Survey of Various Clustering Techniques for Big Data in Data Mining

Kosha Kothari, Ompriya Kale
Department of Computer Science Engineering,
L.J Institute of Technology,Ahmedabad,Gujarat,India.

***Abstarct*— Data mining is the method which is useful for extracting useful information and data is extorted, but the classical data mining approaches cannot be directly used for big data due to their absolute complexity. The data that is been formed by numerous scientific applications and incorporated environment has grown rapidly not only in size but also in variety in recent era. The data collected is of very large amount and there is difficulty in collecting and assessing big data. Clustering algorithms have developed as a powerful meta-learning tool which can precisely analyze the volume of data produced by modern applications. The main goal of clustering is to categorize data into clusters such that objects are grouped in the same cluster when they are "similar" according to similarities, traits and behavior. The most commonly used algorithm in clustering are partitioning, hierarchical, grid based, density based, and model based algorithms. Partitioning algorithms can also be called as the centroid based clustering. Hierarchical algorithms can also be called as the connectivity based clustering. Density based clustering is based on the concept of data reachability and data connectivity. Grid based clustering is based on the size of the grid instead of the data. Model based clustering is based upon the probability distribution. In this survey paper, a review of clustering and its different techniques in data mining is done considering the criteria's for big data.**

***Index Terms*—Clustering Algorithms, Unsupervised learning, Data Mining, Big data.**

I. INTRODUCTION

With the beginning in the era of big data, the data is increasing at rapid speed not only in size but also in variety. There comes challenge and difficulties to handle such large amount of data with the growing data. Big data exhibits different characteristics like volume, variety, variability, value, velocity and complexity [1] due to which it is very difficult to analyze data and obtain information with traditional data mining techniques.

Data mining [2] is the process of analyzing data from different context and encapsulating it into useful information. Data mining consists of extracting, transforming, and loading transactional data into the data warehouse system, storing and managing the data in a multidimensional database system, providing data access to business analysts and information technology professionals, analyzing the data by application software, presenting the data in a useful format. Data mining includes the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, analysis of various clustering algorithms considering the criteria of big data is done.

The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data [3]. Clustering [4] – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The essential requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitive to the order of input records, etc. Data mining is a multi-step process. It requires collecting and converging data for a data mining algorithm, prospecting the data, evaluating the results and taking relevant action. The collected data can be stored in one or more operational databases, a data warehouse or a flat file. The data is mined using two learning approaches. i.e. supervised learning or unsupervised clustering.

Data Mining is a four step: Assemble data, Apply data mining tools on datasets, Interpretation and evaluation of result, Result application.

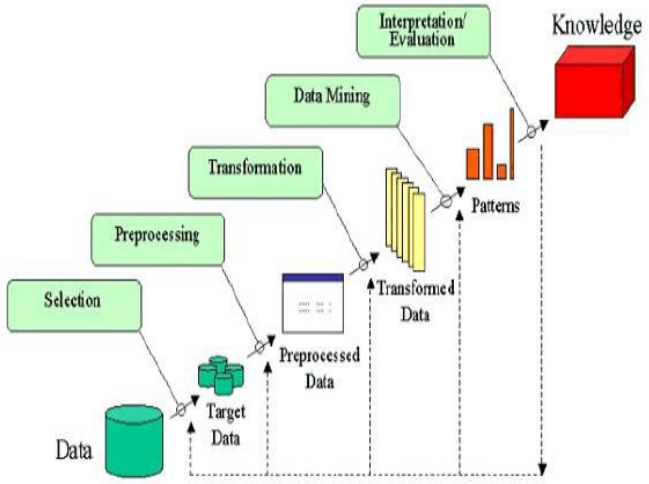

Figure 1: Steps of Data Mining Process

II. CRITERION TO BENCHMARK CLUSTERING METHODS

Distinct criteria need to be used to figure out the relative strengths and weaknesses of every algorithm when surveying clustering methods for big data, with respect to the three-dimensional properties of big data, including Volume, Velocity, and Variety.

- **Volume:** refers to the ability of a clustering algorithm to deal with a large amount of data [5]. It is the most distinct feature that demands specific requirements to all classical technologies and tools used. To guide the selection of 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:** refers to the ability of a clustering algorithm to handle different types of data (numerical, categorical and hierarchical). It deals with the complexity of Big data [1]. 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:** refers to the speed of a clustering algorithm on big data. Big Data are generated at high speed [1]. To guide the selection of a suitable clustering algorithm with respect to the Velocity property [5], the following criteria are considered: (i) Complexity of algorithm and (ii) the run time performance.

The correspondent criterion of big data is explained in detail:

- Type of dataset: Most of the classical clustering algorithms focus either on numeric data or on categorical data. But, the data that is been collected in the real world contains both numeric and categorical attributes. It is difficult for applying classical clustering algorithm directly into these kinds of data. Clustering algorithms work effectively either on purely numeric data or on purely categorical data; most of them perform poorly on mixed categorical and numerical data types.
- Size of dataset: The size of the dataset has a dominant effect on the quality of clustering. When the data size is small some clustering methods are more efficient clustering methods than others.
- Handling outliers/ noisy data: An effectual algorithm will often be able to handle outlier/noisy data because the data in most of the real applications are not pure. Also, noise will make it difficult for any algorithm to cluster an object into a suitable cluster. Thus it will affect the results that are provided by the algorithm.
- Stability: One of the important features for any clustering algorithm is the ability to generate the same partition of the data irrespective of the order in which the patterns are presented to the algorithm. Thus the stability of the algorithm should be quite effective.
- Cluster shape: A good clustering algorithm should be able to handle clusters of arbitrary shape that are produced by real data and their wide variety of data types [5].
- Input parameter: An alluring feature for clustering is the one that has fewer parameters, since a large number of parameters may affect cluster quality because they will depend on the values of the parameters.
- Handling high dimensionality: This is particularly important feature in cluster analysis because many applications require the analysis of objects containing a large number of features. For example, any text documents might contain thousands of terms or keywords as features. Thus it becomes complex due to the curse of dimensionality. Many dimensions may not be relevant. The data becomes increasingly sparse with the increase in number of dimensions, so that the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low.

III. CLASSIFICATION OF CLUSTERING ALGORITHM

Clustering is the essential task of Data Mining. Various algorithms are used for clustering. The most frequently used algorithms in Clustering are Partitioning-based, Hierarchical-based, Density-based, Grid-based and Model-based algorithms.

- **Partitioning-based**: Partitioning [6] algorithms spilt the data into several subsets. The reason of splitting the data into several subsets is that it is computationally not feasible to check every possible subset; there are certain greedy probing schemes used in the form of iterative inflation. Categorically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters. 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, and (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, while objects which are near the center represent the clusters in the K-medoids algorithm. There are many other partitioning algorithms such as PAM, CLARA, CLARANS, PSO and FCM.

- **Hierarchical-based**: Hierarchical[7][9] clustering algorithm groups data objects to form a tree shaped structure. It is also called as connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, and centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach.The hierarchical method has a major drawback though, which relates to the fact that once a step (merge or split) is performed, this cannot be undone. BIRCH, CURE, ROCK, Chameleon, Echidna, SOHAC, SWIFT and HGCUDF are some of the well-known algorithms of this category.

- **Density-based**: In Density-based [8] clustering, data objects are separated based on density regions, connectivity and boundary. They are intently related to point-nearest neighbours. Density based algorithm defined as a connected dense component continue to grow the given cluster as long as the density in the neighborhood exceeds certain threshold. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes as well as provide protection against outliers and noise. Hence the overall density point is evaluated to determine the functions of datasets that influence a particular data point. DBSCAN, OPTICS, DBCLASD, DENCLUE, DESCRY, STDBSCAN and CHRONICLE are algorithms that use such a method to filter out noise (outliers) and discover clusters of arbitrary shape.

- **Grid-based**: The grid-based [10] clustering approach uses a multiresolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space. In 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 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 fulfill the time requirement. The examples of Grid-based are Wave-Cluster, STING, OptiGrid and CLIQUE.

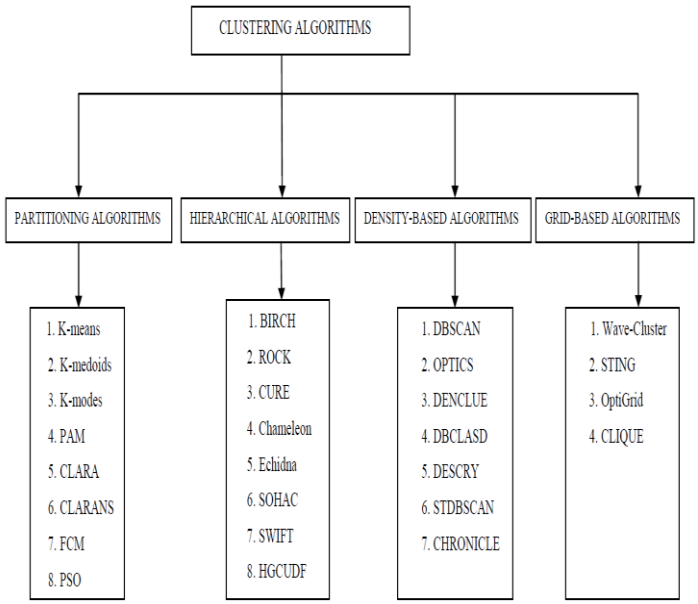

Figure 2: Clustering Algorithm Classification

IV. CONCLUSION

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different number of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering; the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells.

REFERENCES

[1] Avita Katal, Mohammad Wazid, and RH Goudar,
"Big data: Issues, challenges, tools and good practices". In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404-409. IEEE, 2013.

[2] Oded Maimon, Lior Rokach, "Data Mining AND Knowlwdge Discovery Handbook", Springer Science+Business Media.Inc, pp.321-352, 2005.

[3] A.K. JAIN, M.N. MURTY&P.J. FLYNN,"Data Clustering: A Review ",ACM Computer Survey, vol. 31, no. 3, pp. 264–323, 1999.

[4] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" ,International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

[5] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil A. Zomaya, S. Foufou, and A. Bouras ,"A Survey of Clustering Algorithms for Big Data:Taxanomy & Empirical Analysis". IEEE Transactions on Emerging Topics in Computing, JUNE 2014.

[6] Cheng-Ru Lin, Chen, Ming-Syan Syan , "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging" IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.

[7] Pradeep Rai, Shubha Singh, "A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.

[8] Pragati Shrivastava, Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.

[9] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.

[10] Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian,"GDCLU: a new Grid-Density based CLUstring algorithm",IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.102-107, 2012.