International Journal of Innovative Research in Engineering
Volume 3, Issue 5 (September-October 2022), PP: 148-151.
www.theijire.com **ISSN No: 2582-8746**

Cluster Analysis in Data Mining

U.M.Nandhini[1], K. Sowmiya Sri[2], S.Srimega[3], S.Swetha[4]
[1,2,3,4] *Department Of Computer Science, Sri GVG Visalakshi College for women, Udumalpet, Tamil nadu, India.*

How to cite this paper: U.M.Nandhini[1], K.Sowmiya Sri[2], S.Srimega[3], S.Swetha[4]. "Cluster Analysis in Data Mining", IJIRE-V3I05-148-151.

Abstract: *Here, we are going to discuss Cluster Analysis in Data Mining. So first let us know about what is Data Mining & What is clustering in Data Mining then it introduction & the need for clustering in data mining. The blog cover how to define clustering in data mining, the different types of cluster in data mining and why clustering is so important. We are also going to discuss the algorithm and application of cluster in Data Science. Later we will learn about the different approaches in cluster analysis and data mining clustering methods.*

I. INTRODUCTION

Cluster Analysis is the process to find similar group of objects in order to formclusters. It is an unsupervised machine learning – based algorithm that acts on unlabeled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group and used to add clustering methods that are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio - medical and geo - spatial. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi objective optimization. In this article wewill taking you through the type of clustering, different clustering algorithms and a comparison of clustering methods.

II. METHODOLOGY

1. What is Data Mining?
2. What is Clustering in Data Mining?
   2.1 What is Cluster Analysis in Data Mining?
3. Application of Data Mining Cluster Analysis
4. Requirements of Clustering in Data Mining
5. Data Mining Clustering Methods
   5.1 Partitioning Clustering Method
   5.2 Hierarchical Clustering Methods
   5.3 Density - Based Clustering Method
   5.4 Grid - Based Clustering Method
   5.5 Model – Based Clustering Method
   5.6 Constraint – Based Clustering Method
6. What kind of classification is not considered a cluster analysis?
7. Advantages & Disadvantages of Cluster Analysis
8. Conclusion
9. Reference

1. What is Data Mining?

Data Mining is the process of sorting through large data set to identify patternsand relationship that can help solve business problems through data analysis.

Data Mining techniques & tools enable enterprises to predicts future trends &make more informed business decisions. Data Mining is a crucial componentof successful analytics initiatives in organizations.
Example: large database & improve market segmentation.

2. What is Clustering in Data Mining?

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 clusteranalysis, which is based on the similarities of the data.
Example: Retail Marketing.

2.1. What is Cluster Analysis in Data Mining?

Cluster Analysis in Data mining means that to find out the group objects which are similar to each other in the group but are different from the object in other groups. The process of clustering in data analytics, the set of data are divided into groups or classes based on data similarity.

3. Application of Data Mining Cluster Analysis

There are many uses of Data Clustering analysis such as image processing. Data analysis, pattern recognition, market research and many more. Using Data Clustering, companies can discover new groups in the database of customers. They can also classify the existing customer base into distinct groups depending on the patterns of their purchases. Classification of data can also be done based on patterns of purchasing.

Taxonomy or the classification of animals with the help of cluster analysis is very common in the field of biology. Clustering can help identify and group species with similar genetic features and functionalities and also give us an understanding of some of the most commonly found inherent structures of specific populations or species. Areas are identified using the clustering in data mining. In the database of earth observation, lands are identified which are similar to each other.

Based on geographic location, value and house type, a group of houses are defined in the city. Clustering in data mining helps in the discovery of information by classifying the files on the internet. It is also used in detection applications. Fraud in a credit card can be easily detected using clustering in data mining which analyzes the pattern of deception.

4. Requirements of Clustering in Data Mining

Interpretability

The result of clustering should be usable, understandable and interpretable. The main aim of clustering in data analytics is to make sure haphazard data is stored in groups based on their characteristical similarity.

Helps in dealing with messed up data

- Usually, the data is messed up and unstructured. It cannot be analyzed quickly, and that is why the clustering of information is so significant in data mining. Grouping can give some structure to the data by organizing it into groups of similar data objects.
- It becomes more comfortable for the data expert in processing the data and also discover new things. Analyzing data that has already been classified and labeled through clustering is much easier than analyzing unstructured data. It also leaves less room for error.

High Dimensional

Data clustering is also able to handle the data of high dimension along with the data of small size. The clustering algorithms in data mining need to be able to handle any dimension of data.

Attribute shape clusters are discovered

Clustering algorithms in data mining should be able to detect arbitrarily shaped clusters. These algorithms should not be limited by only being able to find smaller, spherical clusters.

Dealing with Erroneous Data

Usually, databases carry a lot of erroneous, noisy or absent data. If the algorithm being used during clustering is very sensitive to this type of anomaly, then it can lead to low-quality clusters. That is why it is very important that your clustering algorithm can handle this type of data without problems.

Algorithm Usability with multiple data kind

Many different kinds of data can be used with algorithms of clustering. The data can be like binary data, categorical and interval-based data.

Clustering Scalability

The database usually is enormous to deal with. The algorithm should be scalable to handle extensive database, so it needs to be scalable.

5. Data Mining Clustering Methods

5.l. Partitioning Clustering Method

In this method, let us say that "m" partition is done on the "p" objects of the database. A cluster will be represented by each partition and m < p. K is the number of groups after the classification of objects. There are some requirements which need to be satisfied with this Partitioning Clustering Method and they are: –

- One objective should only belong to only one group.

- There should be no group without even a single purpose.
  There are some points which should be remembered in this type of PartitioningClustering Method which are:
- Therefore will be an initial partitioning if we already give no. of apartition (say m).
- There is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning.

5.2. Hierarchical Clustering Methods

Among the many different types of clustering in data mining, In this hierarchicalclustering method, the given set of an object of data is created into a kind of hierarchical decomposition. The formation of hierarchical decomposition will decide the purposes of classification. There are two types of approaches for the creation of hierarchical decomposition, which are: –

- **Divisive Approach**

Another name for the Divisive approach is a top-down approach. At the beginning of this method, all the data objects are kept in the same cluster. Smaller clusters are created by splitting the group by using the continuousiteration. The constant iteration method will keep on going until the condition of termination is met. One cannot undo after the group is split ormerged, and that is why this method is not so flexible.

- **Agglomerative Approach**

Another name for this approach is the bottom-up approach. All the groups are separated in the beginning. Then it keeps on merging until all the groupsare merged, or condition of termination is met.

There are two approaches which can be used to improve the HierarchicalClustering Quality in Data Mining which are: –

1. One should carefully analyze the linkages of the object at every partitioningof hierarchical clustering.
2. One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are groupedinto micro-clusters. After grouping data objects into micro clusters, macroclustering is performed on the micro cluster.

- **Density-Based Clustering Method**

In this method of clustering in Data Mining, density is the main focus. Thenotion of mass is used as the basis for this clustering method. In this clustering method, the cluster will keep on growing continuously. At least one number of points should be there in the radius of the group for each point of data.

- **Grid-Based Clustering Method**

In this type of Grid-Based Clustering Method, a grid is formed using the object together. A Grid Structure is formed by quantifying the object spaceinto a finite number of cells.

Advantage of Grid-based clustering method: –

- Faster time of processing: The processing time of this method is muchquicker than another way, and thus it can save time.
- This method depends on the no. of cells in the space of quantized eachdimension.

- **Model-Based Clustering Methods**

In this type of clustering method, every cluster is hypothesized so that it canfind the data which is best suited for the model. The density function is clustered to locate the group in this method.

- **Constraint-Based Clustering Method**

Application or user-oriented constraints are incorporated to perform the clustering. The expectation of the user is referred to as the constraint. In this process of grouping, communication is very interactive, which is provided by the restrictions.

6. What kinds of classification are not considered acluster analysis?

- **Graph Partitioning –** The type of classification where areas are not the same and are only classified based on mutual synergy and relevance is not cluster analysis.
- **Results of a query –** In this type of classification, the groups are created based on the specification given from external sources. It is not counted as a Cluster Analysis.
- **Simple Segmentation –** Division of names into separate groups of registration based on the last name does not qualify as Cluster Analysis.
- **Supervised Classification –** Those type of classification which is classified using label information cannot be said as Cluster Analysisbecause cluster analysis involves group based on the pattern.

7. Advantages & Disadvantages of Cluster Analysis

The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. Disadvantages of clustering arecomplexity and inability to recover from database corruption.

In a clustered environment, the cluster uses the same IP address for Directory Server and Directory Proxy Server, regardless of which cluster node is actually running the service. That is, the IP address is transparent to the client application. In a replicated environment, each machine in the topology has its own IP address. In this case, Directory Proxy Server can be used to provide a single point of accessto the directory topology. The replication topology is therefore effectively hidden from client applications. To increase this transparency, Directory Proxy Server can be configured to follow referrals and search references automatically. Directory Proxy Server also provides load balancing and the ability to switch to another machine when one fails.

III.CONCLUSION

So now we have learned many things about Data Clustering such as the approachesand methods of Data Clustering and Cluster Analysis in Data mining. Going through diverse clustering in data mining example can further assist you to get an in-depth insight into the process.

Reference

1. *Rohit Sharma,Program Director for the UpGrad-IIIT Bangalore, PG Diploma Data Analytics Program.*