International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 10 (2013), pp. 1115-1122

http://www. irphouse.com /ijict.htm

A Study of Hierarchical Clustering Algorithm

Yogita Rani[1] and Harish Rohil[2]

[1]*Department of Computer Science & Application, CDLU, Sirsa-125055, India.*
[2]*Department of Computer Science & Application, CDLU, Sirsa-125055, India.*

Abstract

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 these objects are very dissimilar to the objects that are in other clusters. Clustering methods are mainly divided into two groups: hierarchical and partitioning methods. Hierarchical clustering combine data objects into clusters, those clusters into larger clusters, and so forth, creating a hierarchy of clusters. In partitioning clustering methods various partitions are constructed and then evaluations of these partitions are performed by some criterion. This paper presents detailed discussion on some improved hierarchical clustering algorithms. In addition to this, author have given some criteria on the basis of which one can also determine the best among these mentioned algorithms.

Keywords: Hierarchical clustering; BIRCH; CURE; clusters ;data mining.

1. Introduction

Data mining allows us to extract knowledge from our historical data and predict outcomes of our future situations. Clustering is an important data mining task. It can be described as the process of organizing objects into groups whose members are similar in some way. Clustering can also be define as 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. Mainly clustering can be done by two methods: Hierarchical and Partitioning method [1].

In data mining hierarchical clustering works by grouping data objects into a tree of cluster. Hierarchical clustering methods can be further classified into agglomerative

and divisive hierarchical clustering. This classification depends on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual objects at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level or splitting a cluster from the next higher level. The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram. This tree graphically displays the merging process and the intermediate clusters. This graphical structure shows how points can be merged into a single cluster.

Hierarchical methods suffer from the fact that once we have performed either merge or split step, it can never be undone. This inflexibility is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such techniques cannot correct mistaken decisions that once have taken. There are two approaches that can help in improving the quality of hierarchical clustering: (1) Firstly to perform careful analysis of object linkages at each hierarchical partitioning or (2) By integrating hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters using another clustering method such as iterative relocation [2].

2. Related Work

Chris ding and Xiaofeng He, introduced the merging and splitting process in hierarchical clustering method. They provides a comprehensive analysis of selection methods and proposes several new methods that determine how to best select the next cluster for split or merge operation on cluster. The author performs extensive clustering experiments to test 8 selection methods, and found that the average similarity is the best method in divisive clustering and the Min-Max linkage is the best in agglomerative clustering. Cluster balance was a key factor there to achieve good performance. They also introduced the concept of objective function saturation and clustering target distance to effectively assess the quality of clustering [3].

Marjan Kuchakist et al. gives an overview of some specific hierarchical clustering algorithm. Firstly, author classified clustering algorithms, and then the main focused was on hierarchical clustering algorithms. One of the main purposes of describing these algorithms was to minimize disk I/O operations, consequently reducing time complexity. They have also declared attributes, disadvantages and advantages of all the considered algorithms. Finally, comparison between all of them was done according to their similarity and difference [4].

Tian Zhang et al. proposed al. proposed an agglomerative hierarchical clustering method named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and verified that it was especially suitable for large databases. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points so that best quality clusters can be produced with available resources. BIRCH can

typically produce a good cluster with a single scan of the data, and improve the quality further with a few additional scans of the data. BIRCH was also the first clustering algorithm proposed in the database area that can handle noise effectively. The author also evaluate BIRCH's time/space efficiency, data input order sensitivity, and cluster quality through several experiments [5]

Sudipto Guha et al. proposed a new hierarchical clustering algorithm called CURE that is stronger to outliers, and identifies clusters having non-spherical shapes and wide variances in size. This is achieved in CURE process by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. To handle large databases, CURE employs a combination of random sampling and partitioning. Along with the description of CURE algorithm, the author also described, type of features it uses, and why it uses different techniques [6].

3. Hierarchical Clustering Algorithms

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment, once a merge or split decision has been executed. Then it will neither undo what was done previously, nor perform object swapping between clusters. Thus merge or split decision, if not well chosen at some step, may lead to some-what low-quality clusters. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other techniques for multiple phase clustering. So in this paper, we describe a few improved hierarchical clustering algorithms that overcome the limitations that exist in pure hierarchical clustering algorithms.

3.1 CURE (Clustering Using REpresentatives)

CURE is an agglomerative hierarchical clustering algorithm that creates a balance between centroid and all point approaches. Basically CURE is a hierarchical clustering algorithm that uses partitioning of dataset. A combination of random sampling and partitioning is used here so that large database can be handled. In this process a random sample drawn from the dataset is first partitioned and then each partition is partially clustered. The partial clusters are then again clustered in a second pass to yield the desired clusters. It is confirmed by the experiments that the quality of clusters produced by CURE is much better than those found by other existing algorithms [6].

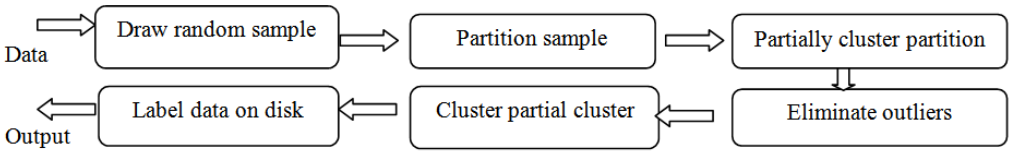

Figure 1: CURE Process. This diagram appears courtesy of [Guha, 2000].

Figure.1 shows how the CURE process is performed. Furthermore, it is demonstrated that random sampling and partitioning enable CURE to not only outperform other existing algorithms but also to scale well for large databases without sacrificing clustering quality. CURE is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the centre of the cluster by a specified fraction[8].

3.2 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH is an agglomerative hierarchical clustering algorithm and especially suitable for very large databases. This method has been designed so as to minimize the number of I/O operations. BIRCH process begins by partitioning objects hierarchically using tree structure and then applies other clustering algorithms to refine the clusters. It incrementally and dynamically clusters incoming data points and try to produce the best quality clustering with the available resources like available memory and time constraints. BIRCH process mainly takes four phases to produce best quality clusters

In this process two concepts are introduced, clustering feature and clustering feature tree (CF tree), which are used to summarize cluster representations. A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering. BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans [5].

3.3 ROCK (RObust Clustering using linKs)

ROCK is a robust agglomerative hierarchical-clustering algorithm based on the notion of links. It is also appropriate for handling large data sets. For merging data points, ROCK employs links between data points not the distance between them. ROCK algorithm is most suitable for clustering data that have boolean and categorical attributes. In this algorithm, cluster similarity is based on the number of points from different clusters that have neighbors in common. ROCK not only generate better quality cluster than traditional algorithm but also exhibit good scalability property [6].

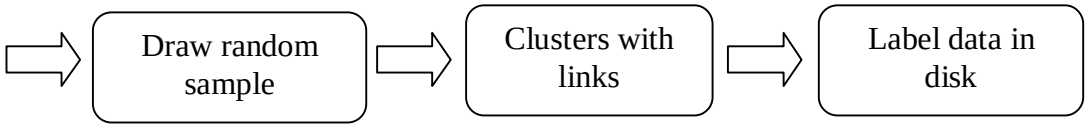

Figure 2: ROCK Process.

The steps involved in clustering using ROCK are described in figure 2. In this process after drawing random sample from the database, a hierarchical clustering algorithm that employs links is applied to sample data points. Finally the clusters involving only the sample points are used to assign the remaining data points on disk to the appropriate cluster.

3.4 CHEMELEON Algorithm

CHEMELEON is an agglomerative hierarchical clustering algorithm that uses dynamic modeling. It is a hierarchical algorithm that measures the similarity of two cluster based on dynamic model. The merging process using the dynamic model facilitates discovery of natural and homogeneous clusters. The methodology of dynamic modeling of clusters that is used in CHEMELEON is applicable to all types of data as long as a similarity matrix can be constructed.

The algorithm process mainly consist of two phases: firstly partitioning of data points are done to form sub-clusters, using a graph partitioning, after that have to do repeatedly merging of sub-clusters that come from previous stage to obtain final clusters. The algorithm is proven to find clusters of diverse shapes, densities, and sizes in two-dimensional space. CHEMELEON is an efficient algorithm that uses a dynamic model to obtain clusters of arbitrary shapes and arbitrary densities [7].

3.5 Linkage Algorithms

Linkage algorithms are agglomerative hierarchical methods that consider merging of clusters is based on distance between clusters. Three important types of linkage algorithms are Single-link(S-link), Average-link (Ave-link) and Complete-link (Com-link). In the Single-link, distance between two subsets is the shortest distance between them. In the Average-link, distance between two subsets is the average distance between them and in the Complete-link, distance between two subsets is the largest distance between them. **S**ingle-link is sensitive to the presence of outliers and the difficulty in dealing with severe differences in the density of clusters. On the other hand, displays total insensibility to shape and size of clusters .Average-linkage is sensitive to the shape and size of clusters. Thus, it can easily fail when clusters have complicated forms departing from the hyper spherical shape. Complete-linkage is not strongly affected by outliers, but can break large clusters, and has trouble with convex shapes [8].

3.6 Leaders–Subleaders

Leaders-Subleaders is an efficient hierarchical clustering algorithm that is suitable for large data sets. In order to generate a hierarchical structure for finding the subgroups or sub-clusters, incremental clustering principles is used within each cluster. Leaders–Subleaders is an extension of the leader algorithm. Leader algorithm can be described as an incremental algorithm in which L leaders each representing a cluster are generated using a suitable threshold value. There are mainly two major features of Leaders–Subleaders. First is effective clustering and second is prototype selection for pattern classification.

In this algorithm, after finding L leaders using the leader algorithm, the next step is to generate subleaders, also called the representatives of the sub clusters, within each cluster that is represented by a leader. This sub-cluster generation process is done by choosing a suitable sub threshold value. Subleaders in turn help in classifying the given new or test data more accurately. This procedure may be extended to more than

two levels. An h level hierarchical structure can be generated in only h database scans and is computationally less expensive compared to other hierarchical clustering algorithms [1].

3.7 Bisecting k-Means

Bisecting k-Means (BKMS) is a divisive hierarchical clustering algorithm. It was proposed by Steinbach et al. (2000) in the context of document clustering. Bisecting k-means always finds the partition with the highest overall similarity, which is calculated based on the pair wise similarity of all points in a cluster. This procedure will stop until the desired number of clusters is obtained. As reported, the bisecting k-means frequently outperforms the standard k-means and agglomerative clustering approaches. In addition, the bisecting k-means time complexity is O(nk) where n is the number of items and k is the number of clusters. Advantage of BKMS is low computational cost. BKMS is identified to have better performance than k-means (KMS) agglomerative hierarchical algorithms for clustering large documents [9].

4. Conclusion

This paper presents an overview of improved hierarchical clustering algorithm. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment, once a merge or split decision has been executed. This merge or split decision, if not well chosen at some step, may lead to some-what low-quality clusters. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other techniques for multiple phase clustering. These types of modified algorithm have been discussed in our paper in detail.

References

[1] Pavel Berkhin (2000), *Survey of Clustering Data Mining techniques* ,Accrue Software, Inc..

[2] Jiawei Han and Micheline Kamber (2006), *Data Mining: Concepts and Techniques*, The Morgan Kaufmann/Elsevier India.

[3] Chris ding and Xiaofeng He (2002), *Cluster Merging And Splitting In Hierarchical Clustering Algorithms.*

[4] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), *A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3,* pp.229- 240.

[5] Tian Zhang, Raghu Ramakrishnan, MironLinvy (1996), *BIRCH: an efficient data clustering method for large databases,* International Conference on Management of Data, In Proc. of 1996 ACM-SIGMOD Montreal, Quebec.

[6] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim (1998), *CURE: An Efficient Clustering Algorithm For Large*
[7] *Databases*, In Proc. of 1998 ACM-SIGMOD lnt. Conference on Management of Data.
[8] G. Karypis, E.H.Han and V.Kumar (1999), *CHAMELEON: Hierarchical clustering using dynamic modeling*, IEEE Computer, 32, pp. 68-75.
[9] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho (2007), *Improving Hierarchical Cluster Analysis: A new method with outlier detection and automatic clustering*, Chemo metrics and Intelligent Laboratory Systems, 87, pp. 208-217.
[10] L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu (2010), *A fast divisive clustering algorithm using an improved discrete particle swarm optimizer*, Pattern Recognition Letters, 31, pp. 1216-1225.