Survey on Hierarchical Clustering Algorithms in Data Mining

[1]T.Kousiga, [2]Dr.R.Shanmuga Vadivu

[1]Assistant Professor, [2]Assistant Professor
Department of Computer science,
P.S.G College of Arts and Science, Coimbatore, Tamil Nadu, India.

***Abstract*: Data mining is a process used to turn raw data into useful information. Various techniques and algorithms have been used for extracting meaningful information from large data sets. Clustering is one of the main techniques for analysis in data mining. It is a process of grouping similar data items together. There are many algorithms used in clustering. In this paper, a survey on hierarchical clustering algorithms and some related research works has been discussed.**

***Keywords:* Clustering, Hierarchical clustering algorithms, Agglomerative, Divisive.**

I. INTRODUCTION

Clustering is similar to classification, but the groups are not predefined. The groups of similar data items are called clusters. The grouping of data is accomplished by finding the similarities and measuring the distance between the data items. The most important aspect is that the similarity within the cluster must be higher than the similarity outside it. High quality clusters can be achieved by using various efficient algorithms. The traditional algorithms in clustering have been broadly classified into two major categories: (1) Hierarchical algorithms and (2) Partitional (non-hierarchical) algorithms. The further sections in this paper are based on the survey of hierarchical clustering algorithms.

II. HIERARCHICAL CLUSTERING ALGORITHMS

Hierarchical clustering algorithms are used to construct the hierarchical relationship among data items to form clusters. When the information on various levels of cluster structure is required, these algorithms work efficiently to interpret results. It merges various levels together in a sequential of steps. The result of hierarchical clustering can be graphically shown in a tree like structure called dendrogram.

Table-1. Traditional Hierarchical Algorithms

Hierarchical-Clustering Algorithms	Constructs the clusters by partitioning the data either in top-down or bottom-up approach in a recursive manner.
I. Agglomerative Algorithm	The object in each cluster is merged successively until obtaining the desired cluster structure.
II. Divisive Algorithm	All the objects in one cluster are divided into sub-clusters successively until obtaining the desired cluster structure.

Advantages:

- Do not require the number of clusters as initial parameters.
- Applicable to any type of data.
- Capable to handle the similarity and distance measures easily.

Limitations in traditional hierarchical clustering algorithms:

- Do not allow to separate overlapping clusters.
- Inability to adjust the clusters once split or merge decision has been executed.
- High computational complexity.
- Computationally time consuming.

To overcome the limitations in traditional hierarchical algorithms, some new algorithms have been developed and implemented in various research areas. They are mentioned below:

(i) CURE (Clustering Using Representatives) – [3] A combination of random sampling and partitioning of data set is used to handle large database and it create a balance between centroids and all data points.

(ii) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) – [3] This process begins with partitioning the objects hierarchically using tree structure and applies other clustering algorithms to refine clusters. The concept of CF-Tree (Clustering Feature Tree) was introduced in this algorithm in order to summarize cluster representations in a height-balanced tree.

(iii) ROCK (RObust Clustering using linKs) – [3] This algorithm employs links for merging the data points. It is well suited for clustering with Boolean and categorical attributes.

(iv) CHAMELEON – [3] This algorithm measures the similarity of two clusters based on dynamic model and the merging process facilitates homogenous clusters.

(v) Linkage Algorithms – [3] It merges the clusters based on distance between the clusters. Single Link finds the shortest distance, Average Link finds average distance and Complete Link finds the largestdistance between the clusters.

Table-2. New Hierarchical Clustering Algorithms

Hierarchical Algorithms	Advantages
(i) CURE (Clustering Using Representatives)	The quality of CURE algorithm is found to be better than other traditional algorithms.
(ii) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)	It was designed to reduce the repetition of I/O operations and much suitable for large databases. It is found that it can provide improved quality of clustering with few additional scans.
(iii) ROCK(RObust Clustering using linKs)	It is found more appropriate to handle large data sets and also exhibits good scalability property.
(iv) CHAMELEON	This algorithm is proven to be better for finding diverse shapes, sizes and density in two-dimensional space.
(v) Linkage Algorithms	The analysis of this algorithm can be performed to find which measure will be more appropriate for the particular field of application.

III. RELATED WORK

The improvements made in existing hierarchical algorithms in some research works are mentioned below:

In a survey of clustering on time-series data, the author explained about the purpose of agglomerative hierarchical algorithm in time-series data. [4] The hierarchical clustering is not restricted to cluster with equal length time-series data. It can also be applied to time-series with unequal length when appropriate distance measure like dynamic time warping (compares discrete sequences with continuous value sequences) is used to compute distance and similarity.

In 2014, [5] proposed a *Mean Gain Ratio (MGR)* for categorical data, which is a new information theory based hierarchical divisive clustering algorithm. MGR is used to implement clustering from attributes view point. The key idea of MGR algorithm is to select a clustering attribute and some equivalence classes from clustering attribute to share much information possible with other attribute partitions. It can be run with or without the specification of number of clusters. They experimental results on nine UCI real life datasets of this algorithm proved that it provides better performance and efficiency than other basic algorithms in clustering. The future work is that it can be applied to either balanced (Votes) or unbalanced class distribution (Zoo data set) and the advantages of MGR and G-ANMI algorithms can be combined to improve the accuracy of MGR.

In 2015, [6] performed hierarchical agglomerative cluster analysis for data reduction applied in psychological data by reviewing specific distance and linkage measures and comparing three linkage measures (single linkage, complete linkage, average linkage). The groups are created sequentially by merging the similar clusters together systematically by using a statistical technique. As a result, average linkage was compromising between single linkage and complete linkage and it was found to be more appropriate for bilingual data but it is sensitive to size and shape of clusters. Single linkage is sensitive to outliers and complete linkage breaks large clusters.

In 2015, [7] developed a novel computationally efficient HCA and HECA hierarchical clustering algorithms using grid and ensemble approach for segmenting multispectral images. The overlapping clusters, complex shaped clusters and clusters with different size and density can be separated effectively by using these algorithms. The advantages of grid and density approaches are combined in HCA algorithm which is based on CCA (Canonical Correlation Analysis) algorithm. The HECA (Hierarchical Ensemble Clustering Algorithm) was proposed by using ensemble method with the combination of different scale of information to overcome the instability of results in HCA algorithm. This algorithm is more effective up to eight wavelengths contained in multispectral images. The use of grid structure has some restrictions over dimension of the data being processed and making use of suitable algorithms to overcome this drawback can be considered for future enhancement.

In 2016, [8] introduced a citation based method called *EigenFactor Recommends (EFRec)* used for the improvement of scholarly navigation. The multiple scales of relevance for different users is made possible with the help of hierarchical structure of scientific knowledge used in this algorithm. The method was implemented and generated over millions of recommendations from millions of articles from various bibliographic databases. By using SSRN, an online A-B comparison was made and found that this approach offers much larger recommendation coverage and performs well with co-citation. In this paper, the ranking and hierarchical clustering aspects are used with EFrec. The extensions or variations from this method can be built on the general framework for further study of scholarly recommendation. If extended, this method will be helpful for dealing with big data.

IV. CONCLUSION

In this paper, the important aspects of hierarchical clustering algorithms are surveyed and some of the recent research studies in various applications based on these algorithms are discussed. The new algorithms are found to be better than traditional algorithms for handling complex datasets. Choosing an appropriate and efficient algorithm is purely based on the dataset or area of application which will be chosen by the researcher. If the nature of cluster and the desired output is previously unknown, then trial and error of various clustering techniques will be the best option. Hierarchical clustering algorithms can be used to discover clusters with different granularities and enables to find meaningful clusters at various levels. The research proposal of novel methods have proven that choosing a combination of two or more appropriate methods and algorithms will give computationally better performance, accurate results and efficiency with certain limitations instead of applying a single clustering technique. There are many other algorithms and methods related to hierarchical clustering which has not presented here.

REFERENCES

[1] Bharati M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305.

[2] Mythili .S, Madhiya .E, "An Analysis on Clustering Algorithms in Data Mining", IJCSMC, Vol. 3, Issue. 1, pg.334 – 340, January 2014.

[3] Yogita Rani and Dr. Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology, ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232.

[4] T.Warren Liao, "Clustering of time series data—a survey", Pattern Recognition 38 (2005) 1857 – 1874, The Journal of Pattern Recognition society.

[5] OdiliaYim , Kylee T. Ramdeen, "Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data", DOI:10.20982/tqmp.11.1.p008,vol. 11 no. 1, 2015.

[6] Hongwu Qin, Xiuqin Ma, Tutut Herawan, Jasni Mohamad Zain, "MGR: An information theory based hierarchical divisive clustering algorithm for categorical data", Knowledge-Based Systems 67 (2014) 401–411.

[7] I. A. Pestunov, S. A. Rylov, and V. B. Berikov, "Hierarchical Clustering Algorithms for Segmentation of Multispectral Images", ISSN 8756-6990, Optoelectronics, Instrumentation and Data Processing, Vol. 51, No. 4, pp. 1–10, 2015.

[8] Jevin D. West, Ian Wesley-Smith, and Carl T. Bergstrom, "A Recommendation System Based on Hierarchical Clustering of an Article-Level Citation Network", IEEE Transactions On Big Data, VOL. 2, NO. 2, APRIL-JUNE 2016.