Journal of American Science

Websites:
http://www.jofamericanscience.org
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Data clustering technique in big data mining

[1]Sandeep Chopade & [2]Dr. Kailash Aseri

[1]Research Scholar, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
[2]Assistant Professor, Department of Computer Science, SunRise University, Alwar, Rajasthan (India)
Email- raghusand@gmail.com

Abstract: Cluster Analysis in Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups. In the process of clustering in data analytics, the sets of data are divided into groups or classes based on data similarity. Then each of these classes is labelled according to their data types. Going through clustering in data mining example can help you understand the analysis more extensively. In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. After the classification of data into various groups, a label is assigned to the group. It helps in adapting to the changes by doing the classification. So if we were to define clustering in data mining, then we can say that the process of cluster in data mining is basically comprising a set of abstract objects into groups of similar objects. The process of dividing and storing them in these groups is known as cluster analysis.

[Chopade, S. and Aseri, K. **Data clustering technique in big data mining.** *J Am Sci* 2023;19(5):77-79]. ISSN 1545-1003 (print); ISSN 2375-7264 (online). http://www.jofamericanscience.org. 07. doi:10.7537/marsjas190523.07.

Keywords: Big Data, Clustering, Mining, Techniques.

Introduction: 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.

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] Luo Enzhen, Wang Guojun, Li Chaoliang. Research on clustering algorithm for multidimensional data deduplication in big data environment[J]. Small Computer Systems, 2016, 37 (3) : 438-442.

[2] Lin Qingxin. Exploring the application of K-means clustering algorithm under big data in network security detection[J]. Network Security Technology and Application, 2017, (3): 92-93.

[3] JI Lianghao. Research on clustering algorithm based on density biased sampling ［J］ Journal of Chongqing Uni-versity of Posts and Telecommunications Natural Science Edition, 2007, 19(6) : 729-732.

[4] Li Xuelong, Gong Haigang. Review of Big Data Systems[J]. Chinese Science: Information Science, 2015, 45(1): 1–44

[5] Zhou Runwu, et al. Parallel optimal sampling clustering K-means algorithm for big data processing [J]. Computer Applications, 2016, 36(2) : 311 - 315.

5/6/2023