Conference Proceeding Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Enhancing K-Means Algorithm Clustering Performance with Improved Time Complexity

Ayush Tiwari
Research Scholar, Department of Computer Science & Engineering, Global Nature Care Sangathan's Group of Institutions, Jabalpur, India

Abstract: Clustering plays a very important role in research area in the field of data mining. Clustering is a method of subdivide a set of data in an important sub classes called clusters. It assists users to recognize the natural cluster from the data set. It is unsupervised arrangement that means, it has no prearranged classes. This paper introduces analysis of many partitioning methodology of clustering algorithms and their relative research by reflecting their profits freely. Implementation of cluster analysis are Economic Science, Document categorization, Pattern Recognition, Image Processing, text mining. No particular algorithm is effective enough to solve problems from different fields. Hence, in this study some algorithms are presented which can be used according to one's necessity. In this paper, different familiar partitioning- based methods – k-means, k-medoids and Clarans – are studied and compared. The study given here survey the behavior, nature and efficiency of these three methods. Data Mining is a method of classifying valid, suitable, new, logical pattern in the data.

Keywords*: Clustering, K-Means Algorithm , K-Medoids Algorithm , Clarans Algorithm*

I. INTRODUCTION

Data Mining is a method of classifying valid, suitable, new, logical pattern in the data. Data Mining is related with solving problem by studying existing data[1]. Clustering is a process of data explorations, a procedure of finding patterns in the data that of our interest in clustering off.unsupervised learning that means we don't know in advance how data should be group together. Various Techniques for clustering are as follows:

- Partitioning Method
- Hierarchical Method
- Grid- based Method
- Density-based Method
- Model-based Method

Among all these methods, this paper is pointed to explore partitioning based clustering methods which are k-means, k-medoids and clarans. These methods are discussed along with their algorithms, strength, benefits and limitations.

Data mining is the latest interdisciplinary field of computational science. Data mining is the process of discovering attractive information from large amounts of data stored either in data warehouses, databases, or other information repositories. It is a process of automatically discovering data pattern from the massive database [1]. Data mining refers to the extraction or "mining" of valuable information from large data volumes . Nowadays, people come across a massive amount of information and store or represent it as datasets. Process discovery is the learning task that works to the construction of process models from event logs of information systems [6]. Fascinating insights, observable behaviours, or high-level information can be extracted from the database by performing data mining and viewed or browsed from various angles. In data mining, many data clustering techniques are used to trace a particular data pattern [2]. Data mining methods for better understanding are shown in Fig. 1.

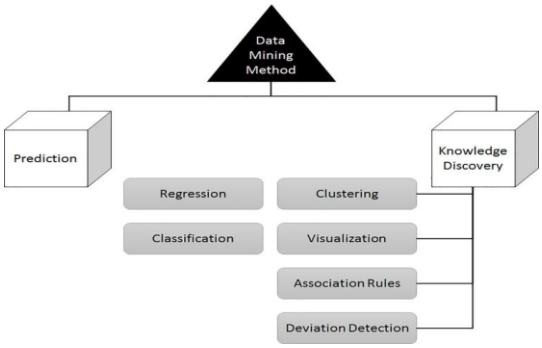

Fig. 1. Methods of Data Mining Techniques.

Ease of Use

Partitioning techniques distributes the object in various partitions where single partition defines cluster. The objects with in single clusters are of similar features where the objects of different cluster have dissimilar features 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 and claran as are partitioning algorithm[1]

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.[2] 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**.**

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

Input:

- K: the number of clusters
- D: a data set containing nobject

Output:

- A set of k clusters

Method*:*

1. Arbitrarily choose k objects from D as the initial cluster centers.

2.Repeat

3. 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;

4.Update the cluster means, i.e., calculate the mean value of the objects for each cluster.

5.Until no change;

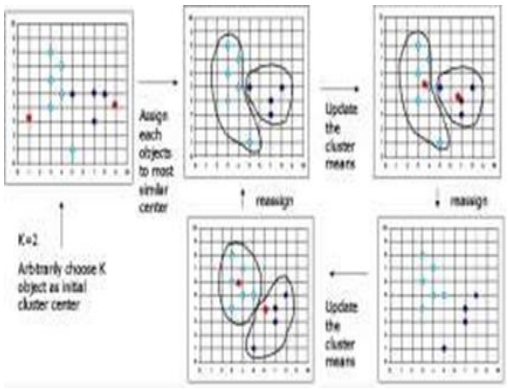

Figure 2.1 Working of K-mean Algorithm

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 method 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 method 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 nobjects.

Outputs*:*

- A set of k clusters.

Method:

a. Randomly choose k objects in D as the initial representative objects.
b. Repeat
c. Assign each remaining object to the cluster with the nearest representative object;
d. Randomly select a non-representative object, O random.
e. Compute the total cost of exchanging representative object, Oj with O random;
f. If S<0 then swap Oj with O random to form the new set of k representative object;
g. Until no change;

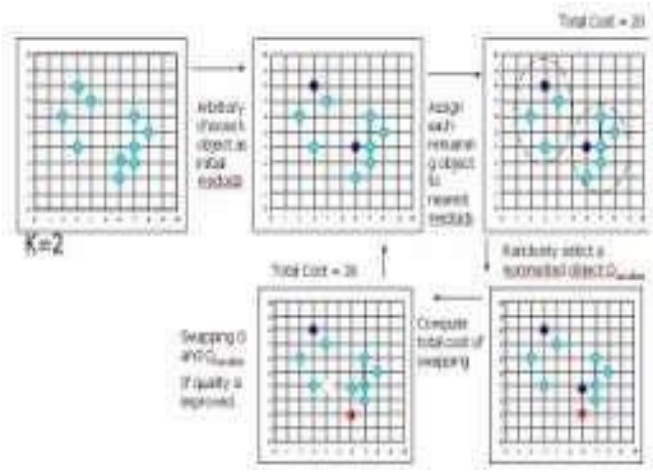

Figure 2.2: Working of K-medoid Algorithm

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).

COMPARISON

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(ikn)	O(ik(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

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.

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.

FUTURE DIRECTIONS AND OPEN ISSUES

To date, Data Mining and information disclosure are advancing an essential innovation for businesses and scientists in numerous domains. Although information mining is extremely powerful, it faces innumerable difficulties during its usage. The problems could be identified with performance, data, strategies, and procedures utilized. The information mining measure becomes effective when the challenges or issues are distinguished accurately and sifted through appropriately.[7]

Some of the following challenges and future directions are:

i. **Efficiency and Scalability of Algorithms:** The data mining algorithms must be proficient and adaptable to extricate data from gigantic sums of information within the database. So, as a future direction, develop a parallel formulation of an Improved rough k-means algorithm to enhance the efficiency of an algorithm.

ii. **Privacy and Security:** Information mining ordinarily leads to genuine issues in terms of information security, protection, and administration. For case, when a retailer reveals his clients purchasing details without their permission. So, as a future direction, there needs to develop a single cache system and DES (Data Encryption Standard) techniques in any Clustering Algorithm to improve the privacy and security of data in the cloud.

iii. **Complex Data Types:** Complex data elements, objects with graphical data, temporal data, and spatial data may be included in the database. Mining of these types of data isn't practical to be done one device.

iv. **Performance:** The execution of the data mining framework depends on the proficiency of calculations and procedures are utilizing. The calculations and strategies planned are not up to the marked lead to influence the performance of the data mining process.

Therefore, as a future direction, we need to introduce a new hybrid approach of an Improved Rough k-means Algorithm, and the Genetic Algorithm will improve the performance and handles the complex data. The combination of Partitioning Clustering and Hierarchical Clustering Algorithms will also increase the accuracy of data analysis.

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, Result and total run time depends or modify, such that the strengths, performance and efficiency of these techniques can be improved.

References

[1] P. Indira priya and D. D. K. Ghosh, "A Survey on Different Clustering Algorithms in Data Mining Technique," Int. J. Mod. Eng. Res., vol. 3, no.1, pp. 267–274, 2013.

[2] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N.Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining,"Int. J. Eng. Res. Appl. www.ijera.com, vol. 2, no. 3, pp. 1379–1384,2012.

[3] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma,"Machine learning techniques for data mining: A survey," 2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013, no. I, 2013

[4] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998.

[5] Shalini S Singh & N C Chauhan ,"Kmeans v/s K-Medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology,2015.

[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, 2012.

* * * * *