Survey on Data Clustering Techniques

[1]Sindhu Madhuri G, [2]Santhosh S, [3]Ganesh

Dept. of Computer Science and Engineering,

JAIN (Deemed-to-be University), Bengaluru, India

Email Id-[1]g.sindhumadhuri@jainuniversity.ac.in, [2]santhosh.s@jainuniversity.ac.in, [3]ganeshdmca@gmail.com

ABSTRACT: ***Data Clustering is an important method used in the process of mining the data. It is a mechanism in which a collection of objects is allocated to classes where they are called clusters. Objects belonging to a particular cluster are identical in that cluster to other objects but dissimilar from objects belonging to other clusters. In case the data is spread across multiple sites, the clustering process becomes difficult and complex. When applied to distributed applications, distributed clustering comes as a relief to the problems of traditional clustering. Some clustering methods like centralized methodology, grid-based technology, density-based technology and their algorithms was discussed in this paper. The partitioned approach divides the data collection into system. It consists with a certain criterion of similarity, the hierarchical method establishes a hierarchy between the clusters, merging data items into groups. This article introduces the density-based strategy of distinguishing clusters of high population density with higher density objects by fusion of data objects onto rows or cells.***

KEYWORDS: Algorithms, Clustering, Grid-based, Hierarchical, K-means.

INTRODUCTION

Clustering is a method where a data set is replaced by a cluster or group of clusters, which are mainly collections of data points that "belong together" in some way. Because of sophisticated data collection methods, vast amounts of data are stored in different databases. The demand for grouping of important information and the extraction of useful data is increasing. Clustering consists in distributing data to groups of forms a crucial part with a grouping affinity but rather difference between them. Characterizing data into smaller clusters would probably results loss in some specifics when interpreting data[1].

The data objects are represented by lower cluster numbers and therefore data are modeled by its own clusters. Cluster analysis has been the arrangement of patterns (usually shown as a survey results suggest or as an item in a multidimensional space) in similarity-based clusters. Similar cluster patterns are closely linked to those in the following clusters. In the areas of engineering, machine learning, image processing, medicine, marketing , data mining, compression of such image or data, facts, etc., application with clustering approaches had already greatly increased.[2]. Types of clustering methods are illustrated in Fig. 1.

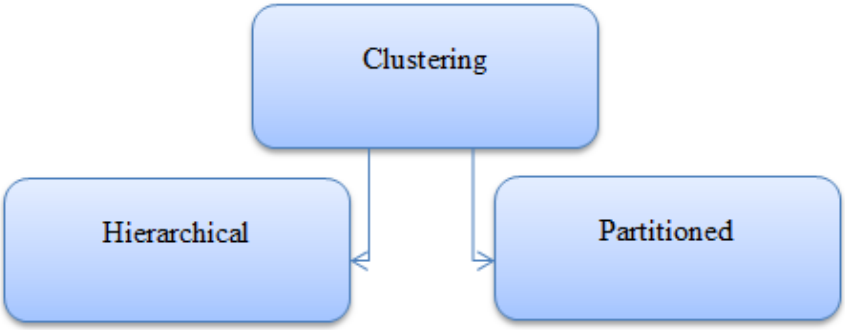

Fig. 1: Types of Clustering

The present clustering algorithms were developed by researchers, many of us designed new data clustering techniques and few of them tested and evaluated Numerous techniques of grouping. Clustering technology attempts to reach patterns in a cluster only by grouping people patterns into either groups or clusters so that the whole patterns even in the cluster become more like patterns of different patterns.[3]

METHODOLOGY

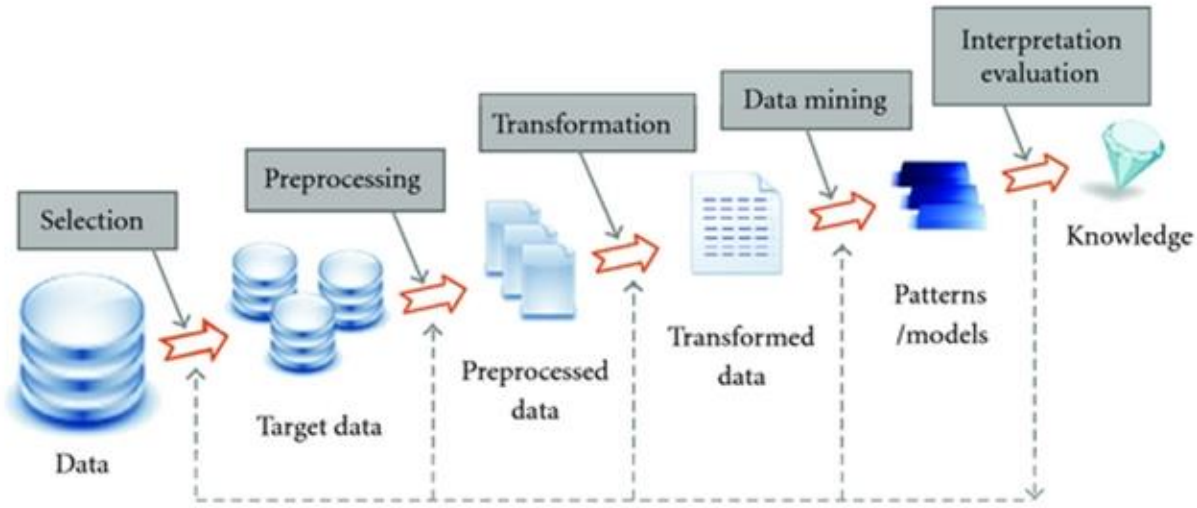

Fig. 2: Data Mining Process

The overall process of data mining is illustrated in Fig.2.

Hierarchical Clustering

Hierarchical clustering is a cluster analysis approach in which cluster hierarchy is generated to decompose the data objects in clusters on the basis of certain criteria. The clusters thus obtained in the hierarchy are known as genograms which show how the clusters are interrelated. There are two different ways to structure the hierarchy. This could be up or down. The large clusters are grouped into small clusters, together with individuals or small groups of large clusters. [4].

It measures the relationship between each point and also the dataframe and in fact takes and is integrated to generate statistics the best classification points. They are often known as one level or vector or the distance. measurement process is repeated. This process will continue until those nodes form a single group. [5][6].

The hierarchical clustering algorithm's step-wise procedure is given below (refer to Fig. 3):

A. Calculate the length matrix for each template pair. As one cluster, believe every sequence.

B. Use the data matrix to locate the most similar pair of clusters. Combine these related pattern pairs into one cluster.

C. If any point seems to be in a cluster, stop it and start through step B otherwise.

You chose the relational algorithm since:

• Resilience enclosed with respect to the level of granularity.

• Can tackle all forms of distances effectively.

• Applicable to almost all attribute types.

• Lot of adaptability.

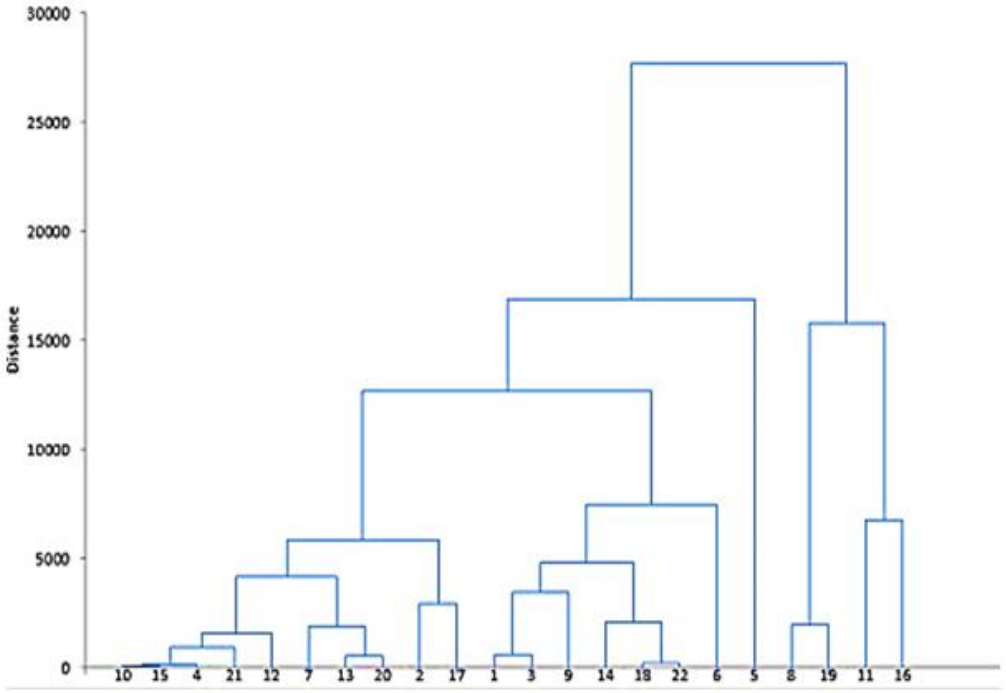

Fig. 3: Example of Hierarchical Clustering Genogram

Partitioned clustering

This follows an iterative optimisation method to minimize an objective function and tests the efficiency of clustering. The centroids of clusters are pretty reliable to minimize optimization problem dependent upon this criterion of optimality[7].

Partitioned approach is one of the clustering analysis techniques where a number of n entities are specified and even the original data is divided into cluster centres where k<=n minimizes an objective partitioning criterion and each cluster contains similar objects but is different from outside clusters objects. An example is shown in Fig. 4.

The resulting k clusters must meet the following two requirements.

- Every cluster must have at least one single object.
- Every object should contribute to a single cluster. K-means, methods are the most common methods of partitioning techniques.

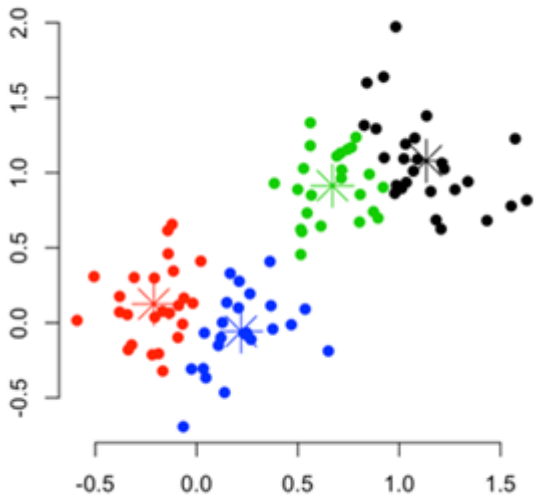

Fig. 4: Example of Partitioned clustering

Hierarchical Clustering Algorithm

Algorithms based on linkage:

For these types of algorithms, the linkage criterion is used. Graphic representation can be used for these types of algorithms. Single-link, average-link and complete link algorithm are mostly three types of algorithms used in this group. Single-link algorithm determines the minimum inter-cluster Automatic wire algorithm uses average category distance while the entire link brain processes a maximum distance as the cluster formation distance. Its distance here between closest values of two galaxies. Single-link algorithms can be used in density-based methods, while square error methods can use complete-link algorithms.

Algorithms based on quality criteria:

There are some disadvantages in pure hierarchical clustering algorithms; one of the disadvantages is that they cannot maintain their quality criteria. Once the various clusters are combined or divided into sub-clusters, they will not return to the previous state or swap objects between clusters. Therefore, if somewhere a merge or split operation is not applied correctly, it may result in poor cluster efficiency. One solution to this problem is the combination hierarchical clustering with other consistency preservation techniques.

There are some algorithms that are described below that can overcome this limitation:

- CURE (Clustering Using Representatives).
- ROCK (Robust Clustering using links).
- Leaders- Sub leaders.
- CHAMELEON Algorithm.
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchy).
- Linkage Algorithm.

Chart framework with self-organization:

Table 1: Analysis of All Clustering Algorithms

Algorithms	Dataset size	performance	Dataset type	Quality	Accuracy	Results	Time complexity
Hierarchical clustering algorithm	Large and tiny	Low at all times	Perfect and arbitrary	Good for tiny datasets	Better when cluster count increases	Good results for arbitrary datasets	O(n2logn)
Self-organization map algorithm	Large and tiny	Low as cluster count increases	Perfect and arbitrary dataset	Good for tiny datasets	High	Good results for arbitrary datasets	O(n)

The autonomy graph is essentially a deep neural network, resulting in a low, discrete statistical analysis. The self-organization map method takes a collaboration strategy to get an un-monitored method of coming to understand the computational models in the brain. In order to preserve data topology properties, this method calls for a neighborhood function. Unlike other neural nets, self-organizing maps often function in two separate ways. The course called the vector quantizing mode uses samples to construct the map. A new cluster is automatically sorted by mapping mode. The points in this algorithm become labelled with a two-dimensional map. A input space with the same size is issued for each position also as input field. [8][9]. A complete analysis is illustrated in Table 1.

CONCLUSION

In this paper, various types of clustering techniques and their algorithms have been discussed. Partition method is a way to separate data into multiple clusters containing similar objects. In hierarchical clustering, cluster hierarchy is created to decompose data objects.

Just after results are compared on the base with all listed variables, all methods allow almost same styles in terms software applications. The appropriate approach for cluster analysis on the basis of multiple factors can also be measured and contrasted with these algorithms.

REFERENCES

[1] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on Neural Networks*. 2005, doi: 10.1109/TNN.2005.845141.

[2] R. Rubina and P. Verma, “Various Techniques of Clustering: A Review,” *IOSR J. Comput. Eng.*, vol. 18, no. 05, pp. 23–28, 2016, doi: 10.9790/0661-1805032328.

[3] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *Journal of King Saud University - Computer and Information Sciences*. 2018, doi: 10.1016/j.jksuci.2017.06.001.

[4] D. T. Larose and C. D. Larose, “ Hierarchical and k -Means Clustering ,” in *Discovering Knowledge in Data*, 2014, pp. 209–227.

[5] L. Akoglu, H. Tong, and D. Koutra, “Graph based anomaly detection and description: A survey,” *Data Min. Knowl. Discov.*, 2015, doi: 10.1007/s10618-014-0365-y.

[6] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Adv. Artif. Intell.*, 2009, doi: 10.1155/2009/421425.

[7] C. C. Aggarwal, *Data classification: Algorithms and applications*. 2014.

[8] H. Cai, V. W. Zheng, and K. C. C. Chang, "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications," *IEEE Trans. Knowl. Data Eng.*, 2018, doi: 10.1109/TKDE.2018.2807452.

[9] P. Sarlin and Z. Yao, "Clustering of the self-organizing time map," *Neurocomputing*, vol. 121, pp. 317–327, 2013, doi: 10.1016/j.neucom.2013.04.007.