International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 375-378
e-ISSN:2278-621X

MODIFICATION OF FAST CLUSTERING ALGORITHM TO CLUSTER THE DATASETS IN DATA MINING

Ashwin D'Souza[1], Jaison D'Souza[2], Karen Maria Buthello[3] and Vanitha T[4]

Abstract - Clustering is similar to as of classification in which data is grouped. A cluster classification is a collection of objects and which are similar between them and are dissimilar to the objects belonging to the clusters.Clustering analysis isthe main analytical methods in data mining. K-means isthe most popular and partition based on clustering algorithm. But it is computationally expensive and the quality of clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have beenproposed in the literature for improving a performance of the k-means clustering algorithm. In this research, the mostrepresentative algorithms K-Means and K-Medoids were examined and analysed based on their basic needs. Thebest algorithm in each category was found out based on how well they perform.

I.INTRODUCTION

Partitioning of objects into homogeneous clusters is anoperation for data mining, this operation is required for an individual classification and data aggregate and distribution of large heterogeneous data setsinto smaller homogeneous subsets that can be easily managed, separatelymodelled and analysed. Objects in the same cluster areidentical to each otherthan objects in different clusters according to therepresent criteria. Clusteranalysis toolsare based onk-means, K-Medoids,and othermethods have also been built into many other statistical analysissoftware packages.

II. EXISTING ALGORITHM

A. K-Means Algorithm

It is an unsupervised learning algorithm with different data analysis applications widelyused for data mining and machine learning purposes. The main goal is to classify data into groups of information. The group is consists of the separation of information examination of specific datasets into k different clusters which combinedevery data entries. Each data entry of the result is related with centroids. Within these sets, the distance of centroid is minimised.The process to achieve the result sets of classified data. It basically consists on several iterations of a particular process, designed to getan optimal minimum solution for all data points.

Process:
First, we need to establish a function for what we want to first minimise, means the distance betweena data point and correspondent to the centroid.

[1] *Aloysius Institute of Management and Information Technology(AIMIT), St Aloysius College(AUTONOMOUS), Mangalore, Karnataka, India*
[2] *Aloysius Institute of Management and Information Technology(AIMIT), St Aloysius College(AUTONOMOUS), Mangalore, Karnataka, India*
[3] *Aloysius Institute of Management and Information Technology(AIMIT), St Aloysius College(AUTONOMOUS), Mangalore, Karnataka, India*
[4] *Aloysius Institute of Management and Information Technology(AIMIT), St Aloysius College(AUTONOMOUS), Mangalore, Karnataka, India*

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \|x_j - c_i\|^2$$

To achieve the result we are splitting the process into several steps. For this, we need a large set of data entries and k.

1. Randomly choose k points as partition centres.
2. To store the information calculates the data point on the set and centres distance.
3. Assign each point to the nearest cluster centre. Theminimum distance is calculated for each point.
4. Update the cluster centre positions by using the following formula

$$c_i = \frac{1}{|k_i|} \sum_{x_{j \in k}}^{k} x_j$$

5. If the cluster centres changes then repeat the process from step 2. Otherwise, you have successfully computed the k-means clustering algorithm and gotpartition member and centroids.

The achieved result is minimum configuration for the selected start point. It is possible that the output isn't an optimal minimum of the selected set of data, but instead a local minimum function. Run the process more than once to get the proper solution.

Advantages
This clustering methodology has some benefits comparing to others. The most important ones are:

i. Lots of Applications- It has several live world implementations on many different subjects.
ii. Fast – Achieves the result in a fast way.
iii. Simple and reliable- It solving the problem for a large set of information.
iv. Efficient– It givesa good solution with relatively low computing complication for clustering problem.
v. Good Solution – Best result set will be provided.
Disadvantages
i. No Absolute Data – It can't be used on data entries because it's notreplicate a mean function.
ii. Result Set – The result set is not good.
iii. Initialization method – The results will change depending upon the initialization action.

B. K-Medoid Algorithm

Method:
1.Initialize i by random selection from n data points of data matrix A
2.Calculate the distance between data points n and medoidi
a. Element by element binary operation:
i.v+v(T) Centred dot and return Y
ii. Compute 2*(A(T)*A) where A is data matrix, A(T) is transpose of data matrix A and return Z

iii. Compute Y-Z and return D
b. Select minimum from D obtaining a random sample from D(n,i) where n is a number of columns and
i is medoid.
3.For each data point o, swap with medoidi, and compute total cost B
4.Compute minimum cost c from total cost B
5.Add c to Final Matrix
Advantages
i. Flexible - We can use K-Medoids with *any* similarity measure, medoids are only used with distances are consistent with the *mean*.
ii. Robustness of medoid–It's used by K-Medoids is roughly comparable to the *median.* It is a
iii. More *robust* estimate of a representative point than mean as used k-means.

III. RESULT

```
Enter no of elements in cluster
9
Enter elements in cluster
2 4 10 12 3 20 30 11 25
Enter value of m1 and m2
2 3
Cluster 1   2
Cluster 2   4 10 12 3 20 30 11 25
average of cluster1 2.0
average of cluster2 14.375
Cluster 1   2 4 3
Cluster 2   10 12 20 30 11 25 11 25
average of cluster1 3.0
average of cluster2 18.0
Cluster 1   2 4 10 3
Cluster 2   12 20 30 11 25 25 11 25
average of cluster1 4.75
average of cluster2 19.875
Cluster 1   2 4 10 12 3 11
Cluster 2   20 30 25 11 25 25 11 25
average of cluster1 7.0
average of cluster2 21.5
Cluster 1   2 4 10 12 3 11
Cluster 2   20 30 25 11 25 25 11 25
average of cluster1 7.0
average of cluster2 21.5
```

Figure 1. K-Means Result

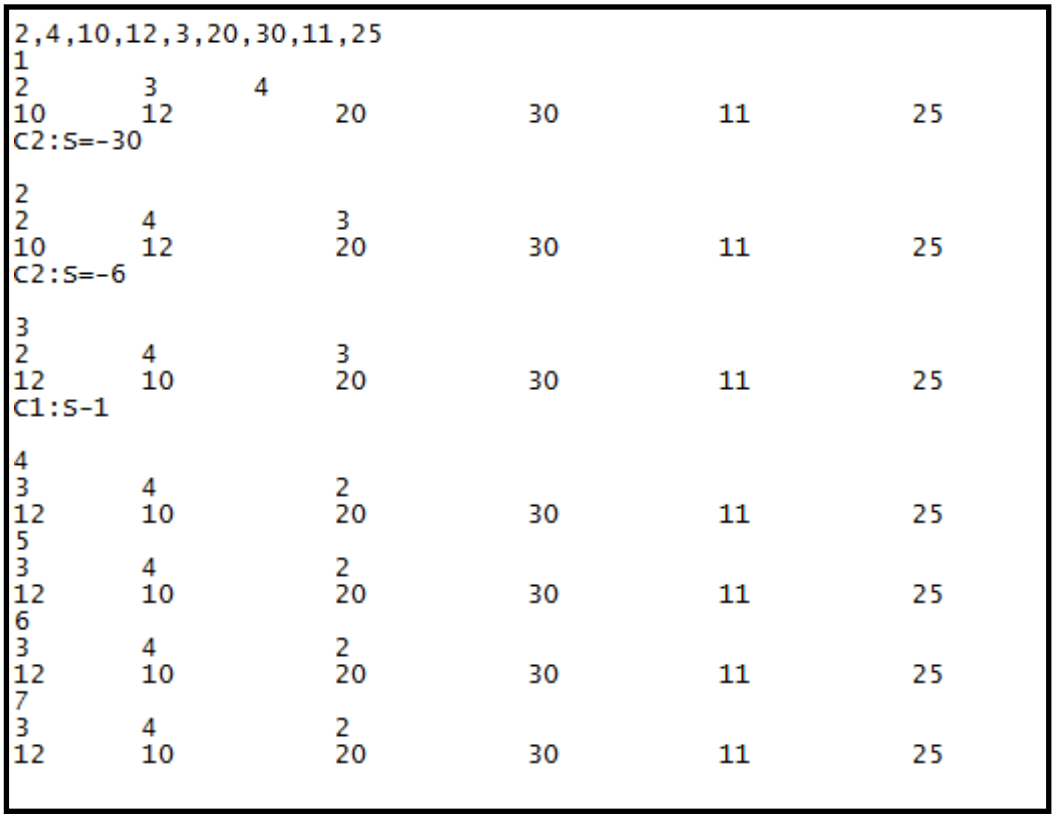

Figure 2. K-Medoid Result

IV. CONCLUSION

Since measuring the similarity between data objects is simple than mapping data objects to data points in feature space, these pairwise similarity clustering algorithms can greatly reduce the difficulty in developing clustering based patternrecognition applications. The advantage of the K-means algorithm is favourable execution time and It'sfault is thatthe user has to know how many clusters are searched forin advance. It is observed thatK-means algorithm is efficient for smaller data sets and K-Medoids to seem to perform better for large datasets.

REFERENCES

[1]"Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra, ACT 2601, Australia(1988)

[2] Sridhar Ramaswamy, Rajeev Rastogi, Kyuseok Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets "

[3] Rose, K., Gurewitz, E. and Fox, G. (1990) "A *Deterministic Annealing Approach to Clustering*, Pattern Recognition Letters"

[4] Zhexue Huang,"A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining"