Implementation of Clustering Algorithms for Real Time Large Datasets

P.Rajyalakshmi , M.Kiran Kumar, G.Uma Maheswari, P.Naresh

Abstract : Now a day’s clustering plays vital role in big data. It is very difficult to analyze and cluster large volume of data. Clustering is a procedure for grouping similar data objects of a data set. We make sure that inside the cluster high intra cluster similarity and outside the cluster high inter similarity. Clustering used in statistical analysis, geographical maps, biology cell analysis and in google maps. The various approaches for clustering grid clustering, density based clustering, hierarchical methods, partitioning approaches. In this survey paper we focused on all these algorithms for large datasets like big data and make a report on comparison among them. The main metric is time complexity to differentiate all algorithms.

Keywords: cluster, grid, density.

I. INTRODUCTION

Clustering is a procedure for assemblage similar data objects of a data set [1][18]. The main objective of this paper is to present a wide-ranging study of clustering approaches. Clustering divides data into groups which is a collection of related data items. Every group treated and called as cluster, which is a collection of objects that are related and dissimilar to objects of remaining clusters. Data mining deals with massive amount of databases that compel with clustering analysis. Clustering used in statistical analysis, geographical maps, decision-making, biology cell analysis machine-learning and in google maps. Under these limitations that clustering method is mainly suitable for the investigation of interrelationships with the data points. Different methods for clustering includes hierarchical methods [1], partitional approaches, density based approaches [5], grid based approaches [7][8] and model based techniques.

Hierarchical methods: Hierarchical methods [1][9] attain a separation of the objects which forms a tree like structure. First cluster is to be formed and then it will divide into small cluster. There are some other clustering methods like BIRCH, ROCK , and Chameleon [4].

A) Agglomerative method- It is a top down approach which initiates with a root node (data object) and further add data objects that will used for that cluster. This process will continue till the final stage where required cluster was obtained.

B) Divisive method: It is a bottom-up approach which initiates with a final cluster (data objects) and in further eliminate data objects that will not used for that cluster. This process will continues till the final stage where required cluster was obtained by doing eliminations.

II. CLUSTERING APPROACHES

1.1 Hierarchical method:

BIRCH & ROCK

Hierarchical clustering forms a cluster of hierarchy like tree which consists of objects, called a dendrogram[1]. Each and every node contains left and right clusters and siblings. This method allows us to explore data on unusual levels of granularity. BIRCH is used to achieve scalability. BIRCH [3] works based on the method called “clustering feature” to sum up a cluster data, and “clustering feature tree (CF-tree)” [4] to characterize a cluster by hierarchical fashion. Steps in BIRCH algorithm.

1: BIRCH scans the entire database to create CF-tree.
2: BIRCH applies clustering algorithms to cluster the left and right nodes of the CF-tree, and eliminates meager clusters, and also merges the intense clusters into big ones.

1.2 Partitional method: K-means, k-medoids.

The k-means [3] is widely using algorithm. Here the total numbers of clusters which are going to formed are initiated first by K [5], co it is called k-means. In the first step assign total number of data objects to k clusters ,and in the next step find a mean of each cluster objects and re assign data objects which are near to the mean.by doing this in each iteration new clusters are formed till there is no chance to form a new cluster.

Steps are

Input: Dataset D and k

Process:
Implementation of Clustering Algorithms for Real Time Large Datasets

- Select k-objects of dataset as cluster centers.
- Allocate each object to k-clusters.
- Find mean of all clusters and depending on mean re assign cluster objects.
- Repeat the process until no change.

**Output:** set of k-clusters.

### Fig 1: K-means clustering

#### 1.3 Grid-based method: STING, Wave Cluster, and CLIQUE

STING [18] deals with numerical data and aims to form a region oriented queries. STING builds data summaries. STING embedding spatial region of the data objects and it is further separated into rectangular cells. These rectangular grids represent different kinds of resolutions. They further make a hierarchical composition.

#### 1.4 Density-Based method:

DBSCAN, OPTICS and DENCLUE are density based methods [3][4]. Essential metrics of DBSCAN are density and connectivity. Both are calculated by means of local distribution related to nearest neighbors. DBSCAN [7] creates spatial data with minimal density.

### Fig 2: DBSCAN Algorithm

### III. COMPARISON OF CLUSTERING METHODS

All algorithms are compared by means of processing time. The following table explains about comparative study.

| Algorithm | Time Taken |
|-----------|------------|
| BIRCH     | Less       |
| ROCK      | High       |
| K-Means   | Less       |
| STING     | Less       |

### IV. RESULTS AND DISCUSSION

All above algorithms are analyzed by means of execution time for large and small dataset. The following graph gives results.

### Fig 3 Comparison of various clustering algorithms

### V. CONCLUSION

This paper presents a assessment of clustering approaches. All approaches applied for large and small data sets and comparative study was done. The dataset was taken from UCI repository. In Hierarchical method BIRCH algorithm, in Partitional method K-means algorithm, in Grid-based method STING, in Density-Based method OPTICS and in Model-based method CLARANS algorithms are compared and analysed. All of them are compared in terms of time complexity for large datasets in bigdata. BIRCH, K-means and COBEWB tooks less ime to compute and CLARANS time complexity is high.

### REFERENCES

1. Yingjie Tian, Dongkuan Xu. "A Comprehensive Survey of Clustering Algorithms", Published online: 12 August 2015 © Springer-Verlag Berlin Heidelberg 2015.
2. “Gupta, G. K. “Introduction to data mining with case studies” PHI Learning Pvt. Ltd”, 2006.
3. Dunham, Margaret H. “Data mining: Introductory and advanced topics” Pearson Education India, 2006.
4. Sabha Firdaus, Md. Ashraf Uddin , “A Survey on Clustering Algorithms and Complexity Analysis”, IICSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.
5. Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Genetic algorithm-based clustering technique." Pattern recognition 33.9 (2000): 1455-1465.
6. Murtagh, Fionn (2002), "Clustering in massive data sets", in Abello, James M.; Pardalos, Panos M.; Resende, Mauricio G. C., “Handbook of massive data sets”, Massive Computing 4, Springer, pp. 513 – 516, ISBN 978-1-4020-0489-6.
7. Murtagh, Fionn (1983), “A survey of recent advances in hierarchical clustering algorithms”, The Computer Journal 26 (4): 354–359.doi:10.1093/comjnl/26.4.354.
8. [http://en.wikipedia.org/wiki/Nearest_neighbor_chain_algorithm](http://en.wikipedia.org/wiki/Nearest_neighbor_chain_algorithm).
9. Zhang T, Ramakrishnan R, LivnyM(1996) “BIRCH: an efficient data clustering method for very large Databases” ACM SIGMOD Rec 25:103–104
10. https://archive.ics.uci.edu/ml/datasets/mushroom.