A Study of Different Partitioning Clustering Technique

Ashish Goel[1]
[1]Department of Computer Science and Engineering
[1]Galgotia Collage of Engineering & Technology, Uttar Pradesh Technical University, Gr. Noida, U.P

Abstract— In the field of software, Data mining is very useful to identify the interesting patterns and trends from the large amount of stored data into different database and data repository. Clustering technique is basically used to extract the unknown pattern from the large set of data for electronic stored data, business and real time applications. Clustering is a division of data into different groups. Data are grouped into clusters with high intra group similarity and low inter group similarity [2]. Clustering is an unsupervised learning technique. Clustering is useful technique that applied into many areas like marketing studies, DNA analysis, text mining and web documents classification. In the large database, the clustering task is very complex with many attributes. There are many methods to deal with these problems. In this paper we discuss about the different Partitioning Based Methods like- K-Means, K-Medoids and Fuzzy K-Means and compare the advantages or disadvantages over these techniques.

Key words: Data Mining, Clustering, K-Means Clustering, Fuzzy K-Means Clustering, K-Medoids Clustering.

I. INTRODUCTION

In the field of data mining, the clustering technique play a very important role that divide the huge amount of data into similar type of group on the basic of requirement. The partitioning clustering algorithm divides the" n" objects into "k" clusters with maximum intra-class similarity or minimum inter-class similarity. These clusters represent the groups of data and provide the representation of many data objects by fewer clusters. Clustering is very useful technique to deal with the statistical data and unsupervised learning.

Clustering is used in the many fields like machine learning, image analysis, pattern recognition, outliers' analysis, market analysis and bio-informatics and many more. Various Developers have used different methods to achieve clustering in different ways. For managing a large dataset using clustering technique should be satisfy some requirement like scalability, discovering arbitrary shape of clusters, handle with the different type of attribute, deal with the noisy or outliers, interpretability and usability.

Clustering methods have different categories like 1) Partitioning Based Methods, 2) Hierarchical Methods, 3) Grid Based Methods, 4) Density Based Methods, 5) Neural Network Based Methods, 6) Constraints Based Methods 7) Model Based Methods.

Clustering algorithms can be classified into two types like:

- Exclusive Clustering
- Overlapping Clustering

In the way of Exclusive, A data point belongs to one cluster then its never belongs to the any other cluster. But in the way of Overlapping, A data point belongs to one cluster may be belong to the two or more clusters with different degree of membership. K-Means and K-medoids is an exclusive clustering algorithm, where Fuzzy K-Means is an overlapping clustering algorithm.

In this paper we focus on the Different Partitioning Based Methods: 1) K-Means, 2) K-Medoids, 3) Fuzzy K-Means. These Methods are discussed with their algorithms, strength and limitations.

A. K-Means

K-Means is based on the hard clustering. K-means is very commonly used partitioning technique that is mostly used for analyze data and trends in the large amount of data. K-Means is one of the most unsupervised learning methods. K-Means referred to the Hard clustering means data points belongs from the one cluster never belongs any other cluster. In other word we can say that Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. K-Means divides the given set of "N" objects into "K" cluster, where the K is the no. of cluster. Each cluster has a centre point i.e. centroid [1, 6]. All the data points are placed in a cluster having centroid nearest (or most similar) to that data points. After processing all data points, k-means, or centroids, are recalculated, and the entire process is repeated. On the basis of the new centroids, all the data points are bound to the clusters. In the each iteration centroid moves from different data points. This process is continuous until no any centroid move. At last we found the K cluster with N data points.

1) Algorithm of the K-Means:

- Step1. Choose the K Number of clusters to partition N data objects.
- Step2.Generate K clusters and determines the cluster's center.
- Step3**.** Assign each object to the cluster to which the object is the most similar; based on the given similarity function.
- Step 4. Update the cluster means (centroid).
- Step5**.**Repeat steps 3 and 4 until no change occurs in the clusters.

2) Advantages:

- Easy to understand and implement.
- With a large number of variables, K-Means may be computationally fast.
- K-means is based on the exclusive clustering then its produce tighter clusters.

3) Disadvantages

- Not applicable to categorical data.
- Unable to handle noisy data and outliers.
- It does not work well with clusters when original data have Different size and Different density of data.
- Result and total run time depends upon initial partition

B. Fuzzy K-Means

In the hard clustering method, divided data belongs to the exactly one cluster. Fuzzy K-Means is based on the soft clustering method where the data elements belong to the more than one. FCM is also an unsupervised clustering algorithm FCM used in the feature analysis, clustering, classifier design, agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition [6]. In the FCM algorithm data analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster. In fact, FCM is a data clustering technique in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the center of a cluster which will have a low degree of belonging to that cluster [4].

1) Algorithm of Fuzzy K-Means:

- Step1. Choose the K Number of clusters to partition N data objects.
- Step2.Assign randomly K cluster centre.
- Step3**.** Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than, the given sensitivity threshold).
- Step4. Compute the center of each cluster.
- Step5**.**For each point, compute its coefficients of being in the cluster.

2) Advantages

- Gives best result for overlapped data set and comparatively better then k-means algorithm.
- Unlike k-means where data point must exclusively belong to one cluster centre here data point is assigned membership to each cluster centre as a result of which data point may belong to more than one cluster centre.

3) Disadvantages

- Apriori specification of the number of clusters.
- With lower value of β we get the better result but at the expense of more number of iteration.
- Euclidean distance measures can unequally weight underlying factors.

C. K-Medoids

K-Medoids is also based on the hard clustering. Like K-Means. Both the k-means and k-medoids algorithms are partitioning [9, 10] the data into groups and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. But In comparison of both K-Means and Fuzzy K-Means, K-Medoids uses the median or PAM (Partition Around Medoids) in place of centroid. .because K-Means uses centroid to represent the cluster and it is not deal with the outliers. This means, a data object with an extremely large value may disrupt the distribution of data. K-medoids method deals with this by using medoids to represent the cluster rather than centroid. A medoid is the most centrally located data object in a cluster in the place of centroid. In the K-Medoids Method, k number of data objects are selected randomly as medoids to represent k cluster and other 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 in the place of centroid which can represent clusters in a better way and again the entire process is repeated. And all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location one by one. This process is continued until no any medoid remaining for move. At last we found the, K clusters that representing a set of n data objects.

1) Algorithm of K-Medoids:

- Steps1 Choose randomly K number of cluster to partitioning of N data objects.
- Step 2 Associate each data point to the closest medoid.
- Step 3 for each medoid.
  - For each non-medoid data point Like O
  - Swap medoid and O and compute the distance.
- Step 4 Select the distance with lowest cost.
- Step 5 Repeat steps 2 to 4 until there is no change in the medoid.

2) Advantages:

- More robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean

3) Disadvantages:

- Relatively 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.
- Relatively not so much efficient.
- Need to specify *k*, the total number of clusters in advance.
- Result and total run time depends upon initial partition.

D. Comparison of K-Means, Fuzzy K-Means and K-Medoids

	K-means	**Fuzzy K-means**	**K-medoids**
Complexity	O (i k n)	O (I k (n)2)	O (i k (n-k)2)
Efficiency	Comparatively more	Comparatively more than K-Medoids	Comparatively less
Implementation	Easy	Less complicated than K-Medoids and Complicated to K-Means	Complicated
Sensitive to Outliers?	Yes	No	No
Necessity of convex shape	Yes	Not so much	Not so much
Advance specification	Required	Required	Required

of no of clusters 'k'			
Does initial partition affects result and runtime?	Yes	Yes	Yes
Optimized for	Separated clusters	Separated cluster and categories data	Separated clusters,

Table 1: Comparison of K-Means, Fuzzy K-Means And K-Medoids

II. Conclusion A

After Studied of Different Partitioning Clustering Technique, it can be concluded that partitioning based clustering methods are suitable for spherical shaped clusters in small to medium sized data sets. K-means, Fuzzy K-means and K-medoids – All three methods find out clusters from the given database. All three methods require specifying *k*, no of desired clusters, in advance. Result and runtime depends upon initial partition for both of these methods. The advantage of k-means is its low computation cost, while drawback is sensitivity to noisy data and outliers and Both Fuzzy k-means or k-medoid is not sensitive to noisy data and outliers, but it has high computation cost. So at last we can say that after above discussion, If data size is small or medium size than we can use the K-Means where the data is not noisy or we can use the Fuzzy K-means or K-Medoid when the data is noisy but K-Medoid is more complex to execute. So we can say that all three Partitioning techniques have some advantages or disadvantages and user can be using these techniques as per the requirement of the project or need of the project condition.

Reference

[1] A.K. Jain, June, 2010, "Data Clustering: 50 Years Beyond K-Means", Volume No. 31, Issue 8, pp: 651-666.
[2] A. K. Jain, 1999" Data clustering: a review". ACM Computing Surveys, Volume No .31, Issue 3, pp: 264– 323.
[3] Abdullah Al-Mudimigh, 2009, "Efficient implementation of data mining: improve customer's behavior", pp 7-10.
[4] Bailey, Ken. 1994. "Typologies and Taxonomies-Numerical Taxonomy and Cluster Analysis".
[5] Bezdek, James C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Volume No.ISBN 0, pp: 306-406.
[6] Bharati R.Jipkate ,2012 "A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms", pp: 2250–3005
[7] Jawed Siddiqi, , July-September 2002, "A framework for the implementation of a Customer Relationship Management strategy in retail sector".
[8] Joseph L. Breeden,, January 1999, "GA-Optimal Fitness Functions, Center for Adaptive Systems Applications", Inc. 1302 Osage, Suite A Santa Fe, NM 87505.
[9] Rakesh Agrawal and Ramakrishnan Srikant, 1994 "Fast algorithms for mining association n rules in large databases". Volume No. 20, pp: 487-499.
[10] T. Velmurugan, 2011 , "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach", Volume No. 10 ,Issue No .3 , pp: 478- 484.a

