ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 6, June 2016

An Overview of Partitioning Algorithms in Clustering Techniques

Swarndeep Saket J, Dr. Sharnil Pandya

***Abstract*— Data Mining as an area of computer science has been gaining enormous importance in various fields of business, spatial planning and predictive analysis. Clustering techniques are application tools to analyze stored data in various fields. It is a process to partition meaningful data into useful clusters which can be understandable and has analytical value. In the present paper after giving a brief outlook of data mining and clustering techniques, we have made a comparative study of various partitioning algorithms so as to study their worth at a level playing field. The analysis shows that not all partitions algorithms are efficient to handle large datasets. The exploration of partitioning algorithms opens new vistas for further development and research.**

***Index Terms*—** Clustering, k-means, k-Medoids, Clarans, Calara

I. INTRODUCTION

Data mining is the technique of exploration of information from large quantities of data so as to find out predictably useful novel and truly understandable complex pattern of data. Such an analysis must ensure that the pattern in the dataset holds good and hither to not known (novel).The technique of exploration must ensure study of a useful pattern which is completely understandable and can be interpreted and comprehended. From computational point of view, it is powerful new technology with great potential to facilitate current market specialize in the foremost necessary info in their information. These tool are helpful to predict to understand the current market, future trends and behavior and enables us to take proactive measures to make more commercial sense. No Doubt, data mining and clustering techqniues have become very useful for large datasets even in social media such as face book and twitter [1]. It should be noted here that there is no dearth of large datasets in real world, not to talk about an abundance of data on the web and virtual stores. Data mining is also called Knowledge Discovery in Database.

A. Introduction to Clustering:

A cluster may be treated as a subset of objects which are similar in nature. It is a 'unsupervised learning process' to group together similar data samples, although, the criteria of classification might differ from each other [2]. To be more precise, a cluster might be defined as collection of data objects with numerous possibilities of classification.

"Clustering can also be used for outlier detection". [3]".Cluster analysis can be used as a complete data processing tool to achieve insight into data distribution."[4] Clustering is used in wide variety of application such as psychology, market research, pattern recognition, data analysis, image processing, city planning and biology [4].

The importance of clustering in data mining can be understood from the fact that according to a survey of K.D. Nuggets in 2011,it has stood third most frequent task in the world .It has helped the experts to erect comprehensible structures in large datasets. Application of clustering have become a very successful tool for classification of documents, Clustering of web log data, recognition of meaningful patterns of data, undertaking spatial data analysis and even creating thematic maps in GIS and image processing.

This paper is organized as follows: Section 1 introduces the data mining and cluster analysis. Section 2 gives an overview of various types of clustering algorithms. In the section 3 gives information about different types of partitioning algorithms with its procedure, advantages and disadvantages. Section 4 summarizes the concluding remarks of the paper.

II. CLUSTERING TECHNQIUES: A BIRDS' EYE VIEW

The clustering techniques differ in various ways and can be categorized as 'Partitioning techniques', 'Hierarchical techniques', 'Density-based techniques', 'Grid based methods' and 'model-based methods'. These methods are peculiar in nature and they handle some or other issues of clustering individually. However clustering techniques do not find favour because of 'computational complexity' and requirement of entries of dataset into the memory [6].

2.1 Hierarchical Clustering Methods:

Hierarchical clustering method seeks to build a' tree based hierarchical taxonomy from asset of unlabeled data. This grouping process is represented in the form of dendrogram. It can be analyzed with the help of statistical method. There are two types of hierarchical clustering methods. They are 1)Agglomerative hierarchical clustering and 2)divisive clustering [7].In the agglomerative approach which is also known as 'bottom up approach', Hierarchical algorithms always result into what is called 'nested set of partitions'. They are called hierarchical because of their structure they represent about the dataset. Divisive and Agglomerative strategies are two important strategies of hierarchical clustering. In case of divisive approach, popularly known as 'top down approach', 'all data points are considered as a single cluster and spitted into number of clusters based on certain criteria'.[8] Examples for such algorithm are BRICH (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) .The most important weakness of hierarchical clustering technique is that it does not scale properly because of time complexity. In addition to this it is difficult to alter ones the

process of analysis has already started.

2.2 Density Based Clustering Methods:

Density based method has been introduced on the basis of density."Benfield and Raftery opined that Density based methods assume that the points that belong to each cluster are drawn from a specific probability distribution"[9].This algorithm can be used for only spherical-shaped clusters. The merit of such clustering is that they have considerable higher density of points than outside the cluster. This method can be effective in handling the noise to some extent provided we can scan the 'input dataset' It only needs one scan of the input dataset. The precondition of this algorithm is that the density parameters should be initialized *aprori*. It permits the given cluster to grow continuously as long as the density of neighborhood exceeds a certain threshold. DBSCAN, DENCLUE and OPTICS are examples of density based methods [10].

2.3 Grid Based Clustering Methods:

A grid based structure is formed by this algorithm by quantizing the object space into finite number of cells. This means that 1)The data space is first partitioned into definite number of cells.2)The cell density for each of the cell is calculated.3)The cells are classified through sorting according to their densities.4)The center of the cluster is identified.5)The distance between the neighboring cells are calculated. The main advantage of the grid based method is fast processing time, irrespective of number of data objects. The main feature of this algorithm is that it does not require computing distances between two data objects. Clustering is performed only at summarized data points. STING. Wave Cluster and CLIQUE are examples of grid based methods.

2.4 Model Based Clustering Methods:

This algorithm is based on hypothesizing a model for every cluster to find best fit of the data according to the mathematical model. It can automatically determine the number of cluster based on standard statics. The method may locate clusters by constructing a density function that reflects the spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics taking outlier. It therefore yields robust clustering method [7]. STASTICAL approach and COBWEB are examples of model based clustering methods.

2.5 Partition Clustering Methods:

Given a database of n objects, it constructs k partitions of the data. Each object must belong to exactly one group. Each group must contain at least one object. Partition technique can improve iterative relocation technique by mining objects from one graph to another. The main objective of partition clustering algorithm is to divide the data points into K partitions. Each partition will reflect one cluster. The technique of partition depends upon certain objective functions .For example 'minimizing the square error criterion'. The weakness of such an algorithm is that whenever the distance between the two points from the center are close to another cluster, the result becomes poor or misleading due overlapping of the data points.

III. TYPES OF PARTITIONING ALGORITHMS:

There are mainly four types of partitioning algorithm includes as K-Mean Algorithm, K-Mediod Algorithm i.e PAM(Portioning Around Medoids), CLARA and CLARANS.

3.1 K-Mean Algorithm:

K-Mean is first developed by James Macqueen in 1967.A cluster is represented by its centroid, which is usually the mean of points within a cluster. "The objective function used for k-means is the sum of discrepancies between a point and its centroid expressed through appropriate distance"[12]. They have convex shapes clusters.

Procedure of K-Mean [13]:-

a) The technique requires arbitrary selection of choose k objects from D as the initial centers, where k is the number of clusters and D is the data set containing n objects.

b) Repeat the first step.

c) Reassign each object to the cluster to which object is most similar. It is based on the mean value of the objects in the cluster.

d.) Calculate the mean value of the objects for each cluster.

e) Until no change

Advantages of K-Mean:

1. If the variables are large, then K-Means most of the time computationally faster than hierarchical clustering methods.

2.K-Means produces tighter clusters than Hierarchical Clustering Method.

Disadvantages of K-Means Partition Algorithm:

1.It is difficult to predict the K Value.

2. More difficulty in comparing quality of cluster.

3.K-Means Algorithm does not work well with global clusters.

3.2 K-Medoid Algorithm:

Partition Around Mediods (PAM) is developed by Kaufman and Rousseuw in 1987. It is based on classical partitioning process of clustering The algorithm selects k-medoid initially and then swaps the medoid object with non medoid thereby improving the quality of cluster. This method is comparatively robust than K-Mean particularly in the context of 'noise' or 'outlier'. K-Medoid can be defined as that object of a cluster, instead of taking the mean value of the object in a cluster according to reference point. K-Medoids can find the most centrally located point in the given dataset.

Procedure of K-Medoid[13] :-

Input:

- K:The number of clusters
- D: A data set containing n objects

Output:

- A set of K clusters

Method:

The following steps are recommended by Tagaram Soni Madhulatha [14]

1. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points(n>K).
2. After selection of the K medoid points, associate each data object in the given data set to most similar medoid.
3. Randomly select non-medoid object O.
4. Compute total cost S of swapping initial medoid

object O.

5. If S>0, swap initial medoid with the new one.
6. Repeat steps until there is no change in the medoid.

Advantages of K-Medoid:

1) It is simple to understand and easy to implement.

2) K-Medoid Algorithm is fast and converges in a fixed number of steps.

3) Partition Around Medoid (PAM) algorithm is less sensitive to outliers than other partitioning algorithms.

Disadvantages of K-Medoid:

1)K-Mediods is more costly than K-Means Method because of its time complexity.

2) It does not scale well for large datasets.

3)Results and total run time depends upon initial partitions

3.3 CLARA(Clustering for Large Application):-

CLARA means clustering large applications and has been developed by Kaufman and Rousseuw in 1990 [15]. 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[16].

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 the specific K medoid 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'.

Advantages of CLARA:

1) CLARA Algorithm deals with larger data sets than PAM (Partition Around Mediods).

Disadvantages of CLARA:

1) The efficient performance of CLARA depends upon the size of dataset.

2).A biased sample data may result into misleading and poor clustering of whole datasets.

3.4 CLARANS:

K-Medoid algorithm does not work effectively for large data sets. Therefore CLARA has been improved and modified so as to used large databases. CLARANS has been developed by Ng and Han in 1994 [17].To overcome the limitations of K-Medoid algorithm clarans is introduced. Clarans (Clustering large Application Based on Randomized Search) is partitioning method used for large database. It is more efficient and scalable than both PAM and CLARA. As in case of CLARA the authors (ibid) have recommended the following steps to be performed:

1) Input parameters numlocal and maxneighbour.

2) Select K object from the database object D randomly.

3)Mark these K object as selected Si and all other as non-selected Si.

4) Calculate the cost T for selected Si.

5) If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.

6) Restart the selection of another set of medoid and find another local optimum.

7) CLARANS stops until returns the best.

According to the authors (ibid) CLARANS uses two parameters – numlocal and maxneighbour. Numlocal means number local minima obtained and maxneighbhour means maximum number of neighbhour examined. 'The higher the value of latter, the closer will be CLARANS to PAM and longer will each search of local minima be'. This is an advantage because the quality of local minima is higher and less number of local minima are to be found out.

Advantages of CLARANS:

1)It is easy to handle outliers.

2) CLARANS result is more the effective as compare PAM and CLARA.

Disadvantages of CLARANS:

1)It does not guarantee to give search to a localized area.

2) it uses randomize samples for neighbhours.

3) it is not much efficient for large datasets.

IV. COMPARISON

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

Parameters	k-means	K-mediods	CLARA	CLARANS
Complexity	O(tkn)	O(tk(n-k)2)	O(ks2+k(n-k))	O(n^2)
Efficiency	Comparatively more	Comparativel yless	Comparatively more	Comparatively more
Implementation	Easy	Complicated	Complicated	Complicated
Sensitivity to Outliers?	Yes	No	No	No
Advance specification of No. of clusters 'k'	Required	Required	Required	Required
Optimized for	Separated clusters	Separated clusters, small dataset	Separated clusters, large dataset	Separated clusters large dataset

Table 1: Comparison of K-means ,K-Medoids, Clara and clarans:

V. CONCLUSION

In the present study an overview has been given on data mining and clustering techniques. The paper analyses four important partitioning algorithms known as K-means, K-Medoids, CLARA and CLARANS. The study presents a comparative table to understand merits and demerits of each of the algorithms. The analysis shows that CLARA and CLARANS are comparatively more efficient and scalable than other algorithms. However algorithms such as K-Means and K-Medoid can be further modified to make them equally efficient and scalable. However, further research is required to study efficiency parameters of each of the partitioning algorithms.

ACKNOWLEDGMENT

We are greatly thankful to our family and friends for the continous support that has provided a healthily environment a drive us to this project. We also thank the Principal and the Management of PIET for extending support to this work.

REFERENCES

[1] Madhuri V. Joseph, "Significance of Data Warehousing and Data Mining in Business Applications, International journal of Soft Computing and Engineering, Vol No:3,Issue no:, March 2013.
[2] J. Kleinberg, "An impossibility theorem for clustering," in Proc. 2002 Conf. Advances in Neural Information Processing Systems, vol. 15, 2002, pp. 463–470.
[3] Yujie Zheng, "Clustering Methods in Data Mining with its Applications in High Education", International Conference on Education Technology and Computer, 2012.
[4] Er. Arpit Gupta, Er.Ankit Gupta, Er. Amit Mishra," RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS", International Journal of Advance Technology & Engineering Research, November 2011, Vol. 1, Issue 1, pp 39-47.
[5] C. Chen, L. Pau, and P.Wang, Eds.,"Handbook of Pattern Recognition and Computer Vision" 1993, pp. 3–32.
[6] Javier Bejar," Strategies and Algorithm for Clustering Large Datasets: A Review
http://upcommons.upc.edu/bitstream/handle/2117/23415/R13-11.pdf
[7] Prabhdip Kaur, Shruti Aggrwal ," Comparative Study of Clustering Techniques", international journal for advance research in engineering and technology, April 2013.
[8] S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, , "A Survey On Partition Clustering Algorithms", January 2011.
[9] Banfield J. D. and Raftery A. E.," Model-based Gaussian and non-Gaussian clustering", Biometrics, 49:803-821, 1993.
[10] A. Gordon, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Bada, Eds "Cluster validation," in Data Science, Classification, and Related Methods. New York: Springer-Verlag, , pp. 22–39, 1998.
[11] P. Berkhin, "Survey of Clustering Data Mining Techniques",. Ttechnical report, AccrueSoftware, San Jose, Cail, 2002.
[12] Pradeep Rai and Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications (0975-8887) Vol 7-No. 12, pp. 1-5, October 2010.
[13] Jiawei Han and Micheline Kamber, "Data MiningTechniques", Morgan Kaufmann Publishers, 2000.
[14] Tagaram Soni Madhulatha. "Comparison between K-Means and K-Medoids Clustering Algorithms", Communications in Computer and Information Science, 2011.
[15] KAUFMAN, L and ROUSSEEUW, P.J. "Finding Groups in Data", New York; John Wiley,1990.
[16] Raymond T. Ng and Jiawei Han," CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transaction on Knowledge and Data Engineerring, Vol No 14,No 5, September/october 2002.
[17] NG, R.T. and HAN, J.," Efficient and Effective Clustering Methods for Spatiial Data Mining, Proceedings of the International Conference on Very Large Data Bases (VLDB '94).Santiago, Chile,144-155, September 1994.
[18] K.H.Wandra and Sharnil Pandya," Centralized Timestamp based Approach for Wireless Sensor Networks, international journal of computer applications,Vol.No91,Issue No2,April 2014.