ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.7

Comparison of Partitioning Algorithms for Categorical Data in Cluster

Rakesh Verma[1], Dr.D.M Puntambekar[2]
Associate Professor[1], Director[2]
School of Computers, IPSA, India

Abstract:
Data mining is the process of extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). There are different types of clusters. Partitioning method, Hierarchical method, Grid-based method, Density-based method, Model based method. This paper presents a study of various partitioning techniques of clustering algorithms and their relative study by reflecting their advantages individually. In this study some algorithms are presented which can be used according to one's requirement. In this paper, various well known partitioning based methods – k-means, k-medoids, Clara and Clarans are compared. The study is given through the Complexity, Efficiency, Implementation, Sensitive to outliers and Optimization.

General Terms: Data Mining, Partitioning Methods

Keywords: Clustering, k-means, k-medoids, Clara, Clarans

1. INTRODUCTION

The purpose of the data mining technique is to mine information from a large data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Cluster is an ordered list of data which have the familiar characteristics.
Among all these methods, this paper is aimed to comparison of different partitioning based clustering methods which are k-means, k- medoids, Clara and clarans. Four methods are discussed along with their algorithms, strength and limitations

2. PARTITIONING TECHNIQUES

Partitioning techniques divides the object in multiple partitions where single partition describes cluster. The objects with in single clusters are of similar characteristics where the objects of different cluster have dissimilar characteristics in terms of dataset attributes. A distance measure is one of the feature space used to identify similarity or dissimilarity of patterns between data objects. K-mean, K-medoid, CLARA and CLARANS are partitioning algorithm .

2.1 K-MEAN

K-mean algorithm is one of the centroid based technique. It takes input parameter k and partition a set of n object from k clusters. The similarity between clusters is measured in regards to the mean value of the object. The random selection of k object is first step of algorithm which represents cluster mean or center. By comparing most similarity other objects are assigning to the cluster**.**

Algorithm:
The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:
- K:the number of clusters
- D:a data set containing n object

Output:
- A set of k clusters

Method:

(a) Arbitrarily choose k objects from D as the initial cluster centers.

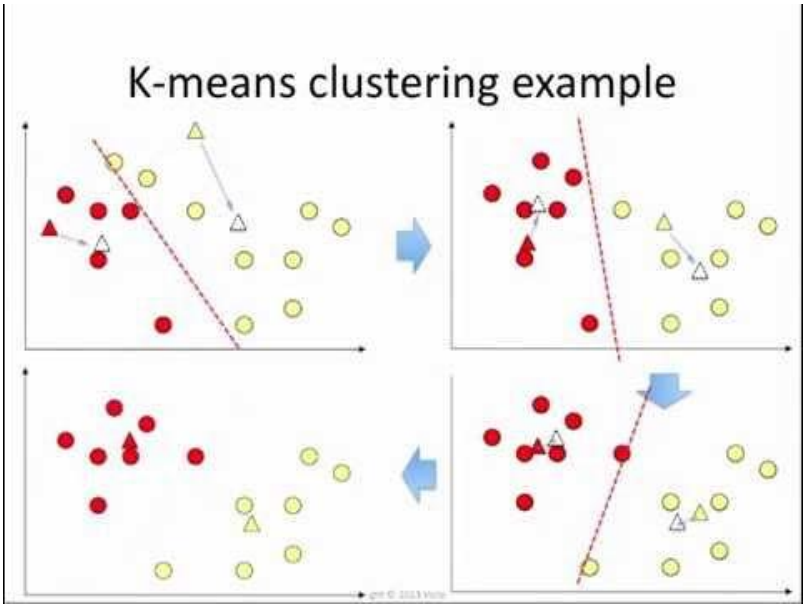
(b) Repeat

(c) Reassign each object to the cluster to which the object is the most similar,

Based on the mean value of the objects in the cluster;

(d) update the cluster means ,i.e., calculate the mean value of the objects for each
Cluster;

(e) Until no change;



2.2 K-MEDOID

The k-means method is based on the centroid techniques to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data.

To overcome the problem we used K-medoids method which is based on representative object techniques. Medoid is replaced with centroid to represent the cluster. Medoid is the most centrally located data object in a cluster.

Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. This process is continued until no any medoid move. As a result, k clusters are found representing a set of n data objects. An algorithm for this method is given below.

Algorithm: PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:

- K: the number of clusters,
- D: a data set containing n objects.

Outputs:

- A set of k clusters.

Method:

(a) Arbitrarily choose k objects in D as the initial representative objects or seeds;

(b) Repeat

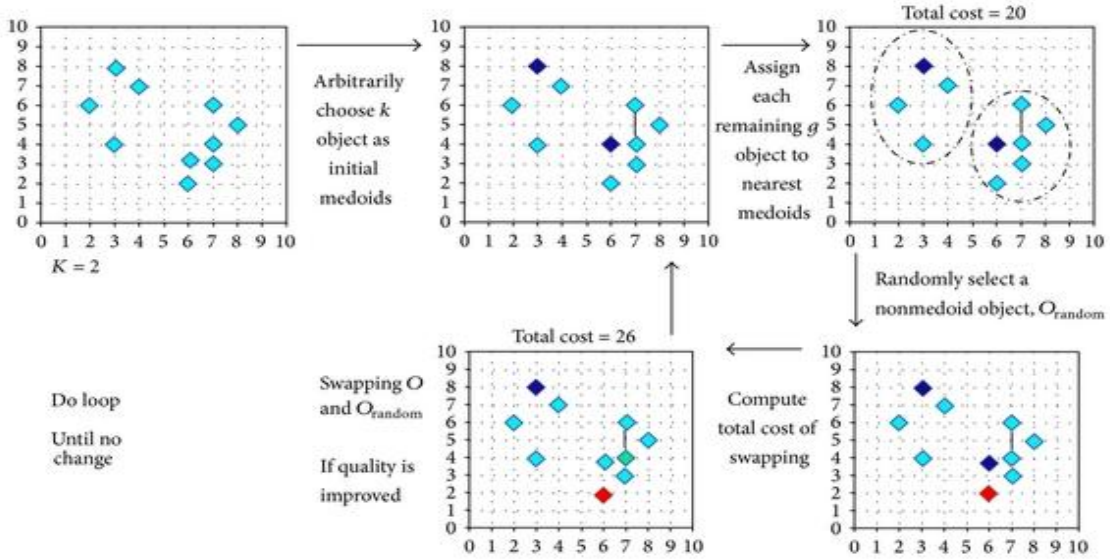
(c) Assign each remaining object to the cluster with the nearest representative object;

(d) Randomly select a non-representative object, Orandom.

(e) Compute the total cost of swapping representative object, Oj with Orandom;

(f) If S<0 then swap Oj with Orandom to form the new set of k representative object;

(g) Until no change;



2.3 CLARA (Clustering for Large Application):-

CLARA means clustering large applications and has been developed by Kaufman and Rousseuw in 1990. This partitioning algorithm has come into effect to solve the problem of Partition Around Medoids (PAM).CLARA extends their K-Medoids approach for large number of object. This technique selects arbitrarily the data using PAM. According to Raymond T. Ng and Jiawei Han the following steps are performed in case of CLARA as given by the authors

1) Draw a sample of 40+2k objects randomly from the entire data set, and call Algorithm PAM to find k medoid of the sample.

2) For each of the object determine specific Kmedoid which is similar to the given object (Oj).

3) Calculate the average dissimilarity of the clustering thus obtained. If the value thus obtained is less than the present minimum we can use it and retained the K-Medoid found in the second step as best of medoid.

4) We can repeat the steps for‘ next iteration‘

2.4 CLARANS

K-medoid algorithm doesn’t work effectively on large dataset. To overcome the limitation of K-medoid algorithm clarans algorithm is introduced[4]. Clarans (Clustering Large Application Based upon Randomized Search) is partitioning method used for large database. Combination of Sampling technique and PAM is used in CLARANS. In CLARANS we draw random sample of neighbours in each step of search dynamically. CLARANS doesn’t guaranteed search to localized area. The minimum distance between Neighbour nodes increase efficiency of the algorithm. Computation complexity of this algorithm is O (n^2).

3. COMPARISION

This table depicts the comparison between k-mean, K-medoid and clarans based on different parameter:

Parameters	k-means	k-medoids	Clara	Clarans
Complexity	O (i k n)	O (i k $(n-k)^2$)	O(ks^2+k(n-k))	O (n^2)
Efficiency	Comparatively more	Comparatively less	Comparatively more	Comparatively more
Implementation	Easy	Complicated	Complicated	Complicated
Sensitive to Outliers?	Yes	No	No	No
Advance specification of No. of clusters ‘k’	Required	Required	Required	Required
Does initial partition affects result and Runtime?	yes	yes	Yes	Yes
Optimized for	Separated clusters t	Separated clusters Small dataset	Separated clusters Large dataset	Separated clusters Large dataset

4. LIMITATION OF EXISTING ALGORITHM

K-Mean

- It is sensible to initial configuration
- Unsuccessful initialization gives empty clusters

K-Medoid

- It is not so much efficient for large dataset.
- It is more costly; complexity is O (i k (n-k)2), where i is the total number of iterations, is the total number of clusters, and n is the total number of objects.
- It has to specify k, the total number of clusters in advance.
- Result and total run time depends upon initial partition.

Clara.

- CLARA Algorithm deals with larger data sets than PAM (Partition Around Mediods).
- The efficient performance of CLARA depends upon the size of dataset.

Clarans

- It doesn’t guarantee to give search to a localized area.

- It uses randomize samples for neighbors.
- It is not so much efficient for large dataset.

4. CONCLUSION

The objects with in single clusters are of similar characteristics where the objects of different cluster have dissimilar characteristics in terms of dataset attributes Several methods have been studied to discover cluster and all these methodologies have been demonstrated in this paper. Partitioning based clustering methods are suitable for categorical based cluster which have small to medium sized dataset. However, to develop the understanding of parameters like Complexity ,Efficiency, Implementation, Sensitive to outliers and Optimization and effects of each parameter of every system needs a very detailed experimentation. The sole purpose of this paper is to help the researchers to select the one according to their need. Future research will focus on using these algorithms together or modify, such that the strengths, performance and efficiency of these techniques can be improved.

6. REFERENCES

[1] Saurabh Shah & Manmohan Singh "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm", International Conference on Communication Systems and Network Technologies, 2012.

[2] T. Velmurugan,and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach" An experimental approach Information. Technology, Journal, Vol, 10, No .3 , pp478-484,2011.

[3] Shalini S Singh & N C Chauhan ,"K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, 2011.

[4] "Data Mining Concept and Techniques", 2nd Edition, Jiawei Han, By Han Kamber.

[5] Jiawei Han and Micheline Kamber, "Data MiningTechniques", Morgan Kaufmann Publishers, 2000.

[6] Abhishek Patel,"New Approach for K-mean and K-medoids algorithm", International Journal of Computer Applications Technology and Research,2013.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn" Data clustering: a review". ACM Computing Surveys, Vol .31No 3,pp.264–323, 1999.