A STUDY AND ANALYSIS OF CLUSTERING TECHNIQUES FOR BIG DATA ANALYSIS

[1]Dukitha.M, [2]Sangeetha .K, [3]Chaya.M,

[1]Asst.Professor,[1,2]III MCA,

[1,2,3] Er.Perumal Manimekalai College Of Engineering

ABSTRACT:

With the beginning of new period data has grown rapidly not only in size but also in variety. There is a complexity in analyzing such big data. Data mining is the technique in which useful information and unseen relationship among data is extracted. Clustering is one of the most important techniques used for data mining in which mining is performed by result out clusters having similar group of data. Broad analysis of these techniques is agreed out and proper clustering algorithm is provided. This paper focuses on a intense study of unlike clustering algorithms highlighting the characteristics of big data.

Keywords: clustering techniques- Partitioning, Density, Grid Based, Model Based, Hierarchical.

INTRODUCTION

Big Data is also **data** but with a **enormous size**. Big Data is a term used to describe a collection of data that is huge in size and yet increasing exponentially with time. The data is so large and difficult that none of the traditional data management tools are able to store it or process it professionally. Big Data are about revolving unstructured, invaluable, imperfect, complex data into usable information. But, it becomes difficult to maintain huge volume of information and data day to day from many dissimilar resources and services which were not available to human space just a few decades back. Very huge quantities of data are produced every day by and about people, things, and their communications. Clustering is the process of grouping the data based on their related properties. The main aim of this paper is to provide different clustering algorithms for Big Data.

CLUSTERING TECHNIQUES:

There are 5 clustering techniques.

1.partitioning based clustering

2.density based clustering

3.grid based cludtering

4.model based clustering

5.hierarchical based clustering

1. Partitioning Method

The partitioning based method divides data objects into a number of partitions (clusters). In this method, data objects are divided into non-overlapping subsets (clusters) such that all data objects into same clusters are closer to center mean values. In this method, all clusters are determined promptly. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. In this method convergence is local and the globally optimal solution cannot be guaranteed. It obtain a single level partition of objects these method are usually based on heuristic creating local optimum solution. Each cluster has atleast one object and each object belongs to only one cluster. Methods like k-mean, PAM (Partitioning Around Medoids), CLARA (Clustering LARGE Applications) and the

Probabilistic Clustering are comes under partitioning clustering. Partitional clustering is considered to be the mainly popular this method is efficient and early adapted for large database .this algorithm uses for large database,this algorithm uses iterative refinement method(hile climbig or greevy method)they converted to a local minimum rather than global minimum

1)number of cluster apriori

2)user to specify starting state of the cluster

class of clustering algorithm also known as iterative replacement algorithm.

This method is efficient and early adapted for large database. This algorithm uses iterative refinement method(hile climbig or greavy method)there are two important method.

- k-means algorithm method
- expectation and maximization method

i.k-means algorithm

K-means algorithm is classical clustering method which is easy to implement. the data about all the object is located in the main memory. *K*-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to locate groups in the data, with the number of groups represented by the variable *K*. This procedure follow a simple and easy way to classify a given data set through a certain The number of clusters (assume k clusters) fixed a priori. The main suggestion is to define k centroids, one for each cluster. as much a7s possible far gone from each other. These centroids shoud be located in a craftiness way because of different location causes different result. So, the improved choice is to place them.

- The centroids of the *K* clusters, which can be used to label new data
- Labels for the preparation data (each data point is assign to a particular cluster)

K-means is a simple algorithm that has been modified to many difficulty domains. As we are going to see, it is a good applicant for extension to work with unclear feature vectors.The method is called k-means because each of the k-cluster is represented by the mean of the object [called centriod with in it

It is also called as centriod method at each step centriod point is assume to be known and each of the remaining points are allocated to the cluster whose centriod is closed to it.once the allocation completed the centriod of the cluster are recomputed by using simple mean and the process is repeated untill there is no change in the cluster. the k-means algorithm uses Euclidean, manhatten distance measure with compact cluster.

ii.exceptation and maximization method

K-mean method does not explicity assume any probability distribution for the attribute value consist similar object. A common task in signal processing is the estimation of the parameters of a probability distribution function. Perhaps the most frequently encountered estimation problem is the estimation of the mean of a signal in noise in maximization and expectation method objects in the dataset have attribute whose value are distributed according to some unknown linear combination of simple probability distribution.

k-means method is used to minimize with in group variation were as EM method is an attempt to maximize expectation of assignment.

EM method consist two iteration

- E STEP
- M STEP
- ***E-STEP***

It invovle estimating the probability distribution of the cluster given data.

- ***M- STEP:***

Involve finding the model parameters that maximize the likely would of the solution.Em method assumes all the attribute are indepedent random variable in a simple case of just two cluster .it object having only one single attribute ,we may assume that the distributed value vary according to normal distribution.

1) mean and standard deviation of the normal distribution for cluster one

2) the mean and standard deviation of the normal distribution for cluster 2.

3) the probability p of a sample belonging to cluster 1 and therefore probability of belonging to cluster 2

2.Density Based Method

The density based method are based on assumption that clusters are high density collection of data. that are separated by low density of data. therefore the basis call density base method is for each data point in a cluster atleast minimum number of points exit within the given distance. there are two importent parameter are used

1 .R (size of the neighbor wood)

2 .N (The minimum points in the neighbor wood)

These two parameters determine the density within the cluster and also determine which object is outlier(or)noise. (The number of concept required for density base clustering)

- Neighbor hood
- Core object
- Proximity
- Connectivity

For these methods a "neighborhood" has to be defined and the density must be calculated according to number of substance in the neighborhood.

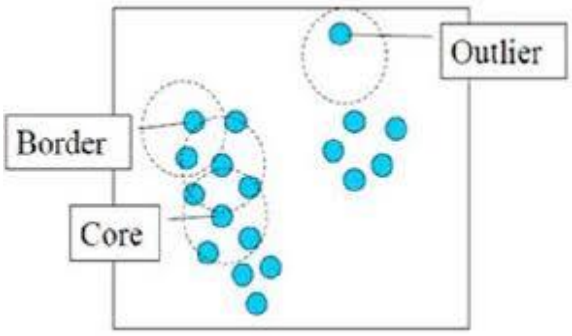

Figure 1.density based cluster

3.Grid Based Method

In this technique measures the object space into a limited number of cells that form a grid organization on which all of the operations for clustering are performed. the object space rather than the data is divided into grid .this method is based on characteristics of data and can deal with non-numeric data. It is based on clustering leaning query answering in multilevel grid structures.

it will generate a minimal description of each cluster. Unlike other clustering methods, Wave Cluster does not require users to give the number clusters applicable to low dimensional space. Grid based methods help in expressing the data at varied level of detail based on all the attributes that have been selected as dimensional attributes.

4.Model Based

Here a model based on probability distribution. the algorithm try to build clusters with a high level of similarity within them and low level of similarity between them based on mean value this algorithm minimized the error function. they optimize the fit among the data and some mathematical model.

Ex: EM (Exceptation and maximization), SOM(Self organizing feature map)

in this model is theorized for each group to locate the best shape of data for a given model.

5.Hierarchical Method

It obtain a nested partition of object resulting in a tree of clusterThese method either start with one cluster and split into the small and small cluster start with each object individual cluster and try to merge large cluster. Hierarchical clustering algorithms have tended to be some what more prominent than others, perhaps because they presuppose very little in the way of data character-istics or of a priori knowledge on the part of the analyst. This method create a hierarchical decomposition of the agreed set of data objects. The tree of clusters so created name as dendrograms.

Every cluster node contains child clusters, sibling clusters partition the points enclosed by their ordinary parent. In hierarchical clustering assign each item to a cluster such that N items then we have N clusters. Find next pair of clusters and merge them into single cluster. Compute distance between new cluster and each of previous clusters. It uses a number of greedy heuristic schemes of iterative optimization. The algorithms to be discuss in this article focus instead on closely what is necessary in order to carry out an agglomeration at any stage of the clustering: this is usually little more than the nearest neighbor points of specified points. hierarchical cluster is classified as

- Agglomeritive approach
- Divisive approach

i.Agglometitive Approach[Bottom Up Approach]

Aggolomeritive approach is a bottom up approach. each object at the start is a cluster by itself and the nearby cluster are rapidly merged. Object are merged into a single large cluster. untill all object are in a single cluster or confident termination condition is satisfied. The single cluster becomes the hierarchi's root. It successively merges the groups that are close to one another, until all the data objects are in same cluster. it finds the two clusters that are closest to each other, and combines the two to forms one cluster.

ii.Divisive Approach[Top Down Approach]

Divisive approach is a top down approach. A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. All object are put it in a single cluster. Than rapidly perform spilting of cluster .and resulting smaller and smaller cluster. untill stopping criteria is met. then successively splits resulting clusters until only clusters of individual objects remember. there are two types:

- Monothetic
- polythetic

Hierarchical cluster based on different distance measures are used

- **Single link algorithm**
- **Complete link algorithm**
- **Centroid link algorithm**
- **Average link algorithm**
- **Ward's minimum variance algorithm**

A.SINGLE LINK ALGORITHM

It determains the distance between two cluster has the minimum of the distance between all pair of points nearest neighbor.

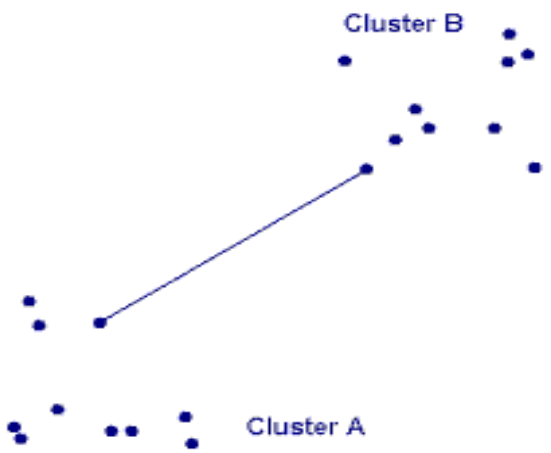

Figure2:Single linkage cluster

the dissimilarity between 2 clusters is the least variation between members of the two clusters.This method produces long chains which form loose, untidy clusters.

B.COMPLETE LINK ALGORITHM

the distance between two cluster is define has a maximum of the pairwise distance. Therefore must be computed the largest choosen. Farthest neighbor.

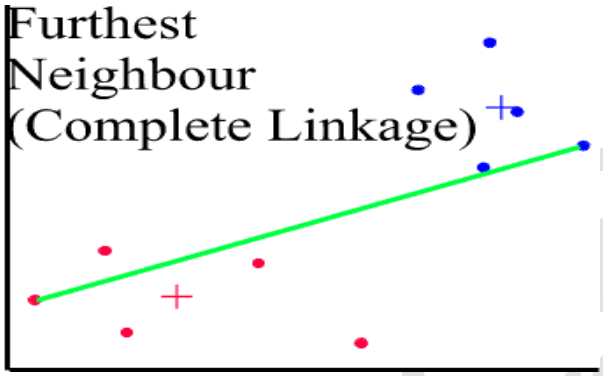

Figure 3:complete linkage cluster

This variation uses the group centroid as the average. The centroid is defined as the center of a cloud of points.

C.CENTROID ALGORITHM

In this algorithm distance between two cluster is determines has the distance between centriod of the cluster. it computed the distance between two cluster has the distance between the average point of the two cluster

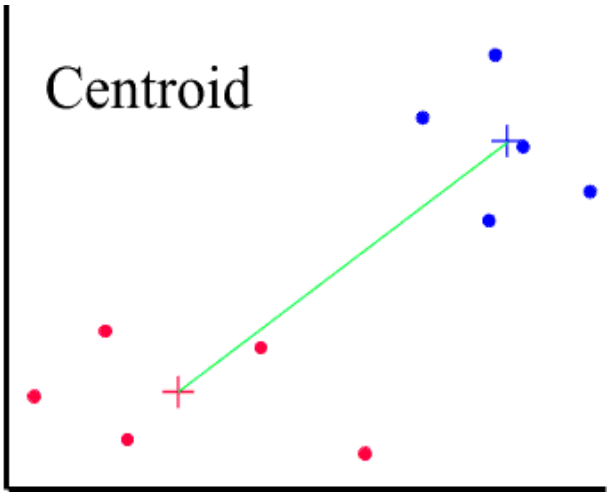

Figure 4:Centriod clustering

And we use the Euclidian distance to complex the centriod. (maximization of method)The centroid method uses the controid (center of the group of cases) to determine the average distance between clusters of cases. greatest similarity between two member of cluster.

D.*AVERAGE LINK ALGORITHM*

It compute the distance between two cluster has the average of all pairwise distance between the object from one cluster to another cluster .that is if there are m element in one cluster n .

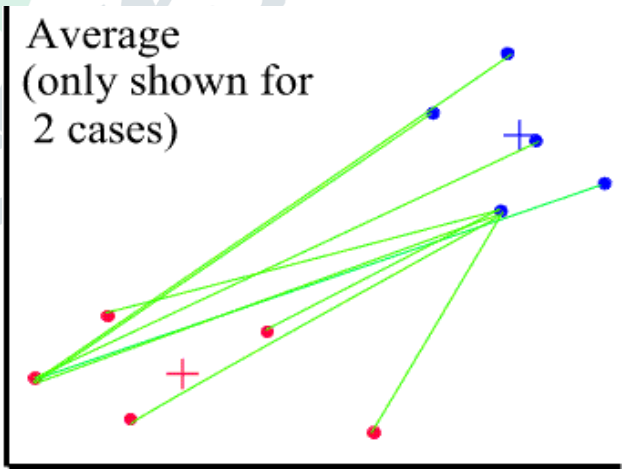

figure5:Average linkage clustering

In the other there are mn distance to be completed. added and divided by the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster

E.WARD'S MININMUM VARIANCE METHOD

Ward's mininmum is the differs between the total within the cluster some of the square for the two cluster separately and within the cluster some of squares resulting from merging the cluster.

Dn(A,B)=NaNbDc(A,B)/(Na+Nb)

Cluster membership is assigned by calculating the total sum of squares deviation from the mean of The principle for union is that it should create the smallest possible add to in the error sum of squares.

Conclusion:

This paper analyzed special clustering algorithms required for processing Big Data. The study discovered that to identify the outliers in large data sets, To perform clustering, various algorithms can be used but to get proper results. essential Big Data Clustering Techniques have been explored of which Map Reduce is of chief importance to my constant research on Big Data Analytics.

References

1. Yasodha P, Ananathanarayanan NR. Analyzing Big Data to build knowledge based system for early detection of ovarian cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.

2. Pandove D, Goel S. A comprehensive study on clustering approaches for Big Data mining. IEEE Transactions on Electronics and Communication System; Coimbatore. 2015 Feb 26-27. p. 1333–8.

3. Park H, Park J, Kwon YB. Topic clustering from selected area papers. Indian Journal of Science and Technology. 2015 Oct; 8(26):1–7.

4. Abbasi A, Younis M. A survey on clustering algorithms for wireless sensor networks. Computer Communications. 2007 Dec; 30(14-15):2826–41.

5.Yadav C, Wang S, Kumar M. Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5.