Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms

Gopi Gandhi
Student (ME CSE IV)
Parul Institute of Engineering and Technology

Rohit Srivastava
Assistant Professor
Parul Institute of Engineering and Technology

ABSTRACT
Clustering plays a vital role in research area in the field of data mining. Clustering is a process of partitioning a set of data in a meaningful sub classes called clusters. It helps users to understand the natural grouping or cluster from the data set. It is unsupervised classification that means it has no predefined classes. This paper presents a study of various partitioning techniques of clustering algorithms and their relative study by reflecting their advantages individually. Applications of cluster analysis are Economic Science, Document classification, Pattern Recognition, Image Processing, text mining. No single algorithm is efficient enough to crack problems from different fields. Hence, 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 and Clarans – are compared. The study given here explores the behaviour of these three methods.

General Terms
Data Mining, Unsupervised learning, Partitioning Methods

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

1. INTRODUCTION
Data Mining is a process of identifying valid, useful, novel, understandable pattern in the data. Data Mining is concern with solving problem by analyzing existing data. Clustering is a method of data explorations, a technique of finding patterns in the data that of our interest. Clustering is a form of unsupervised learning that means we don't know in advance how data should be group together [1]. Various Techniques for clustering are as follows [2]:

1. Partitioning Method
2. Hierarchical Method
3. Grid- based Method
4. Density-based Method
5. Model-based Method

Among all these methods, this paper is aimed to explore partitioning based clustering methods which are k-means, k medoids and clarans. These 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[7]. K-mean, K-medoid and CLARANAs are partitioning algorithm [3].

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 [4]: 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;

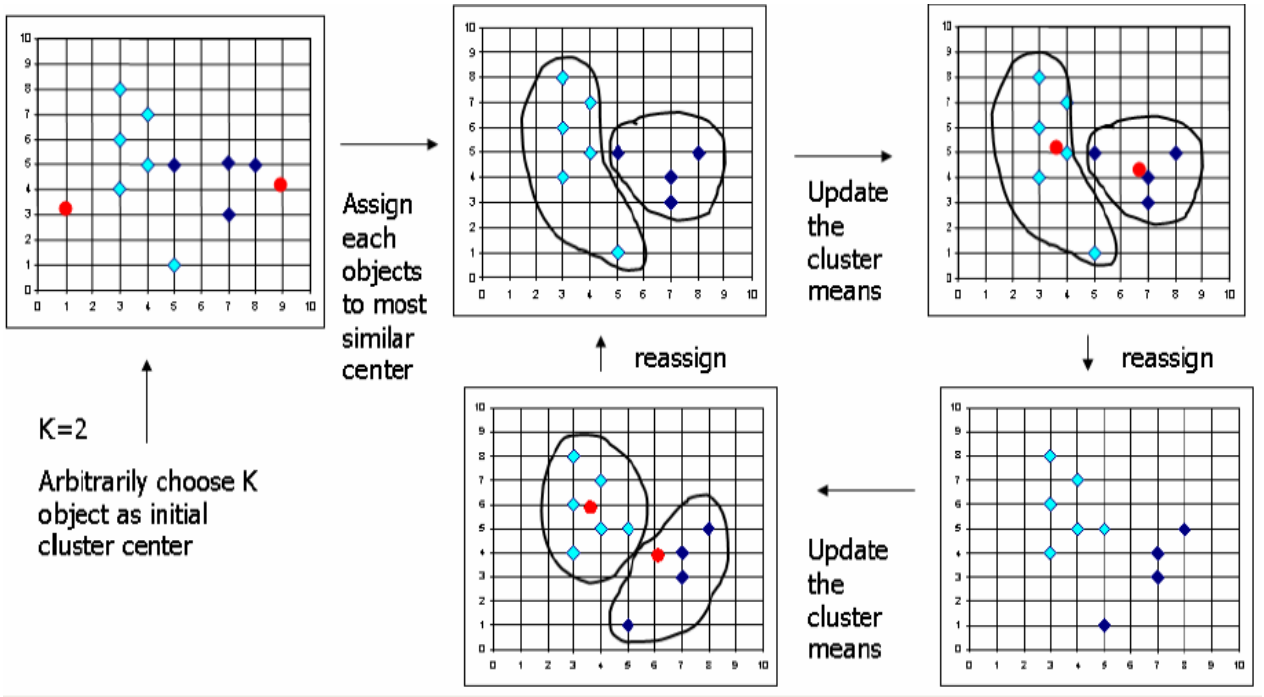

Figure 1: Working of K-mean Algorithm [4].

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[6].

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 [3]. An algorithm for this method is given below.

Algorithm [3]: 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;

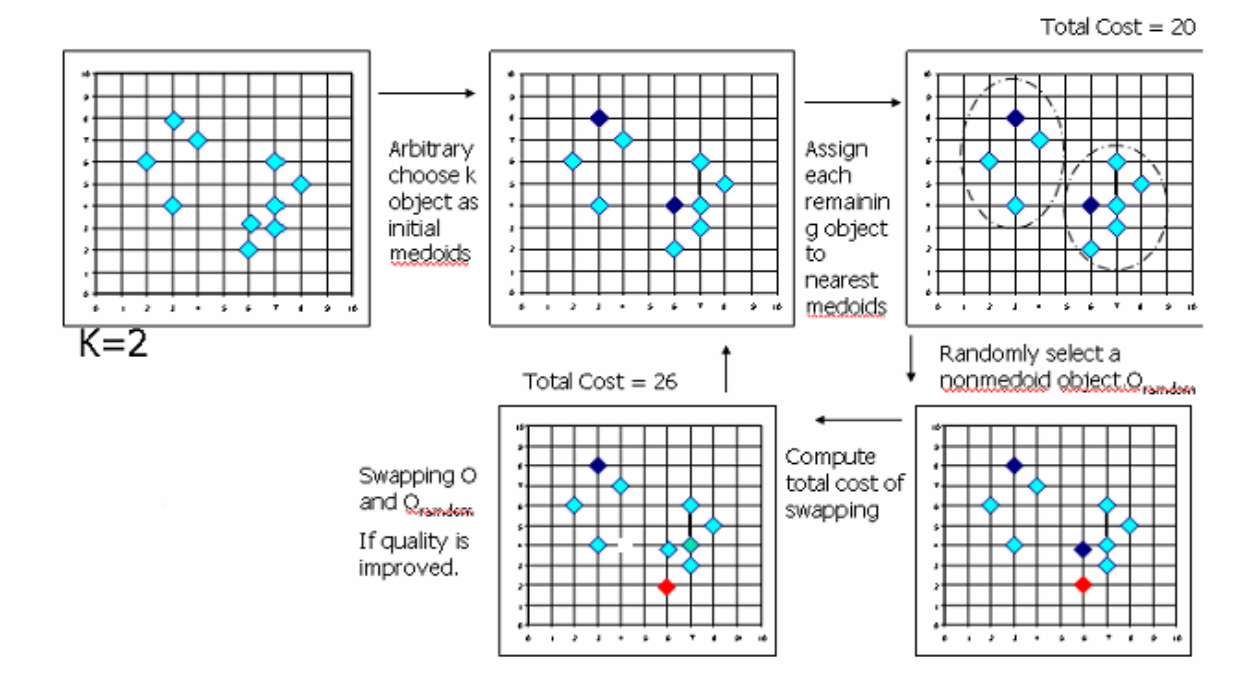

Figure 1: Working of K-medoid Algorithm [5].

2.3 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:

Table 1: Comparison of K-means ,K-medoids & clarans

Parameters	k-means	k-medoids	Clarans
Complexity	O (i k n)	O (i k (n-k)2)	O (n^2)
Efficiency	Comparatively more	Comparatively less	Comparatively more
Implementation	Easy	Complicated	complicated
Sensitive to Outliers?	Yes	No	No
Advance specification of No. of clusters 'k'	Required	Required	Required
Does initial partition affects result and Runtime?	yes	yes	Yes
Optimized for	Separated clusters	Separated clusters, small dataset	Separated clusters, large dataset

4. LIMITATION OF EXISTING ALGORITHM

K-Mean

- It is sensible to initial configuration
- Unsuccessful initialization gives empty clusters.
- Algorithm can apply on spherical clusters.
- The number of cluster should be define in advance
- It is too sensitive to outliers.

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.

Clarans

- It doesn't guarantee to give search to a localized area.
- It uses randomize samples for neighbours.
- It is not so much efficient for large dataset.

5. CONCLUSION

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 spherical based cluster which have small to medium sized dataset. However, to develop the understanding of parameters 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.