PERFORMANCE ASSESSMENT OF IMPROVED FARTHEST FIRST CLUSTERING ALGORITHM ON SMARTPHONE SENSORS DATA

Dr. M. Jayakameswaraiah, Dr. K. Suresh Kumar Reddy, Dr. S. Ramakrishna

1. Assistant Professor, 2. Academic Consultant, 3. Professor
1. School of Computer Science and Applications, Reva University, Bangalore, Karnataka, India.
2. Department of Computer Science and Engineering, S.V.U College of Engineering, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.
3. Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.

Abstract: Research on effective methods to deal with ever larger data sets has been gaining importance in recent years. The goal of clustering is to organize a data collection into clusters, such that items within each cluster are more similar to each other than to items in other clusters. While supervised clustering assumes that some information is available concerning the membership of data items to predefined classes, unsupervised clustering does not require a priori knowledge of data contents. There are many applications of unsupervised clustering in computer vision, pattern recognition, information retrieval, data mining, etc. The objective of this research work is focused on the ethical cluster creation of smartphone sensors data and analyzed the performance of partition based algorithms. Dataset consist of all the information gathered during the network connection established with wristwatch and smartphone, which needs to be, analyzed the performance of the proposed improved farthest first clustering algorithm.

Index Terms: Clustering algorithms, Improved Farthest First algorithm, Density Based Cluster, Filtered Cluster, K-Means, Data Mining.

I. INTRODUCTION

Data Mining discovers unknown relationships in data; in fact it is part of a wider process called “knowledge discovery”. Knowledge discovery defines the phases which should be completed to ensure getting significant results through research. The objective of DM process is to obtain information out of a dataset and convert it into a comprehensible outline. Also, it includes the following: data reprocessing, data management, database aspects, visualization and complexity considerations, online updating, inference and model considerations, interestingness metrics. On the other hand, the actual data mining assignment is the semi-automatic or automatic exploration of huge quantities of information to extract patterns that are interesting and previously unknown. Clustering is one of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Clustering is the method of making group of objects into classes of similar objects.

II. LITERATURE REVIEW

For performing comparison analysis dataset is obtained from UCI data repository. These repositories are very helpful in data analysis. This data can directly process in data mining tools for predicting the results.

2.1 Clustering Techniques

Clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters. The clusters can be formed based upon various parameters depending upon the clustering method. Different Clustering methods have a database of n objects or data tuples.

- Partitioning methods
- Hierarchical based methods
- Density based method
- Grid-based methods
- Model-based methods
- Clustering high-dimensional data
- Constraint-based clustering
2.1.1 Filter Based Clustering

This category of clustering method is for filtering the information or any pattern which are essentially needed. The filtration is carried out based on the keywords that are supplied or some relevant information. Jiang-She Zhang et al proposed a clustering algorithm for the processing of the images. They are computationally stable and insensitive to initialization. They also produce consistent clusters. Thomas et al proposed a collaborative filtering which is a combination of the correlation and singular value decomposition (SVD) to improve accuracy [7,13].

2.1.2 Density Based Clustering

The density based clustering (DBC) groups the objects mainly based on the density of the objects that are reachable and connective. Li Tu et al proposed a framework called D-stream for clustering using the density based approach. Mitra et al suggested an nonparametric data reduction scheme. The procedure followed here is separating the dense area objects from dense area with the aid of an arbitrary object[1,3,15]. The density based clusters (DBC) are robust to noise but the datasets are problematic and requires high densely connected data.

2.1.3 Centroid Based clustering

K-Means is a centroid based clustering method. It partitions the dataset into various clusters based on the mean distance. It is one of the simplest forms of unsupervised algorithm. The main objective of this algorithm is to reduce the squared error. Tapas et al identified that the algorithm works faster as the separation between the clusters increases. This algorithm is applicable for the segmentation of images and data compression. Kanungo et al proposed that the K-means algorithm runs faster as the separation between the cluster increases [2,5,14].

III. IMPLEMENTATION OF IMPROVED FARTHEST FIRST ALGORITHM

First farthest first algorithm proposed by Hochbaum and Shmoys 1985 has same procedure as k-means, this also chooses centroids and assign the objects in cluster but with max distance and initial seeds are value which is at largest distance to the mean of values, here cluster task is different, at initial cluster we get link with extraordinary Session Count, like at cluster-0 more than in cluster-1, and so on. Working as labeled here, it also describes initial seeds and then on basis of “k” number of cluster which we need to know prior[3,15]. In farthest first it takes point Pi then it chooses next point P1 which is at maximum distance. Pjis centroid and p1, p2, …, p, are points or objects of dataset belongs to cluster from equation 1.

\[
\text{Score}(X_i) = \sum_{j=1}^{m} f(x_i) [D]
\]

Farthest first actually solves problem of k-Centre and it is very efficient for large set of data. In farthest first algorithm we are not finding mean for calculating centroid, it takes centroid arbitrary and distance of one centroid. For each \( X_m = [x_{m, 1}, x_{m, 2}, ..., x_{m, m}] \) in D that is labeled by \( m \) categorical attributes, we use \( f(x_{i,j})[D] \) to denote the frequency count of attribute value \( x_{i,j} \) in the dataset. Then, a scoring task is designed for evaluating each point, which is defined as:

1. Step-1: Farthest first traversal (D: data set, k: integer)
2. Step-2: Randomly select first center
3. Step-3: /select centers
4. Step-4: For (i = 2,…, k)
5. Step-5: For (each remaining point) calculate distance to the current center set;
6. Step-6: Select the point with maximum distance as new center
7. Step-7: Apply Euclidean Distance function on each cluster
8. Step-8: /assign remaining points
9. Step-9: for (each remaining point)
10. Step-10: Calculate the distance to each cluster center using Manhattan distance formula
11. Step-11: put it to the cluster with minimum distance
12. Step-12: repeat the steps until each cluster remains

3.1 Euclidean Distance

This Euclidean distance between the two documents \( d_i, d_j \) can be intended as

\[
\text{Euclidean Distance}(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (d_{ik} - d_{jk})^2}
\]

Where, \( n \) is the number of positions present in the vector space model. Euclidean distance gives the dissimilarity between the two documents. If the distance is lesser it indicates they are more similar else dissimilar.

3.2 Manhattan Distance

The Manhattan distance between the two documents \( d_i, d_j \) can be intended as

\[
\text{Manhattan Distance}(d_i, d_j) = \sum_{k=1}^{n} |d_{ik} - d_{jk}|
\]
3.3 Accuracy Measure
To determine the accuracy of the clusters created by clustering algorithms, F-measure is used. Every cluster created by clustering algorithms is considered as the result of a query and the documents in these clusters are treated as set of retrieved documents. The documents in each category are considered as set of appropriate documents. The documents in the cluster that are relevant to that cluster are set of relevant documents retrieved. Precision, Recall and F-measure are calculated for every cluster[12,16,17]. The complete accuracy of the clustering algorithm is the average of the accuracy of all the clusters. Precision, recall and F-measure are considered as follows:

- Precision = relevant documents retrieved / retrieved documents
- Recall = relevant documents retrieved / relevant documents
- F-Measure = (2 * Precision * Recall) / (Precision + Recall)

IV. RESULTS AND DISCUSSIONS

In this section, by smartphone sensors dataset a series of experiments are carried out to categorize the documents into predefined number of categories by using improved farthest first algorithm. The initial centroids for k-means algorithm are chosen randomly by using farthest neighbors. The accuracy of the clusters and proficiency of the algorithm is examined. In this work the tests are carried out with one of smartphone sensors dataset collected from UCI and KDD repositories. The smartphone_sens consists of 18354 Instances and 13 attributes that are timestamp, AccelerationX, AccelerationY, AccelerationZ, MagneticFieldX, MagneticFieldY, MagneticFieldZ, Z-AxisAngle(Azimuth), X-AxisAngle(Pitch), Y-AxisAngle(Roll), GyroX, GyroY, GyroZ.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Number of Clustered Instances</th>
<th>Percentage of Clustered Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>Cluster 0</td>
<td>10644</td>
<td>58%</td>
</tr>
<tr>
<td></td>
<td>Cluster 1</td>
<td>7710</td>
<td>42%</td>
</tr>
<tr>
<td>Density Based Clusterer</td>
<td>Cluster 0</td>
<td>10775</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>Cluster 1</td>
<td>7579</td>
<td>41%</td>
</tr>
<tr>
<td>Filtered Clusterer</td>
<td>Cluster 0</td>
<td>10644</td>
<td>58%</td>
</tr>
<tr>
<td></td>
<td>Cluster 1</td>
<td>7710</td>
<td>42%</td>
</tr>
<tr>
<td>Improved Farthest</td>
<td>Cluster 0</td>
<td>17999</td>
<td>98%</td>
</tr>
<tr>
<td>First Clusterer</td>
<td>Cluster 1</td>
<td>355</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 4.1: Performance comparison of clustering algorithms

Figure 4.1: Performance comparison of Clustering Algorithms on Cluster 0

Figure 4.2: Number of Clustered Instances on Cluster 0
V. CONCLUSION

Clustering has wide applications. It is often used as an individual data mining tool to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis. Clustering not only can act as an individual tool, but also can serve as a preprocessing step for other algorithms which would then operate on the detected clusters. In this experimental research we estimated the performance of the clustering algorithms namely K-Means, Density Based clusterer, Filtered clusterer and our proposed Improved Farthest First Clusterer are evaluated on smartphone sensor data which is taken from the UCI-Machine learning repository. In this research we applied various clustering algorithms on the smartphone sensors data set and compared the clustering accuracy performances using data mining tool. But it is observed that our proposed algorithm gives 98% of clustering accuracy on cluster 0. That means it contributes better results than other standard clusters in data mining system.

REFERENCES