RESEARCH ARTICLE OPEN ACCESS

A Recent Overview of Some Hierarchical Clustering Techniques

Kirti Sharma *, Hemant Verma**
**(Department of Computer Science, S.D.B.C.T Indore M.P. India*
*** (Department of Computer Science, Asst. Professor S.D.B.C.T Indore M.P. India*
Corresponding Auther : Kirti Sharma

ABSTRACT
The process of grouping a set of objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. The choice of clustering algorithm depends both on the type of data available and on the particular purpose of the application. Hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. In this paper we proposed a comparative study of some hierarchical clustering.

Keywords:- Cluster, Divisive Hierarchical, Agglomerative, , Bottom-up, Top-down

Date of Submission: 18-05-2018 Date of acceptance: 01-06-2018

I. INTRODUCTION

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis is an important human activity. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery.

II. MAJOR CLUSTERING METHODS

In general, the major clustering methods can be classified into the following categories.

1. Partitioning methods: Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k ≤ n.

2. Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster.

3. Density-based methods: Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes.

4. Grid-based methods: Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space).

III. TYPICAL REQUIREMENT OF CLUSTERING

The following are typical requirements of clustering in data mining

Scalability: Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.

Ability to deal with noisy data: Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

Incremental clustering: Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data.

High dimensionality: A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling two to three dimensions. Finding clusters of data objects in high dimensional space is challenging, especially considering that such data can be sparse and highly skewed.

IV. HIERARCHICAL METHODS

A hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

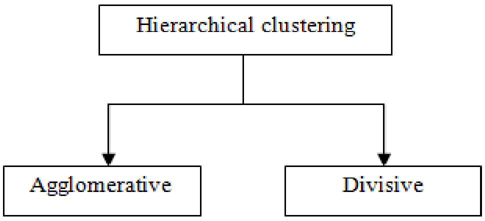

Figure 1. Types of hierarchical clustering

A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering

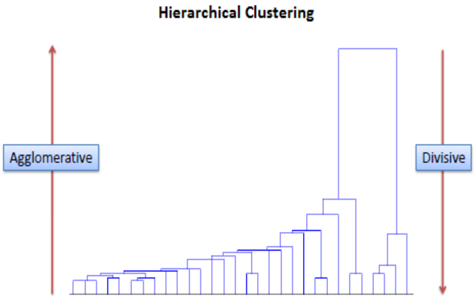

Figure 2. Working of hierarchical clustering

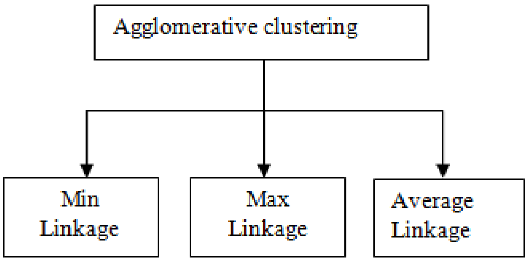

Figure 3. Types of agglomerative clustering

In max complete linkage hierarchical clustering, in each step merge two clusters whose merger has the smallest diameter. In single-link hierarchical clustering in each step merge two clusters whose two closest members have the smallest distance. Average-link clustering compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

V. LITERATURE REVIEW

In 2010 Revati Raman et al proposed “Fuzzy Clustering Technique for Numerical and Categorical dataset”. They presented a modified description of cluster center to overcome the numeric data only limitation of Fuzzy c-mean algorithm and provide a better characterization of clusters. The fuzzy k-modes algorithm for clustering categorical data. They proposed a new cost function and distance measure based on co-occurrence of values. [5]

In 2011 K. Ranjini proposed “Performance Analysis of Hierarchical Clustering Algorithm” They explain the implementation of agglomerative and divisive clustering algorithms by using various types of data. They implements and analysis running time of the algorithms using different linkages (agglomerative) to different types of data are taken for analysis[6].

In 2012 Dan Wei, Qingshan Jiang et al. proposed “A novel hierarchical clustering algorithm for gene Sequences” .The proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. They presented a novel approach for DNA sequence clustering based on a new sequence similarity measure, DMK, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. [7].

In 2013 K. Sasirekha, P. Baby proposed “Agglomerative Hierarchical Clustering Algorithm- A Review”. They showed that data mining hierarchical clustering method are used to build a hierarchy of clusters. They also show that hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [8].

In 2014 Archana Singh and Avantika Yadav proposed “Hybrid Approach of Hierarchical Clustering”. They proposed a hybrid approach of clustering based on AGNES and DIANA clustering algorithms, an extension to the standard hierarchical clustering algorithm. In the proposed algorithm, they used single linkage as a similarity measure. The proposed clustering algorithm provides more consistent clustered results from various sets of cluster centroids with tremendous efficiency [9].

In 2015Olga Tanaseichuk “An Efficient Hierarchical Clustering Algorithm for Large Datasets”. They show that Hierarchical clustering is a widely adopted unsupervised learning algorithm. Standard implementations of the exact algorithm for hierarchical clustering require O(n)2 time and O(n)2 memory and thus are unsuitable for processing datasets with large object. They present a hybrid hierarchical clustering algorithm requiring less time and memory [10].

In 2016 Amit Kumar Kar et al proposed “Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining”. They analyzes the four major clustering algorithms namely: Partitioning methods, Hierarchical methods, Grid-based methods and Density-based methods and comparing the performance of these algorithms on the basis of correctly class wise cluster building ability of algorithm[11].

In 2017 Shubhangi Pandit et al “An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis”. They present work a clustering technique and proposed using fuzzy c-means clustering algorithm for recognizing the text pattern from the huge data base. The proposed work is also committed to advance the approach of clustering for computing the hierarchical relationship among different data objects [12].

VI. COMPLEXITY AGGLOMERATIVE CLUSTERING

Min linkage

$$D_{sl}\left(C_i, C_j\right) = \min_{x,y}\left\{d(x,y) \middle| x \in C_i, y \in C_j\right\}$$

Max linkage

$$D_{cl}\left(C_i, C_j\right) = \max_{x,y}\left\{d(x,y) \middle| x \in C_i, y \in C_j\right\}$$

Average linkage

$$D_{avg}\left(C_i, C_j\right) = \frac{1}{\left|C_i\right| \times \left|C_j\right|} \sum_{x \in C_i, y \in C_j} d(x,y)$$

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its specific properties. So it is very difficult to decide which method is to best for select data set. The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single-link methods are more versatile

REFERENCES

[1]. J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.

[2]. Arun K. Pujari, Data mining Techniques, University Press (India) Private Limited, 2006.

[3]. D. Hand, H. Mannila, P. Smyth, “Principles of Data Mining, Prentice Hall of India, 2004

[4]. Nachiketa Sahoo “Incremental Hierarchical Clustering of Text Documents” May 5, 2006

[5]. Revati Raman Dewangan , Lokesh Kumar Sharma, Ajaya Kumar Akasapu Fuzzy Clustering Technique for Numerical and Categorical dataset International Journal on Computer Science and Engineering (IJCSE) NCICT 2010 Special Issue.

[6]. K. Ranjini Performance Analysis of Hierarchical Clustering Algorithm Performance Analysis of Hierarchical Clustering Algorithm” Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011).

[7]. Dan Wei, Qingshan Jiang et al. proposed “A novel hierarchical clustering algorithm for gene Sequences” 2012 Wei et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License

[8]. K.Sasirekha, P.Baby Agglomerative Hierarchical Clustering Algorithm- A Review International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153.

[9]. Archana Singh and Avantika Yadav “Hybrid Approach of Hierarchical Clustering”World Applied Sciences Journal 32 (7): 1181-1191, 2014 ISSN 1818-4952 © IDOSI Publications, 2014

[10]. Olga Tanaseichuk, Alireza Hadj “An Efficient Hierarchical Clustering Algorithm for Large Datasets” Austin J Proteomics Bioinform &

Genomics - Volume 2 Issue 1 - 2015 ISSN : 2471-0423

[11]. Amit Kumar Kar "A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining". ACEIT Conference Proceeding 2016

[12]. Shubhangi Pandit et al proposed " An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis" Volume 7, Issue 4, April 2017 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.`

Kirti Sharma "A Recent Overview of Some Hierarchical Clustering Techniques "International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.5, 2018, pp.47-50