A Survey on Efficient Hierarchical Algorithm used in Clustering

G.Thilagavathi
PG Scholar
Sri Ramakrishna Engineering College
Coimbatore

D.Srivaishnavi
PG Scholar
Sri Ramakrishna Engineering College
Coimbatore

N.Aparna
PG Scholar
Sri Ramakrishna Engineering College
Coimbatore

Abstract

Clustering is the task of discovering homogenous group of objects. In data mining, the most important method of cluster analysis is hierarchical clustering which builds the hierarchy of clusters.

Hierarchical clustering outputs a hierarchy, a structure that is more informative than the unstructured set of clusters. Hierarchical clustering does not require us to prespecify the number of clusters and most hierarchical algorithms that have been used in Information Retrieval are deterministic.

In this paper, we discussed the hierarchical algorithms in which each one will have specific function with both its advantages and disadvantages.

Index Terms- Clustering, Hierarchical, Agglomerative, Divisive

1. Introduction:

Clustering is the partition of a set of objects into subsets (cluster) such that objects inside one cluster are similar to each other while objects from different clusters are not.

Data clustering is based on the similarity or dissimilarity (distance) measures between data points. Hence, these measures make the cluster analysis meaningful. The high quality of clustering is to obtain high intra-cluster similarity and low inter-cluster similarity as shown in Fig 1. In addition, when we use the dissimilarity (distance) concept, the latter sentence becomes:

The high quality of clustering is to obtain low intra-cluster dissimilarity and
high inter-cluster dissimilarity.

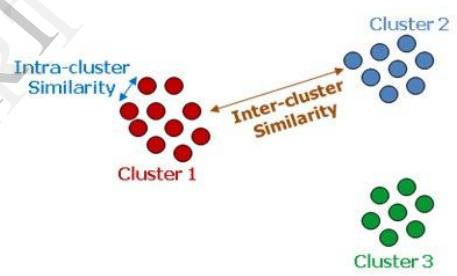

Fig:1 Intra-Cluster similarity and Inter-cluster Similarity.

Hierarchical algorithms create nested relationships of clusters which can be represented as a tree structure called dendrogram. Hierarchical algorithms can be divided into agglomerative and divisive hierarchical algorithms which is shown in Fig2.

1.1.Agglomerative Hierarchical Algorithm:

The bottom-up clustering method starts with each data point in a single cluster. Then it repeats merging the similar pairs of clusters until all of the data points

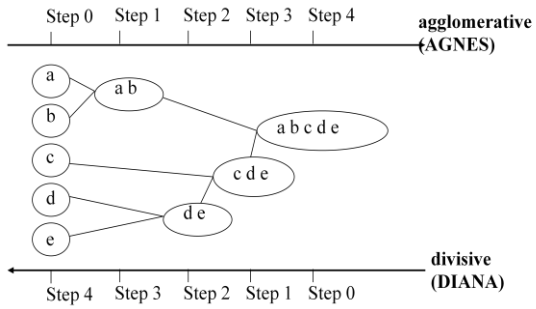

Fig:2 Agglomerative and Divisive Hierarchical Phase

are in one cluster, such as complete linkage clustering and single linkage clustering. CURE[2], ROCK[5], BIRCH[1] and Chameleon[6] are examples of this hierarchical algorithm.

1.2.Divisive Hierarchical Algorithm:

A top-down clustering method and is less commonly used and it reverses the operations of agglomerative clustering, it starts with all data points in one cluster and it repeats splitting large clusters into smaller ones until each data point belongs to a single cluster such as DIANA[2] clustering algorithm.

In this paper, we focus on different hierarchical algorithms. The rest of the paper is organized as follows: Section II defines the different Hierarchical algorithm with its cons and pro.

2. HIERARCHICAL ALGORITHMS:

2.1.BIRCH:

BIRCH [1] is best for finding noise and very effective to handle large data sets by making time and memory constraints explicit. Here data space is not uniformly occupied and hence not every data point is equally important for clustering purposes. It constructs a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering.

Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)

Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree

Compared with prior distance-based algorithms, BIRCH[1] is incremental in the sense that clustering decisions are made without scanning all data points or all currently existing clusters. When compared with other probability-based algorithms, BIRCH tries to make the best use of the available memory to derive the finest possible sub clusters to ensure accuracy and it minimizes the I/O costs to ensure efficiency by organizing the clustering and reducing process using an in-memory balanced tree structure of bounded size. BIRCH has been used for filtering real image.

2.1.1.Disadvantages:

Handles only numerical data.

Sensitive to order of data records.

2.2. CLUBS:

CLUBS (for Clustering Using Binary Splitting) is the simplest hierarchical algorithm. It consists of both divisive and agglomerative phase. During these two phases, by using a least quadratic distance criterion possessing unique analytical properties, the samples are repartitioned. Without requiring input from users, it derives good clusters.

Phase 1: In divisive, Original data set is split recursively into mini clusters through successive binary splits.

Phase 2: In Agglomerative, mini clusters are recombined to final result.

2.2.1.Advantages:

It is robust and impervious to noise.

It provides better speed and accuracy when compared with other methods such as BIRCH, k-means etc.

2.3. CURE:

CURE (Clustering Using Representatives) is a non-spherical cluster and uses multiple representative points to evaluate the distance between clusters and stops the creation of a cluster hierarchy if a level consists of *k* clusters. Representative points are found by selecting a constant number of points from a cluster and then "shrinking" them toward the center of the cluster. Cluster similarity is the similarity of the closest pair of representative points from different clusters. Shrinking representative a point toward the center helps avoid problems with noise and outliers. CURE [2] is better able to handle clusters of arbitrary shapes and sizes. The algorithm cannot be directly applied to large databases. So for this purpose we do the following enhancements:

1. Random sampling
2. Partitioning for speed up
3. Labeling data on disk

2.3.1.Disadvantages:

CURE cannot handle differing densities.

By using the enhancements mentioned above to use this algorithm to large databases.

2.4. ROCK:

ROCK (Robust Clustering using Links) is a hierarchal clustering algorithm to handle the data with categorical and Boolean attributes. A pair of points is defined to be neighbors if their similarity is greater than some threshold. It handles large number of data and it reduces complexity. Clusters are generated by the sample points.With appropriate sample size, the quality of clustering is not affected. ROCK [4]performs well on real categorical data, and respectably on time-series data.

2.4.1Advantages:

Run on real & synthetic data sets. Real data used for comparison to traditional algorithms.Synthetic data used to demonstrate scalability.

2.5.Chameleon (Clustering Using Dynamic Modeling):

Adapt to the characteristics of the data set to find the natural clusters. Use a dynamic model to measure the similarity between clusters. Main property is the relative closeness and relative inter-connectivity of the cluster. Two clusters are combined if the resulting cluster shares certain properties with the constituent clusters. The merging scheme preserves self-similarity. One of the areas of application is spatial data. Chameleon is a two-phase approach.

Phase 1: Uses a graph partitioning algorithm to divide the data set into a set of individual clusters

Phase 2: it uses an agglomerative hierarchical mining algorithm to merge the clusters.

2.6.DIVCLUS-T:

DIVCLUS-T is a divisive hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. In the divisive hierarchical

clustering algorithm, one recursively splits a cluster into two sub-clusters, starting from the set of objects Ω={1,…,n}: given the current partition Pk=(C1,…,Ck), one cluster Cℓ is split in order to find a partition Pk+1 which contains k+1 clusters and optimizes the chosen adequacy measure, based on the inertia criterion. More precisely, at each stage, the divisive hierarchical clustering method DIVCLUS-T[7]:

Phase 1: splits a cluster Cℓ into a bipartition $(A_\ell, \bar{A}_\ell)$ of minimum within-cluster inertia.

Phase 2: chooses in the partition Pk the cluster Cℓ to be split in such a way that the new partition Pk+1 has minimum within-cluster inertia.

2.6.1.Advantages:

DIVCLUS-T is designed for either numerical or categorical data. It provides a simple and natural interpretation of the clusters.

3. Problems and Limitations:

The major problems which occur common in Hierarchical clustering algorithms are no objective function is directly minimized, sensitive to noise and outliers, Difficulty in handling different sized clusters and convex shapes and difficulty in breaking large clusters. In this Birch [1] algorithm is used to handle noise and large dataset. Cure [2] algorithm is used to handle the different sized clusters. Rock [4] and DIVCLUS-T [7] are used to handle the categorical attribute and Chameleon to find the similarity between clusters.

4. Conclusion:

The survey on clustering in data mining has been discussed with different perspectives. The functionalities of different algorithms used in Hierarchal clustering have been explained along with their advantages and disadvantages.

References:

[1] V.S.Jagadeeswaran and P.Uma, ”Detection of noise by efficient hierarchical birch algorithm for large data sets”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2, February 2013

[2] Pavel Berkhin,” Survey of Clustering Data Mining Techniques” Accrue Software, Inc.

[3] Osama Abu Abbas,”Comparison between different clustering algorithms”, The International Arab Journal of Information Technology,Vol 5,No 3,July 2008.

[4] Sudipto Guha,Rajeev Rastogi,”A Robust Clustering Algorithm for Categorical attributes”Information System Vol 25,No 5,Elsevier Science Limited.

[5]V. S. Alves, R. J. G. B. Campello, E. R. Hruschka, “Towards a Fast Evolutionary Algorithm for Clustering”,In Proc. IEEE Congress on Evolutionary Computation,pp. 6240-6247, 2006.

[6] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Elsevier Publication.

[7]MarieChavent, YvesLechevallier and OlivierBriant, ” DIVCLUS-T: A monothetic divisive hierarchical clustering method” Computational Statistics & Data Analysis, 2007, vol. 52.