Clustering In Big Data Environment Using Spark

Varsha Bansal[1], Mr. Kamal Kumar[2]
[1]M.Tech. Scholar, [2]Associate Professor,
[1]M.Tech. Computer Science & Engineering
[1]Ganga Institute of Technology and Management Kablana, Jhajjar, Haryana, India

Abstract: Clustering deals with finding a structure in a collection of unlabeled data. Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. As Big Data is referring to terabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. Clustering helps to visually analyze the data and also assists in decision making. In this paper we have discussed some big data mining clustering techniques and also provides a comparison among them.

Index Terms- Clustering, Data Mining, Clustering Techniques, Big Data**.**

I. INTRODUCTION

Clustering is an important unsupervised learning problem. It enable us to find a structure in a collection of unlabeled data. A simple definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is a collection of objects which are alike between them and are different to the objects belonging to other clusters[1]. Data Stream Mining is the process to extract knowledge structures from continuous, rapid data records.

In data stream mining applications, the goal is to assume the class of new instances in the data stream by giving some knowledge about the class membership or values of previous instances. Machine learning techniques be able to learn this prediction task from labeled examples in an automated fashion. Data stream mining can be measured a subfield of data mining, machine learning and knowledge Discovery. Big Data means a collection of large data that have volume, variety and velocity and require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. The amount of data has been increasing at a faster rate. Since the data has been increasing at a faster rate, it is of great challenge in the direction of manage the data.

I. APPLICATIONS OF CLUSTER ANALYSIS

- In Image processing, Market research, pattern recognition, data analysis, Clustering is used.
- It help marketers discover distinct groups in their customer base. And their customer groups could be characterize based on the purchasing patterns.
- To derive plant and animal taxonomies, categorize genes with similar functionalities and approaching into structures inherent to populations, in the field of biology, clustering is used.
- Clustering also helps in recognition of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- On the trap for information discovery, It also helps in classifying documents.
- Clustering is used in outlier detection applications such as detection of credit card fraud.
- Data Mining As a function, cluster analysis acts as a tool to gain insight into the distribution of data in order to follow the characteristics of each cluster [2].

II. ISSUES AND CHALLENGES

Algorithms suffer for handling hard clustering work without supervision. For example, there is no assumption about the number of clusters in the data stream but in most cases this parameter should be determined.

- The number of expected partitions for the algorithm or the requisite density of the cluster requires a specialist associate. They need to know any repeated pattern again, the quality of clustering is the most important problems of the density and separation of data.
- Accuracy in terms of detecting concept drift.
- Efficiency in terms of speed is a vital problem in data mining clustering.
- Most of the Previous approaches lack precision in detecting outliers.
- **Uncertainity of data:** In many applications we do not have enough data for statistical operations, therefore new methods are needed to manage the precarious data stream in accurate and fast fashion. In many applications (i.e. network monitoring), arbitrary size causes some difficulties to realize accurate groups of data.
- **Data type:** Data stream processing should consider different data types (i.e., a mixture of clear, sequential and different data types).

- **Cluster Validity:** Developments in data stream clustering have finely tuned the need for determining suitable criteria to validate results. Most outcomes of methods are depended to specific application. Employing suitable criteria in deriving results is one of the most important challenges in this field..
- **Space limitation:** Space complexity in some applications (e.g. wireless sensor network monitoring and controlling) can be caused difficulties in processing along with time and concept drift. Small memory sensors are unable to retain large quantities of data, so a new way of data stream clustering should be managed for this limitation.
- **High dimensional data stream:** High dimensional data sets (e.g Image Processing, Personal Equality, Client Preferences Clustering, Network Intrusion Detection, Wireless Sensor Network and Generally Time Series Data) which should be managed through the processing of the data stream. In a large database, data complexity can be increased by the number of dimensions.

III. REQUIREMENTS OF CLUSTERING

clustering is required in data mining due to these following points –

- **Scalability:** we need a highly scalable clustering algorithm to deal with large databases.
- **Ability to deal with different attributes:** On any kind of data such as categorical, interval-based (numerical) data, and binary data these algorithms should be capable to be applied.
- **Attribute shape discovery of clusters:** The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be restricted to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality:** The clustering algorithm should be able to handle low-dimensional data along with the high dimensional space.
- **To handle noisy data:** Some algorithms are sensitive to noisy, missing or erroneous data and may lead to poor quality clusters.
- **Interpretability of final result:** Clustering results should be interpreted, easy, and usable [3].

IV. TECHNIQUES OF CLUSTERING

Hierarchical clustering: Also known as connectivity based clustering. It is based on the idea of objects being more related to nearby objects than to objects farther away. Hierarchical clustering algorithms connect objects in clusters on the basis of their distance. A cluster can be described by the maximum distance needed to connect parts of the cluster. At different distances, individual clusters will create connectivity-based clustering, which is a family of methods, which is different from distance calculations. This distance is based on the choice of works it could have: -

a) Agglomerative (starting with single elements and aggregating them into clusters).

b) Divisive (starting with the complete data set and dividing it into partitions).

Hierarchical clustering techniques use a variety of criteria to decide at each step which clusters should be joined as well as where the cluster should be partitioned into different clusters. This cluster is based on the measure of proximity: There are three cluster proximity measures: single-link, average link and complete link.

Single link: The distance between two clusters should be the smallest distance between two points such that one point is in each cluster.

Complete link: The distance between two clusters should be the largest distance between two points such that one point is in each cluster.

Average link: The distance between two clusters should be an average distance between two points such that one point is in each cluster.

Partitional clustering: These algorithms separate the data points into number of different partitions. These partitions are referred as clusters. The partitional clustering organizes data into single partition instead of representing data into nested structure like hierarchical clustering. Partitional clustering is more useful for large data set in which it is difficult to represent data in tree structure. Partial clustering could be classified as class error clustering, graph theoretical clustering, solving clustering and demand for clustering[8].

Centroid-based clustering: In this clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. K-means clustering gives a formal definition as an adaptation problem. Find cluster centers and assign objects to the nearest cluster center, such as the square distance from the cluster has been reduced. To specify most of these algorithms in advance, the number of cluster K is required, which is considered to be the biggest drawbacks of these algorithms. Also, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. There are many interesting theoretical properties in K-means: - This data space is known as a structure (Voronio diagram). This concept is close to the closest neighbor classification. This can be seen as variation of model-based classification.

Distribution-based clustering: Distribution based clustering model is very closely related to the data. Clusters can easily be easily defined as objects with the same distribution potential. This model of clustering works, such as artificial data sets generated by sampling random objects from a distribution**.** It suffers from the problem of over fitting, unless constraints are put on the model complexity. A more complex model will generally be able to explain the data, which chooses to make the appropriate model complexity naturally difficult. This creates complex models for clustering clusters that can achieve correlation and reliability between properties. However, these algorithms put an additional load on the user: for many real data sets, there may not be a briefly defined mathematical model.

Density-based clustering: In density-based clustering, groups are defined as high density areas compared to the remainder of the data set. Objects in these rare regions - which are essential for different groups - are generally considered to be noise and boundary points. Density-based clustering algorithms find clusters based on the density of data points in an area. The key idea is that each instance of a cluster the neighborhood of a given radius has to contain at least a minimum number of objects i.e. The priority of the neighborhood should be greater than the given threshold. This division is completely different from the algorithm, which uses repeated repositing of points given to a certain number of groups. One of the best density-based clustering algorithms is the DBSCAN.

Grid-Based Clustering: In this approach of clustering approaches, first object space is divided into a limited number of cells which create a grid structure on which all the functions are used for clustering. STING examines the statistical information stored in CLIQUE grid cells. There are usually many levels of this type of rectangular cells that correspond to different levels of resolution, and these cells form a structured structure: At high levels, each cell is divided into several cells at the next lower level. Statistical information about the characteristics in each grid cell is pre-calculated and stored. The goal of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not move the points, but they create several hierarchical levels of the group of objects.

Model-Based Clustering: These algorithms find good approximations of model parameters that best fit the data. They can either be partial or hierarchical, depending on the structure or model they think of the data set and the way in which they refine this model to identify the partition. They are close to density-based algorithms, so they develop special groups to improve the pre-connection model. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

Categorical Data Clustering: These algorithms have been specifically developed for data where Euclidean, or other numerical oriented, distance measures can not be applied. In literature, we gain close access to both partial and hierarchical approaches.

V. TECHNOLOGIES

Clustering is a problem of great practical importance that has been the focus of substantial research in several domains for decades. As storage capacities grow, we have at hand larger amounts of data available for analysis and mining [4]. We can identify two main groups of techniques for mining of huge data base. A group refers to streaming data and implements mining techniques whereas the other group tries to solve this problem directly with efficient algorithms. Recently, many researchers have focused on the data stream on data as a skilled strategy against huge data base mining rather than mining on the entire data base. The main problem in data stream mining means that it is more difficult to develop data to detect such techniques, therefore unchanged methods should be implemented[5]. However, clustering techniques can inspire us to get hidden information. In this survey, we try to clarify: First, the various problem definitions related to data stream clustering in general; Second, the specific difficulties in this area of research; third, the varying assumptions, heuristics, and intuitions forming the basis of different approaches; and how several well-known solutions tackle different problems. The cluster analysis on evolving data stream becomes more difficult, because the data objects in the stream must be accessed in order and can read only once or a small number of times with limited resources. The proposed technique has small memory footprints and also provide empirical evidence of the algorithm's performance on real datasets and synthetic data streams. These are Operational Big Data-No SQL, Analytical Big Data-Map Reduce, Server, Storage and Processing.

VI. CONCLUSION

Existing algorithms are insufficient to face all the challenges raised by the Big Data. There really is no clustering algorithm that can be used to solve all the big data issues. The complexity of implementing these algorithms in terms of time and memory remains a major challenge. However, the mining work is complex due to the specific characteristics of the data stream; It is heavy, even potentially infinite, and besides, continuous, a scan is required, and there is a dynamic change during the time, thus in the real-time usually requires a strong reaction. The data Stream Clustering approach is one of the data mining techniques that can extract knowledge from such data. We provide empirical analysis on the performance of the algorithm in clustering both synthetic and real data streams. Conventional clustering methods are not flexible enough to tackle evolving data. Therefore, in recent years, the demand for efficient data clustering algorithms has led to the publication of several methods. . In practice, each algorithm can be useful based on its applications and properties. So we will Implement technologies to perform clustering using Big Data and compare with existing in terms of certain parameters.

VII. ACKNOWLEDGEMENT

I would like to express my appreciation to Mr. Kamal Kumar (Astt. Professor), for his guidance and support. Without his valuable assistance, this work would not have been completed .I am also thankful to Mrs. Neetu Sharma, the Head of Department Of Computer Science in Ganga Institute Of Technology And Management for her valuable cooperation and assistance.

VIII. REFERENCES

[1] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
[2] http://www.srmuniv.ac.in/sites/default/files/2017/15CS331E-unitIV.pdf
[3] http://www.ijsrd.com/articles/IJSRDV3I70331.pdf
[4] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_99
[5] http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp566-569.pdf
[6] http://shodhganga.inflibnet.ac.in/bitstream/10603/28762/8/08_chapter%202
[7]http://engr.smu.edu/~mhd/dmbook/part2.ppt.
[8] Mu-Yu Lu, SJSU Database System Concepts, Silberschatz, Korth, Sudarshan
[9] Garima Sehgal, Dr. Kanwal Garg "Comparison of Various Clustering Algorithms" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3074-3076
[10] Osama Abu Abbas "Comparisons between data clustering algorithms"
[11] Preeti Baser, Dr. Jatinderkumar R. Saini "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets".
[12] Shafiq Alam, Gillian Dobbie, Patricia Riddle, M. Asif Naeem, "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering",2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 64-68.
[13] B. Rama et. Al., "A Survey on clustering Current Status and challenging issues"(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 9, pp. 29762980, 2010
[14] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.
[15] Stefan Brecheisen, Hans-Peter Kriegel, and Martin Pfeifleisen , Multi-Step DensityBased Clustering , Knowledge and information system (KAIS), Vol. 9 , No. 3 , 2006.
[16] S.Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama," A survey on partition clustering algorithms", International Journal of Enterprise Computing and Business SystemInternational Systems, vol. 1, pp. 1-13, 2011.