Research Study of Big Data Clustering Techniques

S.Mahalakshmi#1, C.SaiAshwini*2 , Meghana S*3
#Assistant Professor, *Student
Dept. of ISE, BMSIT, Bangalore.
1maha.shanmugam@gmail.com
2csaiashwini@ymail.com megsarod@gmail.com3

***Abstract*— Big data refers to a heterogeneous collection of data that is very large, dynamic and complex. Most of the produced data are unstructured and traditional database management tools are unable to handle this type of information. Numerous challenges are in place with big data like storage, transition, visualization, searching, analysis, security and privacy violations and sharing. The new conditions imposed by Big Data impose serious challenges at different level, including data clustering. This paper aims to review and make a concise survey related to clustering techniques in Big Data context with challenges.**

***Keywords*—Big data; dynamic; complex; visualization; clustering**

I. INTRODUCTION

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using database management tools .The exponential growth of data in all fields demands the revolutionary measures required for managing and accessing such data. Big data stores and handles data in different way from traditional data warehouses. Big data comprises massive sensor data, raw and semi-structured log data of IT industries and the exploded quantity of data from social media. Examples of this data include high-volume sensor data and social networking information from web sites such as Google, Face Book, LinkedIn, Yahoo, Amazon and Twitter. Big Data appear in different areas such as health (enhancing the efficiency of some treatments), biomedical, marketing (increasing sales), transportation (reducing costs), business, finance (minimizing risks), management (decision making with high efficiency and speed), social, media, and government services. Big data need big storage and this volume makes operations such as analytical operations, process operations, retrieval operations, very difficult and time consuming. One way to overcome these difficult problems is to have big data clustered in a compact format.Clustering is the task of grouping input data into subsets called clusters.Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to different cluster. Clustering is an unsupervised technique used to classify large datasets in to correlative groups. No predefined class label exists for the data points or instances. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups and the groups are called as clusters.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 metric.

The standard process of clustering can be divided into the following several steps: (1) Feature extraction and selection: extract and select the most representative features from the original data set; (2) Clustering algorithm design: design the clustering algorithm according to the characteristics of the problem; (3) Result evaluation: evaluate the clustering result and judge the validity of algorithm; (4) Result explanation: give a practical explanation for the clustering result .Clustering task is a expensive as many of the algorithms require iterative or recursive procedures and most real life data is high dimensional. Such clustering techniques aim to produce a good quality of clusters. Therefore,they would hugely benefit everyone from ordinary users to researchers and people in the corporate world, as they could provide an efficient tool to deal with large data such as critical systems. Clustering is widely used in variety of applications like marketing, insurance, surveillance, fraud detection and scientific discovery to extract useful information .A good clustering method will produce high quality clusters with high intra-class similarity low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. Each clustering algorithm has its own strengths and weaknesses, due to the complexity of information. Big Data clustering techniques can be classified into two categories Single machine clustering techniques and Multiple-machine clustering techniques which includes Datamining clustering algorithms, dimension reduction techniques, parallel classification and the MapReduce framework.

II. BIG DATA CHALLENGES

The 5 V's of Big Data are:

1. **Volume** – The volume of big data is exploding exponentially day to day. The amount of data is at very large scale .The amount of information being collected is so huge that modern database management tools are unable to handle it and therefore become obsolete. The major challenge is how to determine relevance among the large volumes of data and how to create value from data that is relevant.
2. **Velocity** – Data is being produced at an exponential rate .It is growing continuously in terabytes and petabytes. Velocity includes both the challenge as to how fast data is being produced and how fast the data must be processed to meet demand.
3. **Variety** – Data is being created in different forms- unstructured, semi structured and structured data. This data is heterogeneous is nature. The major challenge is processing of big data to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application.
4.**Variability and Veracity -** It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set. The data being generated is uncertain in nature. It is hard to know which information is accurate and which is out of date. Veracity deals with uncertain or imprecise data.
5. **Value**- All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services.

Big Data is made of structured, semi-structured and unstructured information. Structured data are the basic data types such as integers, characters, and arrays of integers or characters. They are used in relational databases. The most common form of structured data or structured data records is a database where specific information is stored based on a methodology of columns and rows. Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers. Relational databases and spreadsheets are examples of structured data. Unstructured Data refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. It refers to any data that has no identifiable structure. Unstructured information is 90% of Big Data and is human information like emails, videos, tweets, Face book posts, call center conversations, closed circuit TV footage, website clicks. Unstructured data have no predefined format: email, books, journals, documents, videos, photos. Semi-structured data are a combination of two previous types of data, they are generally represented using XML. Most of produced data are unstructured and traditional database management tools are unable to handle this type of information.

To meet the growing demands of data, we should increase the capacity and performance of tools and methods. in order to manage large volume of data while keeping an acceptable resource needs, we have to improve clustering algorithms by reducing, their complexity in terms of time and memory. Big Data is an emerging trend and there is immediate need of data clustering techniques to analyze massive amount of data in near future.

III. CLUSTERING ALGORITHMS AND THEIR CHALLENGES WITH BIG DATA

Big Data clustering techniques can be classified into two categories single machine clustering techniques and multiple machine clustering techniques.

A. Single machine clustering techniques

***a.Partitioning based clustering*:** algorithm divides a data set in a single partition using a distance to classify points based on their similarities. These clusters should fulfill the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. Examples of this type of classification algorithms are K-means, k-mediods and FCM.

K-means Clustering:

1) Arbitrarily choose k objects from D dataset as the initial centroids
2) Repeat
3) Reassign each object to the clusters to which the object is the most similar, based on the mean value of the objects in clusters
4) Update the cluster means, that is, calculate the mean value of the objects for each cluster.
5) Until no change

K-medoids:

It varies from the k-means algorithm mainly in its representation of the different groups or clusters.

Algorithm-

1) Starts from an initial set of medoids and clusters are generated by which are close to respective medoids.
2) The algorithm iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.

Challenges:

i. Poor at handling noisy data and outliers.
ii. Works only on numeric data.
iii. Empty cluster generation problem.
iv. Random initial cluster center problem.
v. User has to provide the value of k

***b. Hierarchical based clustering*:** This method builds clusters in a hierarchical order , it forms nested clusters organised in a hierarchical tree. It forms clusters by recursively or iteratively partitioning the instances in either a top-down or bottom-up fashion. Hierarchical clustering method is of two types:

1. Agglomerative- this is a bottom up approach. In this approach initially each object is considered as a separate individual cluster. It then merges two or more suitable clusters to form new clusters. This merging of clusters is done recursively until a desired cluster structure or a stopping criterion (desired number of clusters k) is reached.

2. Divisive- this is top down approach. In this approach, initially the entire dataset is considered as one cluster. The cluster is then divided in to sub-clusters, which in turn are successively divided into more sub-clusters. This process is repeated until the stopping criterion (desired number of clusters k) is met.

Examples of this type of classification algorithms are BIRCH, CURE and Chameleon.

BIRCH (Balance Iterative Reducing Clustering) is the first clustering algorithm which removes the noisy data or outliers. This algorithm is also called as hybrid clustering. It overcomes the difficulties of hierarchical method. It makes full utilization of the memory and minimizes the I/O cost.

Challenges:

i. If an operation (merge or split) is performed, it cannot be undone i.e. no backtracking is possible.
ii. Inability to scale well.
iii. It is order-sensitive and may generate different clusters for different orders of the same input data.
iv. May not work well when clusters are not spherical.

***c. Density based clustering*:** Density-based clustering method is based on the concepts of density, connectivity and boundary. This method forms clusters based on the density of data points in a region and continue growing a given cluster as long as the density (number of objects or data points) in the neighborhood is exceeding some threshold.

DBSCAN: Density-based spatial clustering of applications with noise. This algorithm forms clusters using two parameters: - Eps ε: Maximum radius of the neighborhood - MinPts: Minimum number of points in an Eps neighborhood of that point.

Algorithm

1) Select a point p
2) If p is core point then
3) Retrieve and remove all points density-reachable from p with respect to Eps and MinPts;
4) Output them as a cluster.
5) Until all points have been processed

DENCLUE: Density-based Clustering analytically models the cluster distribution according to the sum of influence functions of all of the data points. The influence function can be seen as a function that describes the impact of a data point within its neighbourhood. Then density attractors can be identified as clusters. Density attractors are local maxima of the overall density function. In this algorithm, clusters of arbitrary shape can be easily described by a simple equation with kernel,and to separate the dense region into two half spaces. After each step of a multi-dimensional grid construction defined by the best cutting planes.

Challenges:

i. Unsuitable for high-dimensional datasets due to the curse of dimensionality phenomenon.
ii. Its quality depends upon the threshold set.

d. Grid based clustering: Complys with the three stages: firstly is to divide the space into rectangular cells to obtain a grid of cells of equal size, and then delete the low density of cells, and finally combine adjacent cells having a high density to form clusters. This method partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed.

STING (Statistical Information Grid based) explores statistical information stored in grid cells. Statistical information regarding the attributes in each grid cell is pre-computed and stored.

Wave Cluster does not require users to give the number of clusters applicable to low dimensional space. It uses a wavelet transformation to transform the original feature space resulting in a transformed space where the natural clusters in the data become distinguishable.

Challenge: Depends only on the number of cells in each dimension in the quantized space.

***e. Model based clustering*:** Model based clustering method optimizes the fit between the given data and some (predefined) mathematical model.

EM (Expected Maximization): EM finds the maximum-likelihood (ML) estimates for the parameters of the data model. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E steps.

Challenges:

i. The processing time is very slow in case of large data sets.
ii. Complex in nature.

***f. Dimension reduction*:** Its purpose is to select or extract optimal subset of relevant features for a criteria already fixed. The selection of this subset of features can eliminate irrelevant and redundant information according to the criterion used. This selection or extraction makes it possible to reduce the size of the sample space and makes it all more representative of the problem. For large sets of data, dimension reduction is usually performed before applying the classification algorithm to avoid the disadvantages of high dimensionality.

Challenges :

i. Don’t offer an efficient solution for high dimensional datasets.
ii. Should be performed before applying the classification algorithm.

B. Multiple machine clustering techniques

a. Parallel clustering: The processing of large amounts of data imposes a parallel computing to achieve results in reasonable time. the parallel classification divides the data partitions that will be distributed on different machines. This makes an individual classification to speed up the calculation and increases scalability. The parallel clustering methods includes Parallel k-means, Parallel Fuzzy c-means,

Challenges: Complexity of Implementing the algorithms can’t be done easily.

***b. MapReduce based clustering*:** MapReduce is a task partitioning mechanism for a distributed execution on a large number of servers. Principle is to decompose a task (the map part) into smaller tasks. The tasks are then dispatched to different servers, and the results are collected and consolidated (the reduce part). In step Map the input data is analyzed, cut into sub problems and delegated to other nodes (which can do the same recursively). This will be processed later using the Map function which has a pair (key, value) that associates a set of new pairs (key, value). Then comes the stage Reduce, where the lowest nodes reach their results back to the parent node that had asked them. It calculates a partial result using the Reduce function (reduction) involving all the corresponding values for the same key to a unique pair (key, value). Then goes back information in turn.

Challenges:

i. Need more resources.
ii. Implementing each query as a MR program is difficult.
iii. No primitives for common operations(selection/extraction).

IV. HARDWARE PLATFORMS FOR BIG DATA

***Apache hadoop*:** Apache Hadoop is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant. The programming model used in Hadoop is MapReduce. Major drawbacks of MapReduce is its inefficiency in running iterative algorithms. MapReduce is not designed for iterative processes.

***Spark*:** Spark is a next generation paradigm for big data processing developed by researchers atb the University of California at Berkeley. It is an alternative to Hadoop which is designed to overcome the disk I/O limitations and improve the performance of systems.

***High performance computing (HPC)*:** HPC clusters also called supercomputers are machines with thousands of cores. They can have a different variety of disk organization, cache, communication mechanism. They are not as scalable as Hadoop or Spark clusters but they are still capable of processing terabytes of data.

V. CONCLUSION

This paper describes the different clustering techniques and the algorithms with the challenges they pose with Big data. Different clustering algorithms are used to manage large sets of data. It shows that these algorithms are insufficient to face all the challenges of big data. No clustering algorithm performs well for all the evolving criteria. Parallel classification is very useful for big data clustering,but the complexity of implementation is a great challenge. However, the MapReduce framework can provides a very good basis for the implementation of such parallel algorithms. The clustering technique also plays a significant role in data analysis and data mining applications. In order to manage large volume of data while keeping an acceptable resource needs, we have to improve clustering algorithms by reducing their complexity in terms of time and memory.

VI. REFERENCES

[1] A. Fahad, N. Alshatri and Z. Tari, “A Survey of Clustering Algorithms for Big Data: Taxonomy”, IEEE Transactions on Emerging Topics in Computing 2014

[2] Btissam Zerhari, Ayoub Ait Lahcen and Salma Mouline, “Big Data Clustering: Algorithms and Challenges”, International Conference on Big Data, Cloud and Applications BDCA'15 , At Tetuan, Morocco , conference paper may 2015

[3] Apurva Juyal Dr. O. P. Gupta,”A Review on Clustering Techniques in Data Mining”,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014

[4] Keshavanse, Meena Sharma,”Clustering methods for Big data analysis”,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015

[5] S.M. Junaid, K.V. Bhosle,” Overview of Clustering Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 11, November 2014

[6] DongkuanXu and YingjieTian, “A Comprehensive Survey of Clustering Algorithms”, Annals of Data Science, Springer-Verlag Berlin Heidelberg August 20

[7] C. YADAV, S. WANG, et M. KUMAR, “Algorithm and approaches to handle large Data-A Survey,” International Journal of computer science and network, vol 2, issue 3, 2013.

[8] Manish Kumar Kakhani, Sweeti Kakhani and S.R. Biradar, “Research Issues in Big Data Analytics”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 8, August 2013

[9] . Justin Samuel, Koundinya RVP, KothaSashidhar and C.R. Bharathi, A Survey on Big Data and its Research Challenges, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 8, May 2015

[10] MuneshKataria, Ms.Pooja Mittal, Big Data : A Review, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July-2014, pg. 106-110.

[11] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data mining with Big Data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no 1, p. 97-107, 2014.

[12] Abdul Raheem Syed, Kumar Gillela, Dr. C. Venugopal, “The Future Revolution on Big Data”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013

[13] http://www.information-management.com/gallery/Big-Data-Required-Software-Applications-10026664-1.html

[14] http://www.slideshare.net/ijtetjournal/icicce0153

[15] http://www.technologytransfer.eu/article/98/2012/1/What_Is_Big_Data_and_Why_Do_We_Need_It_.html

[16] Dilpreet Singh and Chandan K. Reddy, “A Survey on Platforms for Big Data Analytics”, Journal of Big Data, 1:1, 8, 2014.