International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 370-373
e-ISSN:2278-621X

ANALYSIS OF DATA USING K-MEANS AND K-MEDOIDS ALGORITHMS

Swathi Dsouza[1], Jevita Deena Dsouza[2] & Vanitha T[3]

Abstract- Clustering is dividing of data set based on their behavior. Cluster analysis is analyzing between data based on their characteristics and groups similar elements into the cluster. K-Means algorithm is simple partitioning clustering technique and unsupervised clustering algorithm that cluster n group of elements based on behavior. K-Medoids algorithm is a partitioning clustering technique that is modified from K-means algorithm. In this paper, the two partitioned methods K-Means and K-Medoids algorithms are analyzed and most efficient algorithm are examined.
Keywords- Clustering, K-Medoids, K-Means, Cluster Analysis, Partitioning Custering.

1. INTRODUCTION

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Clustering technique focuses on partitioning collection of data into cluster of similar elements between themselves. Clustering is considered as unsupervised learning technique to find a pattern in a group of unknown data. Relying on the type of data set available the convenient clustering method is selected. Dissimilarities and likenesses are assessed based on characteristic values depicting the articles also, regularly include distance measures. Clustering is an data mining instrument has its root in numerous application zones, for example, biology, security, business, intelligence also, Web search. Cluster analysis or basically grouping is the way toward partitioning an arrangement of data points into subsets. Every subset is a bunch to such an extent that items in a group are like one another, yet dissimilar to other groups. The arrangement of groups resulting from a clustering analysis can be referred as clustering.

1.1 Partitioning

Partitioning based clustering generates a partition of the data such that each elements in a cluster that are related within themselves other than their elements in other cluster. Partitioning method is an iterative and effective when size of the dataset is smaller. In general, the real essential grouping strategies can be ordered into Partitioning Hierarchical, Density-based, Grid-based Methods.
Most existing popular methods of clustering are categorized as Partitioning, Grid based, Hierarchical, Model based methods.

2. K-MEANS ALGORITHM

It is an unsupervised learning algorithm with different data analysis applications widely used for mining data 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 combined every data entries. Each data entry of the result is related with centroids. Within these sets, the distance of centroid is minimized. The process to achieve the result sets of classified data. It basically consists on several iterations of a particular process, designed to get an optimal minimum solution for all data points.

2.1 Algorithm

The K-means calculation for partitioning, where each cluster's center is supplanted by the mean estimation of the elements in the cluster. K-means algorithm takes inputs k defines how many cluster that user wants to form and ddenotes n objects in a data set.
Method:

Step 1: Initial cluster center is defined by choosing k clusters.
Step 2: The k points is selected to the closest centroid of the cluster using Euclidean distance
Step 3: Recalculate the center point of the k points.
Step 4: Steps 2 to 3 is repeatedtill the convergence condition is met.

[1] Aloysius Institute of Management and Information Technology, Mangalore, Karnataka, India
[2] Aloysius Institute of Management and Information Technology, Mangalore, Karnataka, India
[3] Aloysius Institute of Management and Information Technology, Mangalore, Karnataka, India

2.2 Choosing The Quantity Of Clusters

The quantity of cluster should coordinate the data. A wrong decision of the quantity of bunches will invalidate the entire procedure. An exact approach to locate the best number of cluster is to attempt K-means bunching with various number of clusters and measure the subsequent whole of squares.

Introducing the position of the clusters

- Forgy: set the places of the k clusters to k perceptions picked randomly from the dataset.
- Random partition: allocate a cluster randomly to every perception and register means as in step3.

Forgy method:

Number of clusters 4;

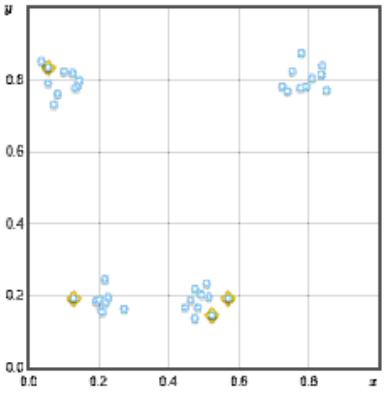

Fig1. Initializeclusters

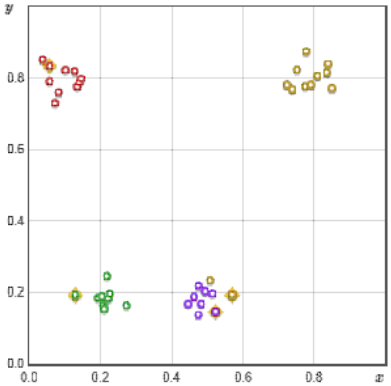

Fig 2. Assign data points to closer cluster

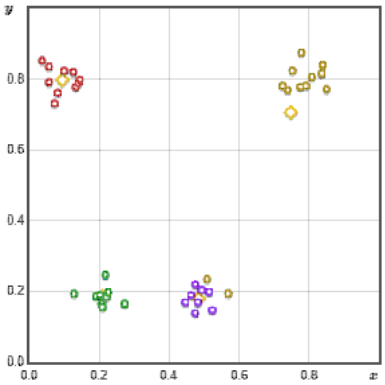

Fig 3.Calculate center of each cluster

Random Partition:

Number of cluster is 4.

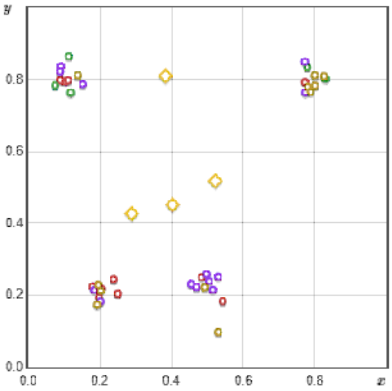

Fig 1.Initialize clusters

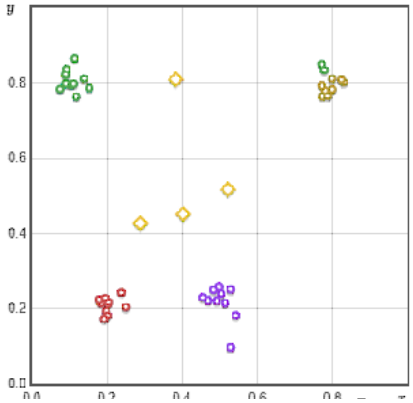

Fig 2.Assign data points to closer cluster

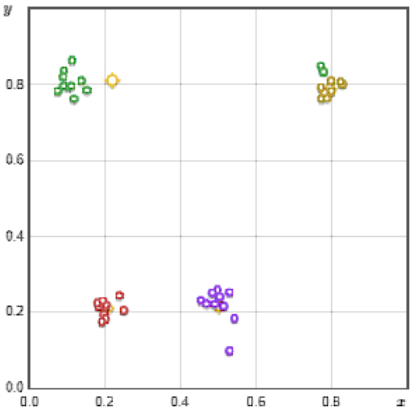

Fig 3.Calculate center of each cluster

Advantages

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

- Lots of Applications- It has several live world implementations on many different subjects.
- Fast – Achieves the result in a fast way.
- Simple and reliable- It solving the problem for a large set of information.
- Efficient– It gives a decent arrangement with moderately low figuring difficulty for bunching issue.
- Great Solution – Best outcome set will be given.

Disadvantages

- No Absolute Data – It can't be utilized on information sections since it's not repeat a mean capacity.
- Result Set – The outcome set isn't great.
- Initialization method – The results will change depending upon the initialization action.

3. K-MEDOID

The K-medoids or Partitioning AroundMedoids algorithm is a partitional cluster algorithm which is somewhat changed from the K-means. They both endeavor to limit the squared-blunder however the K-medoids algorithm is heartier to commotion than K-means algorithm. InK-means algorithm, they pick means as the centroids yet in the K-medoids, data point are chosen be the medoids. A medoid can be characterized as that object of a bunch, whose normal disparity to every one of the articles in the cluster is insignificant.

The K-means calculation is sensitive to anomalies in light of the fact that with the end goal that objects are far from most of the information, and consequently when appointed to a bunch they can dramatically distort the mean estimation of the Cluster.

3.1 Algorithm

Rather than centroids utilizes medoids the most central objects (the best representatives) of each cluster.
This allows using only dissimilarities d(r, s) of all pairs (r, s) of the objects.
Method:

- Randomly select k objects $m_1,...,m_k$ as initial medoids.
- Until the point when the most extreme number of iteration is come to or no improvement of the target function has been found do:
  - Compute the clustering based on $m_1,...,m_k$ by clustering based on each point to the closest medoid and figure the evaluation of the objective capacity.
  - For all sets (m_i , x_s), where x_s is a non-medoid guide, attempt toward enhance the objective capacity by taking $x_s$to be another medoid point and mi to be a non-medoid point.

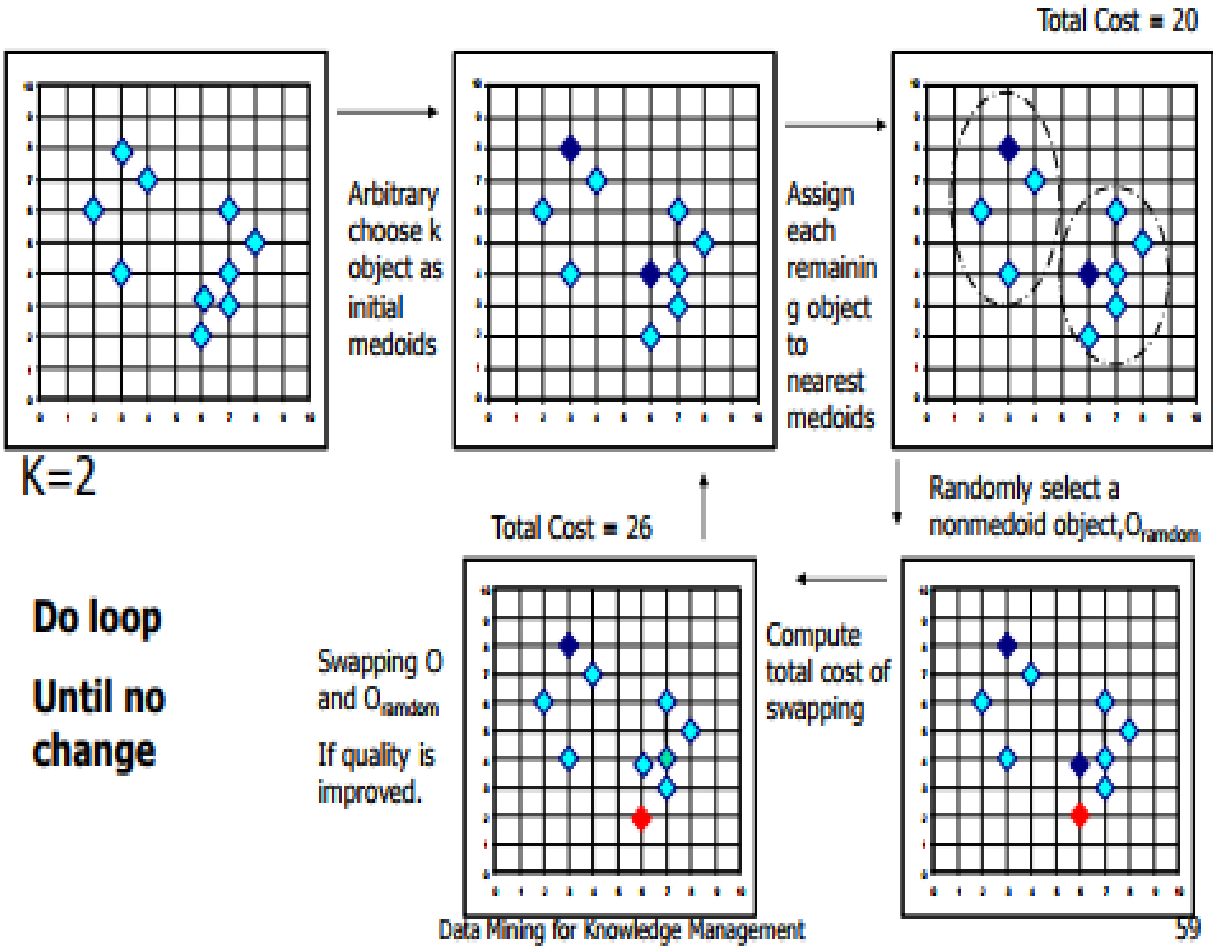

Fig : K-Medoid (PAM)

Advantages:

- Easy to understand and execute.
- Quick and convergent in a predetermined number of steps.
- Normally less delicate to outliers than k-means.
- Allows using general dissimilarities of objects.

Disadvantages:

- Different initial sets of medoids can lead to different final clustering. It is thus advisable to run the procedure several times with different initial sets of medoids.

- The resulting clustering depends on the units of measurement. If the variables are of different nature or are very different with respect to their magnitude, then it is advisable to standardize them.

4. COMPARISON

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

K-medoid could be heartier to commotion and anomalies when contrasted with k-means because of the fact that it limits a whole of general pairwise dissimilarities rather than a total of squared Euclidean separations.K-means and K-medoids both requires clusters k to be specified in the input. K-medoid is less influenced by the outliers in the data and computationally more expensive. K-means selects initial clusters as k where in k-medoid replacing the mean of the cluster with modes.Medoid is more robust than k-means in the presence of noise and outliers.

5. CONCLUSION

In this paper we conclude that the final obtained results from both K- mean and K-Medoids clustering algorithms with respect to the formed number of clusters and distance metric.The outcomes looked at between K-Medoids and K-Means demonstrate that the time taken in cluster head determination and space complexity quality of covering of group is vastly improved in K-Medoids than K-Means. As well as the dataset result shows that K-Medoids is better in all aspects such as execution time, non-sensitive to outliers and reduction of noise but with the limitation that the complexity is greater as compared to K-Means.

6. REFERENCES

[1] Hae-Sang Park, Chi-HyuckJun:"A simple and fast algorithm for K-medoids clustering"
[2] Parvesh Kumar, Siri KrishanWasan:Comparative Analysis of k-mean Based Algorithms
[3] MahendraTiwari,RandhirSingh:"Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data"
[4] Shailendra Singh Raghuwanshi, PremNarayanArya:"Comparison of K-means and Modified K-mean algorithms for Large Data-set"
[5] Dr. T. Velmurugan :"Efficiency of k-Means and K-Medoids Algorithms forClustering Arbitrary Data Points"
[6] "Data Mining Concept and Techniques",2nd Edition, Jiawei Han, By Han Kamber
[7] TagaramSoniMadhulatha :"Comparison between K-Means and K-Medoids Clustering Algorithms"
[8] Preeti Arora, DeepaliDr,ShilpraVarshney: "Analysis of k-means and k-medoid algorithm for Big Data"