internship studio

Welcome to

INTERNSHIP STUDIO

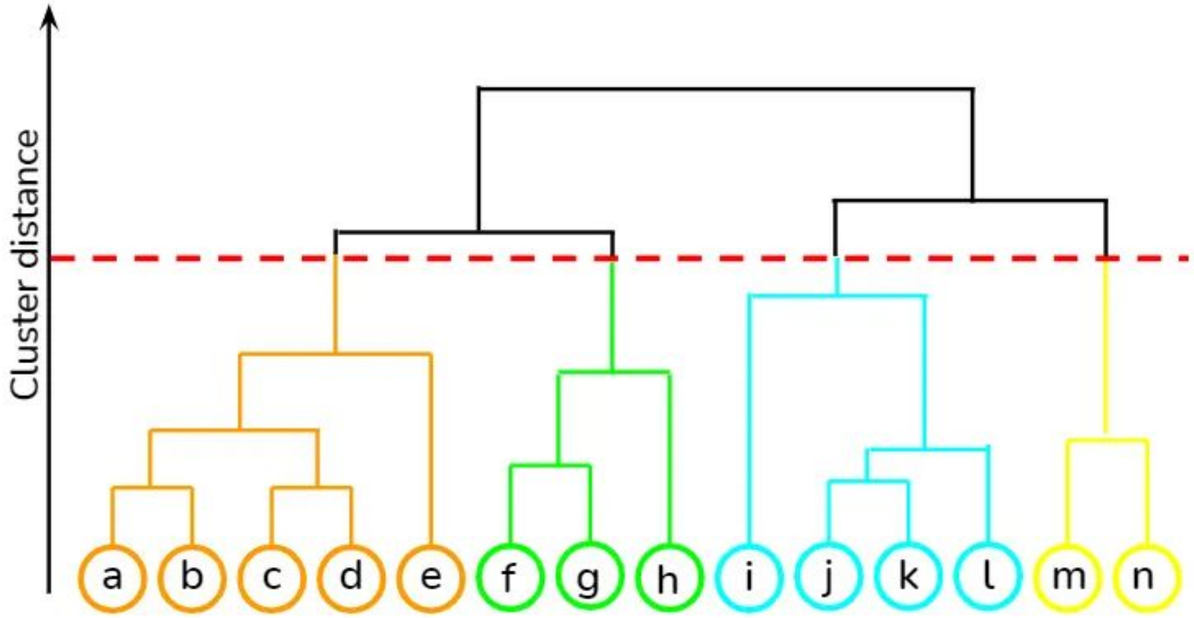
Module 04 | Lesson 05

Hierarchical clustering

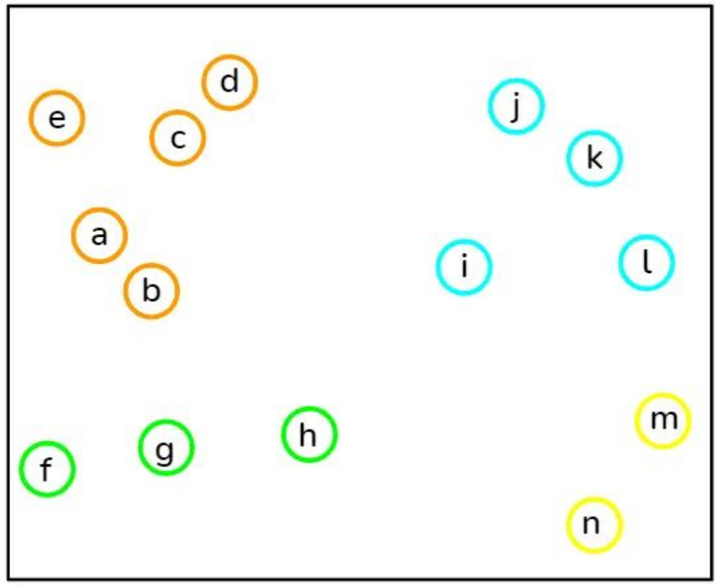
Hierarchical Clustering

A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.



Dendrogram



Applications

There are many real-life applications of Hierarchical clustering. They include:

Bioinformatics: grouping animals according to their biological features to reconstruct phylogeny trees.

Business: dividing customers into segments or forming a hierarchy of employees based on salary.

Image processing: grouping handwritten characters in text recognition based on the similarity of the character shapes.

Information Retrieval: categorizing search results based on the query.

Hierarchical Clustering Types

There are two main types of hierarchical clustering:

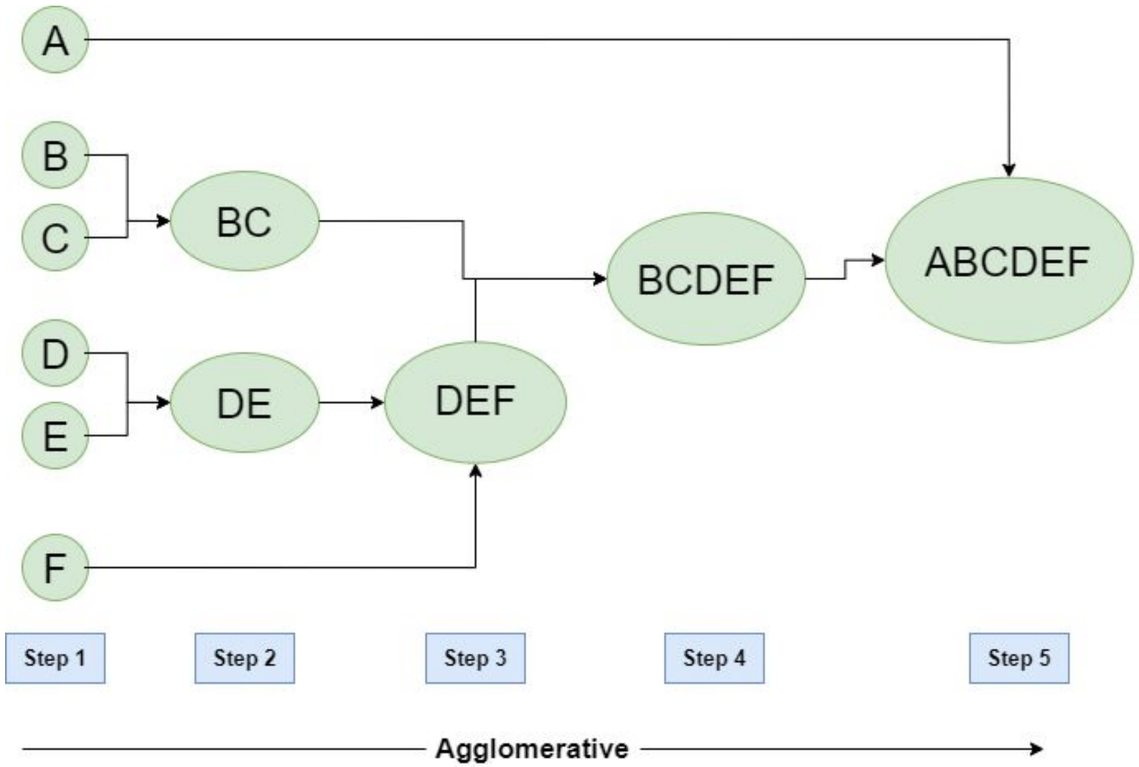
Agglomerative: Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

The algorithm for Agglomerative Hierarchical Clustering is:

1. Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
2. Consider every data point as an individual cluster
3. Merge the clusters which are highly similar or close to each other.
4. Recalculate the proximity matrix for each cluster
5. Repeat Steps 3 and 4 until only a single cluster remains.

Hierarchical Clustering Types

Agglomerative:



Step-1: Consider each alphabet as a single cluster and calculate the distance of one cluster from all the other clusters.

Step-2: In the second step comparable clusters are merged together to form a single cluster. Let's say cluster (B) and cluster (C) are very similar to each other therefore we merge them in the second step similarly to cluster (D) and (E) and at last, we get the clusters [(A), (BC), (DE), (F)]

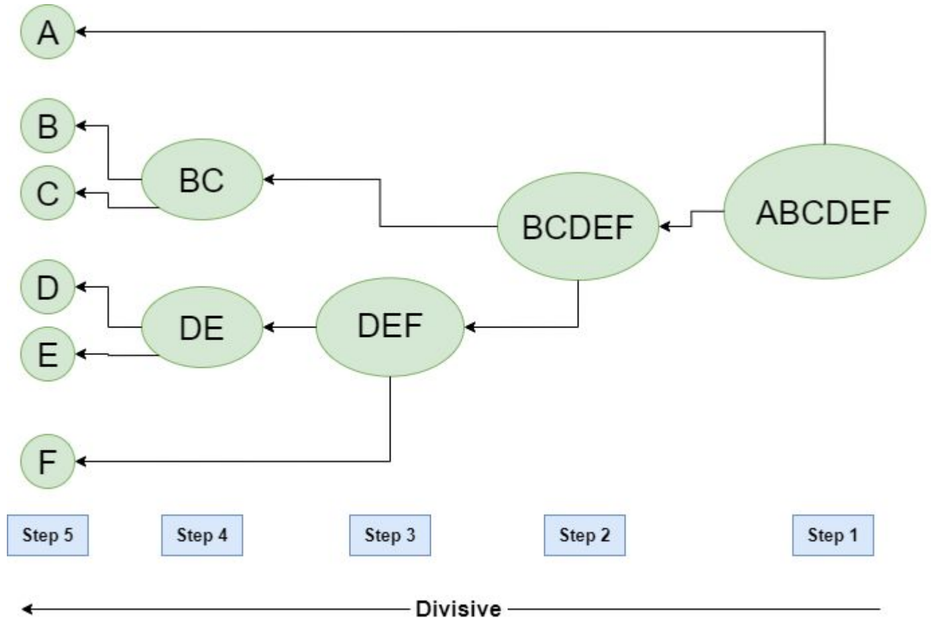
Step-3: We recalculate the proximity according to the algorithm and merge the two nearest clusters([(DE), (F)]) together to form new clusters as [(A), (BC), (DEF)]

Step-4: Repeating the same process; The clusters DEF and BC are comparable and merged together to form a new cluster. We're now left with clusters [(A), (BCDEF)].

Step-5: At last the two remaining clusters are merged together to form a single cluster [(ABCDEF)].

Hierarchical Clustering

Divisive: We can say that Divisive Hierarchical clustering is precisely the opposite of Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable. In the end, we are left with N clusters.



Advantages

- ✔ The ability to handle non-convex clusters and clusters of different sizes and densities.
- ✔ The ability to handle missing data and noisy data.
- ✔ The ability to reveal the hierarchical structure of the data, which can be useful for understanding the relationships among the clusters.

Disadvantages

- ✔ The need for a criterion to stop the clustering process and determine the final number of clusters.
- ✔ The computational cost and memory requirements of the method can be high, especially for large datasets.
- ✔ The results can be sensitive to the initial conditions, linkage criterion, and distance metric used.