International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1387 – 1391

Big Data Clustering Algorithm and Strategies

Nithya P
Computer Science and Engineering
Government College of Engineering,
Salem ,India
e-mail:pnithyame@gmail.com

Kalpana A M
Computer Science and Engineering
Government College of Engineering,
Salem ,India
e-mail:Kalpana.gce@gmail.com

***Abstract*—-** In current digital era extensive volume ofdata is being generated at an enormous rate. The data are large, complex and information rich. In order to obtain valuable insights from the massive volume and variety of data, efficient and effective tools are needed. Clustering algorithms have emerged as a machine learning tool to accurately analyze such massive volume of data. Clustering is an unsupervised learning technique which groups data objects in such a way that objects in the same group are more similar as much as possible and data objects in different groups are dissimilar. But, traditional algorithm cannot cope up with huge amount of data. Therefore efficient clustering algorithms are needed to analyze such a big data within a reasonable time. In this paper we have discussed some theoretical overview and comparison of various clustering techniques used for analyzing big data.

Keywords-*Big data, clustering algorithm, unsupervised learning.*

*****

I. INTRODUCTION

In the current digital era huge amounts of multidimensional data have been collected in various fields such as marketing, bio-medical and geo-spatial fields.It also includes twitter messages, emails, photos, video clips, sensor data etc. that are being produced and shared. It is estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created every day. Facebook alone records 10 billion message transfers, 4.5 billion "like" button clicks and 350 million photos upload per day. This data explosion makes data sets too large to store and analyze using traditional data base technology.

These massive quantities of data are generated because of the development of internet, the raise of social media, use of mobile and the information of Internet of things (IOT) by and about people, things and their interactions [1],[2]. Storing such huge amount of data is no longer a serious issue, but how to design solution to understand this big amount of data is amajor challenge [1].Operations such as analytical operations, process operations, retrieval operations are very difficult and huge time consuming.Unsupervised Machine Learning or clustering is one of the significant data mining techniquesfor discovering knowledge in multidimensional data[8].

Machine learning (ML) techniques focus on organizing data into sensible groups. It can be categorized into Supervised Machine learning and unsupervised machine learning. Supervised machine learning is the task of inferring a function form labeled training data whereas unsupervised machine learning is the task of inferring a function to describe hidden structure from "unlabeled" data.Clustering belongs to unsupervised machine learning. It is one of the popular approaches in data mining and has been widely used in big data analysis. The goal of Clustering involves dividing similar data points in a group and dissimilar data points into dissimilar groups. It is an unsupervised classification task which produces labeled classification of objects without any prior knowledge[3]. The similarity between data points is defined using some inter-observation distance measures and similarity measures [13].

Clustering analysis is broadly used in many applications such as Pattern recognition, Image analysis, Bioinformatics, Machine Learning, Image processing and information retrieval etc.

Many optimized clustering algorithms are performed in single machine environment, but for processing huge amount of data parallel processing and cloud infrastructure is needed. Parallelization of clustering algorithms has become paramount for processing big data[4].To achieve Parallelization the data or computation is broken down into parallel tasks and the task executes the same function on different data[5][12].

The properties that are important to the efficiency and effectiveness of a novel algorithm are as follows [7]:

- Generate arbitrary shapes of clusters rather than be limited to some particular shape.
- Ability to deal with different types of attribute.
- Handle large volume as well as high dimensional features with acceptable time and space complexities.
- Able to detect outliers.
- Does not sensitive to the order of input.
- Decrease the trust of algorithm on user dependent parameters.
- Show good data visualization and helps users to interpret the data.
- Can be scalable.

II. CLUSTERING ALGORITHM CATEGORIES

Clustering algorithm can be broadly classified as follows

- Partitioning clustering
- Hierarchical clustering
- Density based clustering
- Grid based clustering
- Model based clustering
- Evolutionary clustering

A. Partitioning clustering:

Partition clustering is the approach that partitions data set containing n observations into k groups or clusters. This algorithm uses iterative process to optimize the cluster centers,

as well as number of clusters.K-Means, K-medoids, PAM, CLARA are the partitioning clustering methods used to analyze large data sets.

B. Hierarchical clustering:

Hierarchical clustering algorithm does not require pre specification of number of clusters to be generated. It is divided into two types, 1.Agglomerative Hierarchical Clustering and 2. Divisive Hierarchical Clustering .In Agglomerative HierarchicalClustering, all sample objects are initiallyconsidered as individual clusters(size of 1).Then the most similar clusters are iteratively merged until there is just one big cluster .In divisive clustering, all sample objects in a single cluster (size n). Then at each step the clusters are partitioned into a pair of clusters. The algorithm stops until all the observations are in their own clusters.

Hierarchical clustering is static, that is, data points assigned to a cluster cannot be re-assigned to another. It fails to separate overlapping clusters due to lack of information regarding the shape and size of clusters.BIRCH, CURE,ROCK and Chameleon are some of the well-known algorithms of this category.

C. Density based clustering:

Clustering separates data objects based on their regions of density, connectivity and boundary. The clusters are defined as connected dense component which grow in any direction that density leads to. Density based clustering is good for filtering out noise and discovering arbitrary shaped clusters. DBSCAN, OPTICS, DBCLASD, DENCLUE are algorithms that filter out noise and discover clusters of arbitrary shape.

D. Grid based clustering:

This method divides space data objects into a number of cells and performs the clustering on the grids. The performance of grid based clustering depends on the size of the grid, which is usually much less than the size of the database. However for highly irregular data distributions, grid may not be sufficient to obtain required clustering quality. STINGS, Wave –Cluster are typical examples of this category.

E. Model based clustering:

This model assumes that the data are coming from distribution that is mixture of two or more components. This method optimizes the fit between the given data and some (predefined) mathematical model. An advantage of this model is that it can automatically determine the number of clustersbased on statics, taking noise into account. MCLUST, EM and COBWEB are best known model based algorithms.

F. Evolutionary clustering:

Clustering approaches use particle swarm optimization, genetic algorithmand other evolutionary approach for clustering. This approach use stochastic and use an iterative process. This algorithm starts with a random population of solutions, which is a valid partition of data with a fitness value. In iterative steps evolutionary operators are applied to generate the new populations until the termination condition.

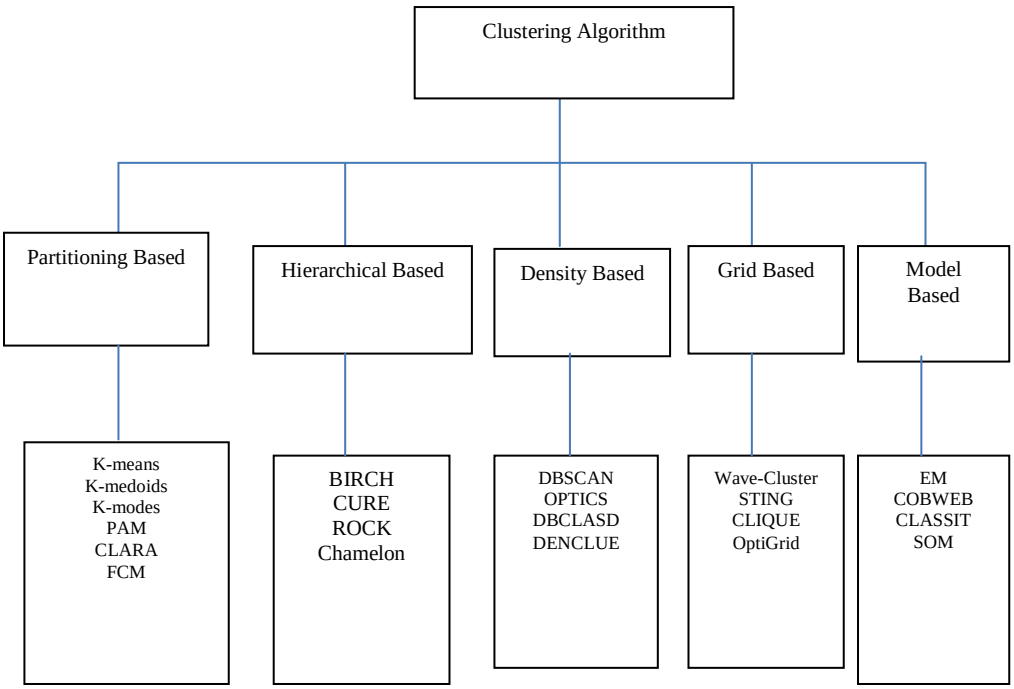

Figure 1: Clustering algorithm Categories

TABLE 1: COMPARITIVE ANALYSIS OF VARIOUS ALGORITHMS WITH MERITS AND DEMERITS.

Type of clustering	Technique	Type of data set Used(Variety)	Merits	Demerits
Partitioning Algorithm	ELM K-means	Numerical	1. Relaxed to appreciate and implement. 2. Produce additional thick clusters than the hierarchical technique especially when clusters are circular.. 3. Suitable for processing large datasets.	1. Deprived at usage of noisy data and outliers. 2. Works only on numeric data. 3. Haphazard preliminary cluster center problem. 4. Not appropriate for non-spherical clusters. 5. User has to provide number of clusters.
	K-modes and K-Prototype algorithm[5]	Mixed Numeric and Categorical data		
	k-mediods	Categorical		
	PAM	Numerical		
	CLARA	Numerical		
	CLARANS	Numerical		
	FCM	Numerical		
Hierarchical algorithm	BIRCH	Numerical	1. It is more adaptable 2. Less robust to noise and outliers. 3. Appropriate to any characteristic type.	1. If a process (merge or split) is performed, it cannot be undone. 2. Incompetence to scale well.
	CURE	Numerical		
	ROCK	Categorical and Numerical		
	Chameleon	All types of data		
	ECHIDNA	Multivirate data		
Density based algorithm	DBSCAN	Numerical	1. Resilient to outliers. 2. Does not necessitate the amount of clusters. 3. Forms clusters of uninformed shapes. 4. Unresponsive to organization of data objects	1. Inappropriate for high-dimensional datasets due to the expletive of dimensionality singularity. 2. Its quality be contingent upon the threshold set
	OPTICS	Numerical		
	DBCSLASD	Numerical		
	DENCLUE	Numerical		
Grid based algorithm	Wave-cluster	Special data	1.Fast handling time. 2.Self governing of the amount of data objects.	1. Be dependent only on the amount of cells in each dimension in the quantized space.
	STING	Special data		
	CLIQUE	Numerical		
	OptiGrid	Special data		
Model based algorithm	EM	Special data	1. Vigorous to noisy data or outlier 2. Fast handling speed 3. It decides the amount of clusters to produce	Multifarious in nature.
	COBWEB	Numerical		
	CLASSIT	Numerical		
	SOMs	Multivariate data		

III. CLUSTERING BENCHMARK

Explicit standards need to be used in big data for the three main characteristics such as Volume, Velocity and variety[6]. Some clustering criteria needs to be considered while evaluating the efficiency of clustering with respect to volume are as follows:

- Where the datasets can be used
- Management of High dimensional data
- Managing noisy data in the dataset.

For the appropriate clustering procedure with respect to the variety method, we have to consider the criteria's such as

- Dataset Categorization
- Shape of the clusters.

The criteria used to measure the property of velocity are

- Algorithm Complexity
- Performance of the algorithm during runtime.

IV. BIG DATA

Big data refers to a collection of massive amount of data which are complex, growing with multiple and autonomous sources are high spot of big data. Big data can be defined as tremendous amount of data with larger number of attributes collected or accumulated together from different sources and for different purpose [3].These data comes from different resources are collected for different purpose are combined together both logically and illogically to derive meaningful insights. The technological revolution has made large memory spaces cheaper and easy to acquire [1]. But exploring interesting phenomena from big data is difficult since it requires massively parallel software running on tens, hundreds or even thousands of servers.

There are three characteristics which makes challenge for discovering unknown and unidentified patterns and information are:

A. Data with heterogeneous and diverse dimensionality

Big data has diverse dimensionality, since the data's are collected from different sources and locations. They have

different scheme and rules for data storing. Data aggregation from various different resources is a major challenge [3].

B.Autonomous sources with distributed and decentralized control

The main characteristics of big data applications are distributed and decentralized controls. All sources are able to gather information without depending on any other module.

C.Complex and evolving relationship

The complexity and the relationship between the data are getting increased with the volume of data.The centralized information system contains sample features which treat individual as independent entity without considering social connections.In the dynamic world, with respect to temporal,spatial and other factors,the features are getting evolved to represent the individuals and social connections.

V. BIG DATA CLUSTERING ALGORITHMS

Traditional algorithm cannot cope up with huge amount of data.Unlike traditional clustering algorithm, volume of data must be taken into account is huge, so this requires significant changes in architecture of storage system. Most traditional clustering algorithms are designed to handle either numeric or categorical data with limited size. But big data clustering deals with different kinds of data such as text, image, video,mobile devices, sensors, etc. Moreover the velocity of big data requires big data clustering techniques that have high demand for the online processing of data[9]. Thus the issue with the big data clustering is how to speed up and scale up the clustering algorithm with minimum yield to the clustering quality.

There are three ways to speed up and scale up big data clustering algorithms.

1. Reduce the iterative process using sampling based algorithms. Since this kind of algorithm performs clustering on the sample data set instead of performing on the whole data set, complexity and memory space needed for processing decreases.
2. Reduce dimensions using randomized techniques. Dimensionality of data set influences the complexity and speed of clustering algorithms. Random projection and global projection are used to project dataset from high dimensional space to a lower dimensional apace.
3. Implement parallel and distributed algorithms, which use multiple machines to speed up the computation in order to increase the scalability[10].

VI. INSTANCES OF CLUSTERING TECHNIQUES IN BIG DATA

There are two main types of techniques designed for large scale data sets, which are based on the number of computer nodes that have been used.

A. Single Machine techniques
B. Multi Machine techniques

Due to the nature of scalability and faster response time multi machine clustering techniques have attracted more attention.

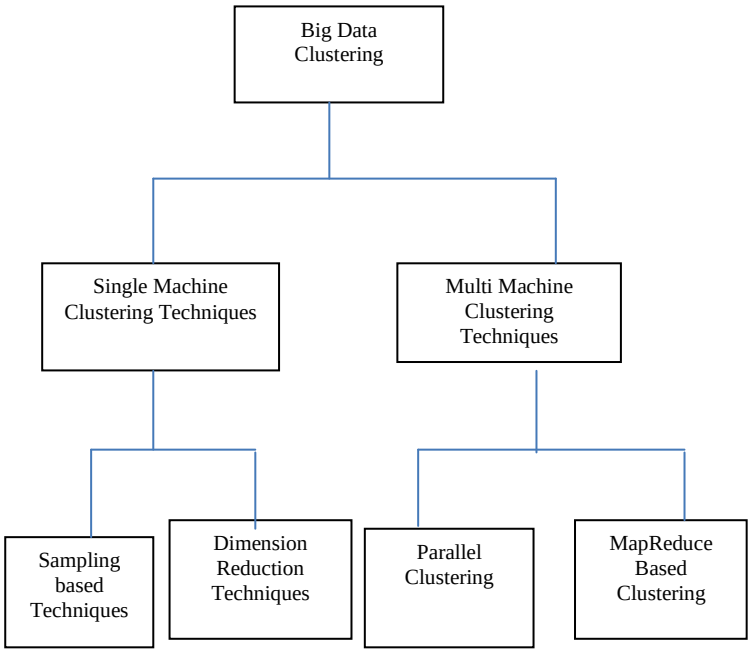

Figure 2. List of Big data Clustering Techniques.

A. Single Machine Clustering Techniques

Single Machine clustering technique uses resources of single machine since it runs run in one machine. Sampling based techniques and dimension reduction techniques are two common strategies for single machine technique.

A.1 Sampling Based Technique

The problem of scalability in clustering algorithms is in terms of computing time and memory requirements. Sample based algorithms handle one sample data sets at a time and then generalize it to whole data set. Most of these algorithms are partitioning based algorithms. A few examples of this category include CLARANS, BIRCH and CURE.

A.2 Dimension Reduction Techniques

The number of instances in the data set influences the complexity and the speed of the clustering algorithms[9]. However objects in data mining consist of hundreds of attributes. High dimensionality of the data set is another influential aspect and clustering in such dimensional spaces requires more complexity and longer execution time[9]. One approach to reduce the dimension is projection. A data set can be projected from high dimensional space to a lower dimensional space. Principal Component Analysis (PCA) is one method used to reduce the dimensionality of a data set. The reduced representation of data set use less storage space and faster execution time. Subspace clustering is an approach which reduces the dimensions in the data set. Many dimensions are irrelevant in high dimensional data. Using this algorithm relevant dimension is obtained to reduce the complexity. Other methods such as CLIQUE and DRCC can also be used to reduce the dimensions in high dimensional data set.

B. Multiple Machine Clustering Techniques

In this age of data explosion, parallel processing is essential to process a massive volume of data in a timely manner. Single Machine clustering Techniques with a single processor and a memory cannot handle the tremendous

amount of data.Algorithms that can be run on multiple machines are needed. Multiple machine clustering technique divides the huge amount of data into small pieces. These pieces are loaded on different machines and solved using the resources of these machines. Parallel processing application includes both conventional and data intensive applications. Data intensive applications are I/O bound and devote largest fraction of execution time to the movement of data. Map Reduce are common parallel processing models for computing data intensive applications.

The steps involved in parallel clustering

1. Input data are partitioned and distributed over different machines.
2. Each machine performs local clustering on its input data.
3. The information of machine is aggregated globally to produce global clusters for the whole data set.

B.1 Parallel Clustering

There are three kinds of parallelism mechanisms used in data mining algorithms are as follows[12].

1. Independent parallelism-the whole data is operated on each processor and no communication between processors
2. Task Parallelism-Different algorithms are operated on each processor
3. SPMD(Single Program Multiple data) Parallelism-The same algorithm is executed on multiple processors with different partitions.

B.2 MapReduce Based Clustering

It is one of the most efficient big data solutions which process massive volume of data in parallel with many low end computing models. This programming paradigm is a scalable and fault tolerant data processing tool that was developed for large scale data applications.

MapReduce model hides details of parallel execution, which allows users to emphasis only on data processing strategies. The MapReduce model consists of two phases: Mappers and reducers. The mappers are designed to generate a set of intermediate key/value pairs. The reducer is used as a shuffling or combining function to merge all of intermediate values associated with the same intermediary key.

VII. APPLICATION

Due to the verity and complexity of data, big data clustering algorithm is widely used in the following areas such as[1],[11]:

- Image Segmentation in medicine area.
- Load Balancing in parallel computing.
- Gene expression data analysis.
- Community Detection in the network.

VIII. CONCLUSION

This paper providesliterature review about clustering in big data and various clustering techniques used to analyze big data.The traditional single machine techniques are not powerful enough to handle large volume of data.Parallel clustering is potentially useful for big data clustering. Parallel clustering algorithm has to be developed by the researchers in order to provide scalability and high throughput. But the complexity of implementing such algorithm is a challenge. In future parallel clustering algorithms are implemented for a particular application and the experimental result will be compared to propose a better algorithm.

REFERENCES

[1] Min chen, Simone A. Ludwig and Keqin li," Clustering in Big data",CRCPress,United states, 2017.

[2] AdilFahad,NalaaAlshatri,ZahirTari,AbdullahAlamri,IbrahimKhalil,AlbertY.Zomaya,Sebti Foufou and Abdelaziz Bouras,"A survey of clustering Algorithms for big data: Taxonomy and Empitical Analysis",IEEE transactions on Emerging Topics in computing,Vol 2,Issue 3,pp.267-279, September 2014.

[3] V W Ajin,Lekshmy D Kumar,"Big data and clustering algorithms",IEEE International Conference on Research Advances in Integrated Navigation Systems,pp.1-5, April 2016.

[4] Tao,Yue Zhang, and Kwei-jay lin,"Efficient Algorithms for web Services Selection with End-to –End QOS Constraints",ACM Transactions on the web,Vol.1,No.1,Article 6, May 2007.

[5] Justin O'Sullivan,D Edmond, and A TerHofstede,"What's in a service?"Distributed and Parallel databases",vol.12,Issue 2,pp.117-113,September 2002.

[6] Dr.Meenu Dave and Hemant Gianey,"Different Clustering Algorithms for Big Data Analytics:AReview",IEEE International conference on system modeling and Advancement in Research Trends, November 2016.

[7] Rui Xu and D Wunsch,"Survey of Clustering Algorithms,"IEEE Transaction on.Neural Network, Vol.16, no.3, pp.645-678,May 2005.

[8] Shirkhorshidi A S,Aghabozorgi S,Wah T Y,Herawan T," Big Data Clustering:A Review",) Computational Science and its Applications-ICCSA 2014.Lecture Notes in Computer Science,Vol 8583.Springer, 2014.

[9] Xu Z &Shi Y, "Exploring big data analysis :Fundamental Scientific problems" Annals of data Science,vol.2(4),pp.363-372,2015.

[10] Kim W,"Parallel Clustering Algorithms:Survey", Spring 2009.

[11] Aggarwal C C,& Reddy K," Data Clustering :Algorithms and applications", CRCPress, United States ,August 2013.

[12] Zhang J,".A parallel clustering Algorithm with MPI-K means".Journal of computers 8,no 1,pp10-17,2013.

[13] Jain A K, Dubes R C,"Algorithms for Clustering Data",Upper Saddle River,NJ,USA:Prentice-Hall,1988.