clustering is not good recommended for high dimension because its proportional to its time complexity. K-means clustering overcomes the performance with other techniques by their faster performance even with large data sets [6]. The improved Y-means has overcome the drawbacks of K-mean and other techniques by self defined cluster variable k and reduces the time complexity and iteration times. Fuzzy c mean has higher time complexity due to the fuzzy calculation and has same accuracy like K-means and Y-means. When the clusters is non-determined, K-means and FCM clustering techniques cant be evaluated for the process, leaving the chance to other clustering. The Subtractive clustering yields better alternative for mountain clustering with the use of the data points instead of grid points and it shows improved accuracy than mountain clustering.
Future Work: In our paper work we have compared the four algorithm and used improved algorithm Y-mean to prove its accuracy among other types. Even though our comparison is not able to cover all the factors like slope over regression, cost function etc of these data clustering algorithms. As a future forward steps, we like to compare between these algorithms using different parameters other than used parameters. The normalized and unnormalized data yields better difference in the results, so it will be added in our future work.

REFERENCE


3. Khalid Hammouda, Prof. Fakhreddine Karray,”A Comparative Study of Data Clustering Techniques”, University of Waterloo, Ontario, Canada N2L 3G1

