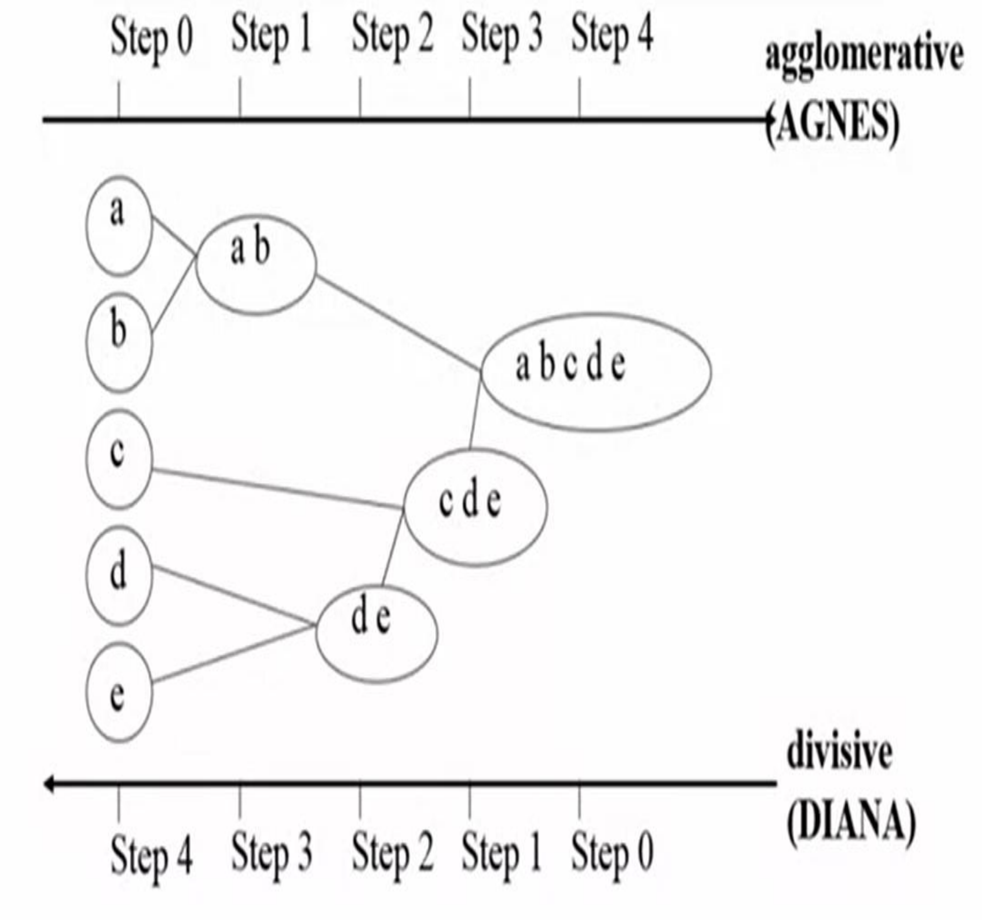
Hierarchical Clustering

CSED, TIET

Hierarchical Clustering-Introduction

- Hierarchical Clustering generates a clustering hierarchy (drawn as a dendrogram).
- It is better than partitioning methods as:
  1. We need not to specify the number of clusters K in advance.
  2. It does not require iterative refinement.
  3. It is more-deterministic.
- Two categories of algorithms:
  1. Agglomerative (AGNES- Agglomerative Nesting)-start with singleton clustering, continuously merge two clusters at a time to build a bottom-up hierarchy of clusters.
  2. Divisive (DIANA- Divisive Analysis)-start with a huge macro cluster, split it into two continuous clusters at a time generating a top-up hierarchy of clusters.



Agglomerative Clustering

- Among clustering based techniques, Hierarchical agglomerative clustering is commonly used due to:
  - its ability to create natural clusters,
  - possibility to view data at different threshold levels
  - and no prior knowledge of number of clusters.
- Agglomerative clustering works on the distance matrix (of order nxn) instead of original feature space (of order nxk)
- The hierarchical agglomerative clustering algorithm begins by considering each word as a separate cluster, so for N words in a class, there are N clusters.
- At each next step, the algorithm merges the most similar clusters and updates the distance between the new cluster and each old cluster.
- The process continues until the distance between the clusters is less than the pre-defined threshold value.

Agglomerative Clustering-Algorithm

Input: Distance matrix (containing distance between n-objects) of order nxn.

Output: Hierarchy of clusters.

1. For Distance matrix Di, do
2. Find the pair of objects with the least distance i.e. least dissimilar pair and merge both the objects into a single cluster.
3. update distance matrix by calculating the distance between old and new clusters using some linkage rules.
4. repeat steps 2 and 3 until all the distances in the distance matrix are less than a specified threshold θ.

Linkage Rules

- At each fusion stage of clusters, a linkage method decides how the inter-cluster distances are calculated.
- A number of linkage rules are described in the literature:

1. Single-linkage
2. Complete-linkage
3. Average-Linkage
4. Centroid-linkage
5. Ward-linkage

Single-Linkage

- In single method, the distance between two clusters is equal to the least distance between an object in one cluster and an object in another cluster.

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X, Y) = min_{x \epsilon X, y \in Y} d(x, y)$$

- The single linkage (also called nearest neighbor or minimum) method employs a friends of friends clustering method.
- The single linkage clustering method tends to produce long and elongated chain-like clusters when the objects are quite close to each other.
- The clusters produced by the single linkage method are internally heterogeneous, but the objective of clustering is to produce homogeneous clusters

Complete-Linkage

- In the Complete linkage (also called farthest neighbor or maximum) method, the distance between the clusters is equal to the greatest distance between the object in one cluster and an object in another cluster.

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X,Y) = max_{x \epsilon X, y \in Y} d(x,y)$$

- The complete linkage method tends to produce more balanced and compact clusters with equal diameters as compared to the single linkage method.

Average-Linkage

- The average linkage method is an attractive trade-off between single and complete linkage methods where the distances are averaged at every amalgamation step.
- In the average linkage method, the distance between two clusters is equal to the average distance between each object in one cluster to every object in the other cluster.

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X,Y) = \frac{1}{|X||Y|}\sum_{x \epsilon X, y \in Y} d(x,y)$$

- The average linkage method employs a more central measure as compared to single and complete linkage methods (as they use single pairwise distance).
- But it is computationally quite expensive.

Centroid Linkage

- In the centroid linkage method, the distance between two clusters is equal to the distance between each centroids (means) of both the clusters.

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X,Y) = d(C_X, C_y)$$

Where Cx and Cy denote mean of data points of cluster X and cluster Y respectively.

- We can also use Grouped Average Agglomerative Clustering (GAAC) which is the weighted average of the distance between centroids of two clusters.

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X,Y) = \frac{N_x C_X + N_y C_y}{N_x + N_y}$$

where N_x and N_x are number of objects in cluster X and Y.

Ward-Linkage

- Ward linkage rule minimizes the total within cluster variance.
- The distance between two clusters X and Y is updated as:

$$Distance\ between\ cluster\ X\ and\ cluster\ Y = d(X, Y) = \frac{N_x N_y}{N_x + N_y} d(C_X, C_y)$$

Agglomerative Clustering -Numerical Example

Consider a distance matrix between seven data points. Use Average –linkage based agglomerative clustering to find the clusters (until all clusters are grouped together).

	A	B	C	D	E	F	G
A	0						
B	0.25	0					
C	0.1428	0.3333	0				
D	0.1428	0.3333	0.25	0			
E	0.5	0.6	0.375	0.375	0		
F	0.5	0.6	0.375	0.375	0	0	
G	0.5555	0.6363	0.4444	0.4444	0.1428	0.1428	0

Solution

Now, the objects E and F will be grouped together as they have minimum distance.

The distance of this new cluster will be updated with all old clusters.

For instance $d\big((E,F),A\big) = \frac{d(E,A)+d(F,A)}{2\times 1} = \frac{0.5+0.5}{2} = 0.5$

	A	B	C	D	E,F	G
A	0					
B	0.25	0				
C	0.1428	0.3333	0			
D	0.1428	0.3333	0.375	0		
E,F	0.5	0.6	0.375	0.375	0	
G	0.5555	0.6363	0.4444	0.4444	0.1428	0

Solution-Contd…

Now, two new clusters will be formed

- E,F and G
- A and C

	A,C	B	D	E,F,G
A,C	0			
B	0.2916	0		
D	0.1964	0.3333	0	
E,F,G	0.2749	0.6121	0.3981	0

Solution-Contd…

Now, one new clusters will be formed

- A,C and D

	A,C,D	B	E,F,G
A,C,D	0		
B	0.3055	0	
E,F,G	0.4382	0.6121	0

Solution-Contd…

Now, one new clusters will be formed

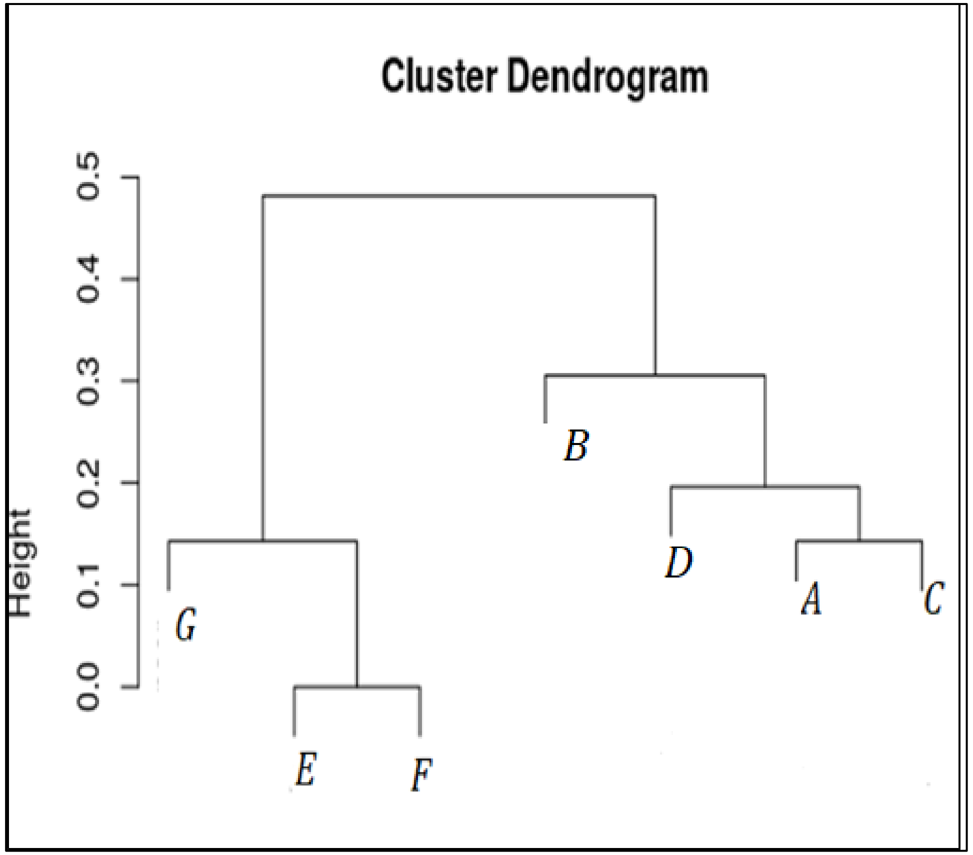
- A,C, D and B

	A,C,D,B	E,F,G
A,C,D,B	0	
E,F,G	0.4817	0

- In the next step all the objects will be clustered together.

Solution-Contd…

- Clustering dendrogram is used to find the optimal distance threshold. i.e. the distance above which the agglomerative clustering should stop further grouping clusters.
- The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram.
- The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.



Divisive Clustering

- DIANA (Divisive Analysis) is developed by Kaufmann and Rousseeuw in 1990
- It is inverse of Agglomerative Clustering.
- Divisive clustering is a top-down approach.
  - The process starts at the root with all the points as one cluster.
  - It recursively splits the high level clusters to build the dendrogram.
  - Can be considered as global approach.
  - More efficient than agglomerative clusters for smaller datasets.
  - For large datasets, it is computationally quite expensive (of the order of $O(n^2)$)

Divisive Clustering (Contd....)

- Choosing which cluster to split
  - Check the sum of squares errors of the clusters and choose the one with the largest value.
- Splitting Criterion: determine how to split
  - We may find the split that result in greater reduction in SSE as a result of split.