International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2022, Submitted in June 2022, iajesm2014@gmail.com

Study on The Big Data Clustering Technique

Anju, Research Scholar, Department of Computer Science, Monad University, Hapur, Uttar Pradesh (India)
Dr. Kailash Kumar Assistant Professor, Department of Computer Science, Monad University, Hapur, Uttar Pradesh (India)

Abstract:

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

Keywords: Big Data, Clustering, Mining, Techniques.

Introduction: Clustering is an essential data mining and tool for analyzing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big Data is referring to terabytes and petabytes 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. This study is aimed to review the trend and progress of clustering algorithms to cope with big data challenges from very first proposed algorithms until today's novel solutions. The algorithms and the targeted challenges for producing improved clustering algorithms are introduced and analyzed, and afterward the possible future path for more advanced algorithms is illuminated based on today's available technologies and frameworks.

Clustering is a kind of unsupervised machine learning technology, which is used to mine the intrinsic similarity of data and divide the data set into several subsets. Each data subset is a cluster, the samples within the cluster are similar to each other, and the samples between different clusters are not similar. In general, the similarity of samples is characterized by Euclidean distance, Markov distance, Manhattan distance, Pearson distance, Chebyshev distance, cosine similarity, Jaccard similarity and probability density. Clustering techniques have been widely used in real life, such as customer grouping in commercial activities, gene sequence classification in bioinformatics, spam identification in the Internet, and analysis of industry electricity usage behavior in the electricity market. With the advent of the era of big data, data collection and storage has become relatively easy and convenient. Large-scale data sets of GB-level and even TB-level storage are emerging one after another. The size of data sets of big data is growing at an unimaginable speed, which brings great challenges to data processing. Therefore, clustering research for large data sets is constantly emerging. So far, clustering algorithms for different types of small and medium-sized data sets have made a historic breakthrough in clustering accuracy. However, these algorithms still have many problems when dealing with large data sets. The main defects are high computational complexity and long computing time, which is unacceptable.

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

Simply it is the partitioning of similar objects which are applied to unlabelled data.

To deal with this avalanche of data, it is necessary to use powerful tools for knowledge discovery. Data mining techniques are well-known knowledge discovery tools for this purpose [3]–[9]. Clustering is one of them that is defined as a method in which data are divided into groups in a way that objects in each group share more similarity than with other objects in other groups [1]. Data clustering is a well-known technique in various areas of computer science and related domains. Although data mining can be considered as the main origin of clustering, but it is vastly used in other fields of study such as bio informatics, energy studies, machine learning, networking, pattern recognition and therefore a lot of research works has been done in this area [10]–[13].

Challenges of big data have root in its five important characteristics [15]:

Volume: The first one is Volume and an example is the unstructured data streaming in form of social media and it rises question such as how to determine the relevance within large data volumes and how to analyze the relevant data to produce valuable information.

Velocity: Data is flooding at very high speed and it has to be dealt with in reasonable time. Responding quickly to data velocity is one of the challenges in big data.

Variety: Another challenging issue is to manage, merge and govern data that comes from different sources with different specifications such as: email, audio, unstructured data, social data, video and etc.

Variability: Inconsistency in data flow is another challenge. For example in social media it could be daily or seasonal peak data loads which makes it harder to deal and manage the data specially when the data is unstructured.

Complexity: Data is coming from different sources and have different structures; consequently it is necessary to connect and correlate relationships and data linkages or you find your data to be out of control quickly.

Properties of Clustering :

1. Clustering Scalability: Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can’t get the appropriate result which would lead to wrong results.

2. High Dimensionality: The algorithm should be able to handle high dimensional space along with the data of small size.

3. Algorithm Usability with multiple data kinds: Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. Dealing with unstructured data: There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. Interpretability: The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Methods:

The clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method

- Model-Based Method
- Constraint-based Method

Partitioning Method: It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

- **Agglomerative Approach:** The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.
- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.
- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

Density-Based Method: The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

Grid-Based Method: In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Model-Based Method: In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

Constraint-Based Method: The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.
2. It can be used for exploratory data analysis and can help with feature selection.
3. It can be used to reduce the dimensionality of the data.
4. It can be used for anomaly detection and outlier identification.
5. It can be used for market segmentation and customer profiling.

Conclusion After continuous research, some research results have been achieved in big data, such as big data search, big data storage, big data mining, etc., but still cannot meet the needs of current big data. Researching real-time, highly robust new high-efficiency clustering algorithms for big data has become a key task to be solved in the deep exploration of the hidden value of big data. In the field of data mining, the final result of many clustering algorithms is sensitive to the correct setting of parameters, which leads to these algorithms far from being called mature and practical intelligent machine learning algorithms. In the big data environment, it is necessary to study and design a more efficient intelligent automatic clustering algorithm. Therefore, the clustering algorithm for big data needs constant research to meet the needs of current big data.

References:

1. T. C. Havens, J. C. Bezdek, and M. Palaniswami, “Scalable single linkage hierarchical clustering for big data,” in Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on. IEEE, 2013, pp. 396–401.

2. “YouTube Statistic,” 2014. [Online]. Available: http://www.youtube.com/yt/press/statistics.html.

3. P. Williams, C. Soares, and and J. E. Gilbert, “A Clustering Rule Based Approach for Classification Problems,” Int. J. Data Warehous. Min., vol. 8, no. 1, pp. 1–23, 2012.

4. R. V. Priya and A. Vadivel, “User Behaviour Pattern Mining from Weblog,” Int. J. Data Warehous. Min., vol. 8, no. 2, pp. 1–22, 2012.

5. T. Kwok, K. A. Smith, S. Lozano, and D. Taniar, “No Title,” in Parallel Fuzzy c-Means Clustering for Large Data Sets, 2002, pp. 365–374.

6. H. Kalia, S. Dehuri, and A. Ghosh, “A Survey on Fuzzy Association Rule Mining,” Int. J. Data Warehous. Min., vol. 9, no. 1, pp. 1–27, 2013.

7. O. Daly and D. Taniar, “Exception Rules Mining Based on Negative Association Rules,” in Proceedings of the International Conference on Computational Science and Its Applications (ICCSA 2004), 2004, pp. 543–552.

8. M. Z. Ashrafi, D. Taniar, and K. A. Smith, “Redundant association rules reduction techniques,” Int. J. Bus. Intell. Data Min., vol. 2, no. 1, pp. 29–63, 2007.

9. D. Taniar, W. Rahayu, V. C. S. Lee, and O. Daly, “Exception rules in association rule mining,” Appl. Math. Comput., vol. 205, no. 2, pp. 735–750, 2008.

10. Meyer, F. G., and J. Chinrungrueng., “Spatiotemporal clustering of fMRI time series in the spectral domain,” Med. Image Anal., vol. 9, no. 1, pp. 51–68, 2004.

11. J. Ernst, G. J. Nau, and Z. Bar-Joseph, “Clustering short time series gene expression data.,” Bioinforma. 21, vol. 21, no. suppl 1, pp. i159 – i168, Jun. 2005.

12. F. Iglesias and W. Kastner, “Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns,” Energies, vol. 6, no. 2, pp. 579–597, Jan. 2013.