COMPARISON OF CLUSTERING TECHNIQUES FOR BIG DATA

[1]Apoorva Verma, [2] Dr. Preeti Tiwari
[1]Student[MCA], [2] Associate Professor
International School of Informatics and Management

Abstract

Big Data in general terms means lots and lots of data coming from many sources, some of them disparate, most of them unstructured all of them contain valuable insights. Several approaches have been developed like or are in the development to harness the implied power in this data and one of them is cluster analysis. Clustering is one of the most fundamental techniques used in data mining. Clustering helps to visually analyze the data and also assist in decision make. It is widely used in variety of application like marketing, insurance, surveillance, fraud detection and scientific discovery to extract useful information. In this paper we have discussed various clustering techniques used in analysis of big data and gave a comparison among them.

Keywords: Hierarchical Based Clustering, Agglomerative Algorithm, Divisive Algorithm, outliers.

I Introduction

Cluster Analysis or Clustering is the task of grouping set of objects in such a way that objects in the same group are more similar to each other than to those in the group. [1] Cluster Analysis itself is not one specific algorithm but the general task to be solved. Data Mining is a process designed to explore large amount of data also known as “Big Data”.

We live on-demand, on-command Digital universe with data prolife ring by Institutions, Individuals and Machines at a very high rate.
We are living in the era of data deluge and as a result the term “Big Data” is appearing is many contexts from meteorological, genomics, complex physics simulations, biological and environmental research, finance and environmental research, finance and business to health care.

In this paper we have tried to give the most popular Big Data’s clustering techniques. Most of the papers focus on single technique for Big Data but here in this paper the goal is to make a broad and general synthesis concerning the Big Data Clustering Techniques. Clustering algorithms have emerged as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications. In particular, their main goal is to categorize data into clusters such that objects are grouped in the same cluster when they are similar according to specific metrics.

Section II of this paper specifies Types of Clustering Techniques

Section III of this paper specifies Criteria for using clustering techniques for big data

Section IV is Conclusion.

II Types of Clustering Techniques

A Hierarchical Based Clustering

Hierarchical based clustering is also known as connectivity based clustering. It is a method in which hierarchies of clusters are constructed. It consists of two methods: -
i)Agglomerative Algorithm: - It is a bottom up approach. It starts by merging each object which are closer to each other or by merging a smaller no of clusters into a larger cluster until all the objects are merged and a termination condition is met. These methods will produce a hierarchy from which the user still needs to choose the appropriate clusters. They are not very robust towards outliers which will either show up as additional clusters or

even cause other clusters to merge (known as SLINK-Single Linkage Clustering).
ii)Divisive Algorithm: - It is a top down approach. It starts by splitting a larger cluster into smaller clusters until there remain only clusters of one data object and the termination condition is met.

B Partition Based Clustering
In such algorithms, all clusters are determined promptly. Initial groups are specified and reallocated towards a union. In other words, the partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster.
These clusters should fulfill the following requirements:
(1) each group must contain at least one object.
(2) each object must belong to exactly one group in the K-means algorithm for instance, a center is the average of all points and coordinates representing the arithmetic mean. In the K-medoids algorithm, objects which are near the center represent the clusters. There are many other partitioning algorithms such as K-modes, PAM, CLARA, CLARANS and FCM.

C Density Based Clustering
In this type of clustering, the data objects are separated based on their connectivity, boundary or their region which plays a vital role in finding non-linear shape structure based on the density. This type of clustering helps to separate low dense region (noise data) from high dense region of clusters. The most popular method is DBSCAN. It features a well-defined cluster model called "density-reachability". OPTICS, DBCLASD and DENCLUE are other algorithms which are used such to filter out noise (outliers) and discover clusters of arbitrary shape.

D Grid Based Clustering
The space of the data objects is divided into grids. The main advantage of this approach is its fast processing time, because it goes through the dataset once to compute the statistical values for the grids. The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. The performance of a grid-based method depends on the size of the grid, which is usually much less than the size of the database. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain the required clustering quality or full the time requirement. Wave-Cluster and STING are typical examples of this category of this approach is its fast processing time, because it goes through the dataset once to compute the statistical values for the grids. The accumulated grid-data make grid-based
clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. The performance of a grid-based method depends on the size of the grid, which is usually much less than the size of the database. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain the required clustering quality or full the time requirement. Wave Cluster and STING are typical examples of this category.

III Clustering techniques selection criteria for Big Data

When evaluating clustering methods for big data, specific criteria need to be used to evaluate the relative strengths and weaknesses of every algorithm with respect to the three-dimensional properties of big data, including *Volume*, *Velocity*, and *Variety*.

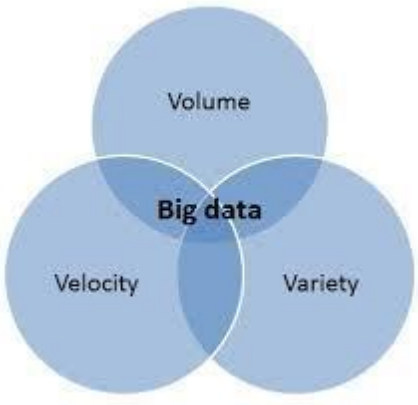

Figure1: 3V's of Big Data [8]

Volume: -Volume refers to the incredible amounts of data generated each second from various sources. To select a suitable clustering algorithm with respect to the *Volume* property, the following criteria are considered: (i) size of the dataset, (ii)

handling high dimensionality and (iii) handling outliers/ noisy data.

Variety: - Variety is defined as the different types of data we can use. To guide the selection of a suitable clustering algorithm with respect to the *Variety* property, the following criteria are considered: (i) type of dataset and (ii) clusters shape.

Velocity: - Velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. To guide the selection of a suitable clustering algorithm with respect to the *Velocity* property, the following criteria are considered: (i) complexity of algorithm and (ii) the run time performance.

Earlier big data was big task for the data analytics as the question was how to analyze such a huge amount of data how to filter it and dig the data which was of their use. But with the help of the above mentioned clustering techniques the task became much more simple convenient and time saving. Thus by understanding the various parameters and criteria and other details about data we can easily compare and decide as to which clustering technique will be best for various types of data. Given below table is a comparative study which shows as to which clustering technique should be used for big data.

Table1: Comparison between various clustering techniques for big data.

Clustering techniques	Algorithm name	volume	variety	Velocity	
		Data set	High dimensionality	Avoid outliers	Data set size hierarchical
hierarchical	WARDS	Small	No	No	Numerical
partial	K-MEANS	Large	No	No	Numerical
	K-Mediods	Small	Yes	Yes	Categorical
	CLARA	Large	No	No	Numerical
	CLARANS	Large	No	No	Numerical
Density based	DBSCAN	Large	No	No	Numerical
	OPTICS	Large	No	No	Numerical
Grid based	STIRR	Large	No	No	Numerical

IV Conclusion

This paper concludes that various clustering techniques can be used for big data depending upon the data variety, volume and velocity. Considering various parameters and criteria one can decide which clustering should be used on different type of data. Clustering techniques have not only made big data analytics easy but also reduced lot of time and effort. All these recent techniques are compared on the basis of execution time and cluster quality and their merits and demerits are provided. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, and pattern recognition. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often proximity according to some defined distance measure. Data clustering algorithms can be hierarchical or partitioned. Density based clustering is designed for building clusters of arbitrary shapes. It builds clusters automatically i.e. no need to mention the number of clusters and naturally removes outliers. Grid based clustering mainly concentrates on spatial data. Thus this paper concludes that depending on various criteria, different techniques for clustering can be used for big data.

V References

[1] Olga Kurasan, Vikiton Medvedev, Paval Stefano Vic, “Strategies for Big Data Clustering” in IEEE in 2014.

[2] Miss Harshada, S Deshmukh, Prof P.L Ranteka, “Cluster Analysis for Big Data” in IJARCET in 2015.

[3] Uranus Kazeml, “Clustering Methods in Big Data” in mat journals in 2017.

[4]Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Zebti FouFou, Abdelaziz Bouras, “A Survey of clustering algorithms for big data:a taxonomy and empirical analysis” in IEEE in 2014.

[5] Keshav Sanse, Meena Sharma, “Clustering methods for Big data analysis” in IJARCET in 2015.

[6] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta,”A Comparative Study of Various Clustering Algorithms in Data Mining,” International Journal of Engineering Research and Applications (IJERA),in 2012.

[7] Arockiam, L., S.S. Baskar, and L. Jeyasimman. 2012. Clustering Techniques in Data Mining.

[8] https://bigdataldn.com/big-data-the-3-vs-explained/