Hierarchical Clustering

Definition:

A hierarchical clustering method works by grouping data objects into a hierarchy or tree of clusters. Representing data objects in this form is useful for data summarization and visualization. Hierarchical clustering methods can be either **agglomerative** or **divisive**:

Agglomerative (bottom up merging strategy):

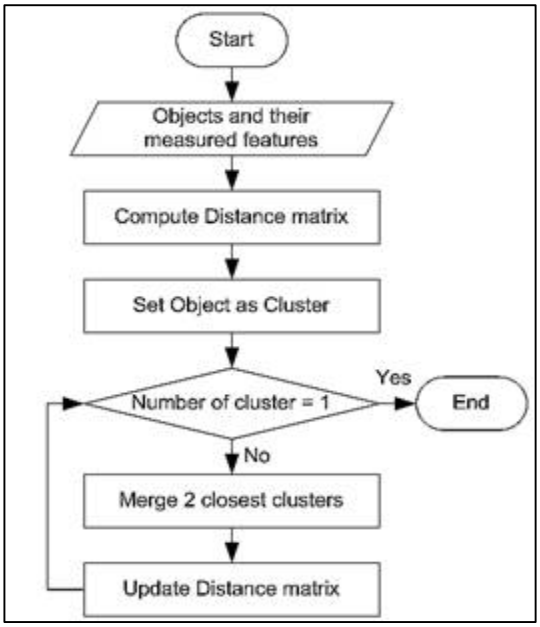
- Initially, all data objects are assigned their own cluster.
- These clusters are iteratively merged to form larger clusters based on a distance measure (similarity/dissimilarity measure) between the clusters.
- This process recursively continues until all the data objects are in a single cluster or certain termination conditions are satisfied. The single cluster becomes the hierarchy's root.
- Because two clusters are merged per iteration, where each cluster contains at least one object, an agglomerative method requires at most n iterations (n is the number of data objects).

Divisive (top down splitting strategy):

- Initially, all the data objects are assumed to be in a single cluster.
- This cluster is iteratively split into smaller least-similar clusters based on a distance measure (similarity/dissimilarity measure) between the clusters.
- The partitioning process continues until each cluster at the lowest level either contains only one data object or the objects within a cluster are sufficiently similar to each other.
- Since once cluster is split into 2 parts per iteration, the divisive method requires at most n iterations (n is the number of data objects).

Dendrogram: A tree structure called a dendrogram is commonly used to represent the hierarchical structure of the clustered data. It shows how objects are grouped together (in an agglomerative method) or partitioned (in a divisive method) step-by-step. The height of a dendrogram corresponds to the distance between two cluster.

The process-flow of Agglomerative hierarchical clustering:



Input Parameter to the Algorithm:

The distance measure is required to be mentioned as an input parameter to the hierarchical clustering algorithm. The distance measure measures the distance between two clusters. Three popular measures are discussed below:

Single Linkage

- The similarity of the closest pair of data points belonging to different clusters is found & these data objects are then grouped as a new cluster.
- The distance between the two clusters here is the distance between the nearest points in the different clusters.
- Single linkage generally provides non-convex cluster.

Complete Linkage

- The distance measure between two clusters here is the the maximum distance between data objects of the the nearest clusters.
- The resulting cluster is formed based on the minimum distance of the maximum distance measures calculated.
- Complete Linkage generally produces clusters that are spherical (in shape).

Average Linkage

- The distance measure between the two clusters is the average distance between each point in one cluster to every point in the other cluster.
- The average distance is advantageous in that it can handle categorical as well as numeric data
- Average linkage is less affected by outliers.

Disadvantage:

- Hierarchical clustering methods can encounter difficulties regarding the selection of merge or split points. Thus, merge or split decisions, if not well chosen, may lead to low-quality clusters.
- Hierarchical clustering methods tend to be overly sensitive to outliers or noisy data depending on the distance measure of choice.

Notes:

- There are several orthogonal ways to categorize hierarchical clustering methods. For instance, they may be categorized into algorithmic methods, probabilistic methods, and Bayesian methods. Agglomerative, divisive, and multiphase methods are algorithmic methods.
- An agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a minimal spanning tree algorithm.
- The dendrogram is cut at different levels to get few bigger clusters or many smaller clusters.
- A challenge with divisive methods is how to partition a large cluster into several smaller ones. If the cluster is very very large, it is computationally prohibitive to examine all possibilities. Due to such challenges in divisive method, agglomerative methods are used more.
- A good way to evaluate the cluster is to assess whether a non-random structure exists in the data after clustering is performed.
- It is desirable to estimate the ideal number of clusters even before a clustering algorithm is used. This can significantly help evaluate the cluster.
- The quality of hierarchical agglomeration can be improved by analyzing the object linkages at each hierarchical partitioning (e.g., in Chameleon Clustering), or by first performing micro-clustering and then operating on the micro- clusters with other clustering techniques such as iterative relocation (as in BIRCH).

Hierarchical Clustering implementation in R:
https://github.com/DataAstrologer/DataScience/blob/master/R/Machine%20Learning%20Algorithm%20Templates/Unsupervised%20Learning/Clustering/Hierarchical%20Clustering.Rmd

Hierarchical Clustering implementation in Python:
https://github.com/DataAstrologer/DataScience/blob/master/Python/Machine%20Learning%20Algorithm%20Templates/Unsupervised/Clustering/Hierarchical%20Clustering.ipynb