Cluster Analysis

- Hierarchical Clustering
  - AGNES
  - DIANA

1

Hierarchical Clustering

- Use distance matrix as clustering criteria
- This method does not require the number of clusters ***k*** as an input
- Hierarchical methods can be

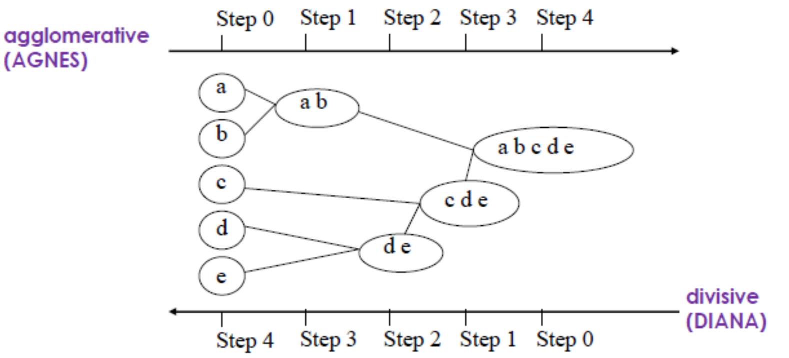
  Agglomerative: bottom-up approach

  Divisive: top-down approach

2

Hierarchical Clustering

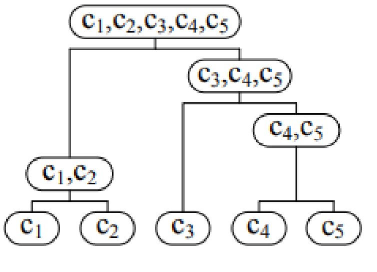
- Agglomerative and divisive algorithms on a data set of five objects **{a, b, c, d, e}**

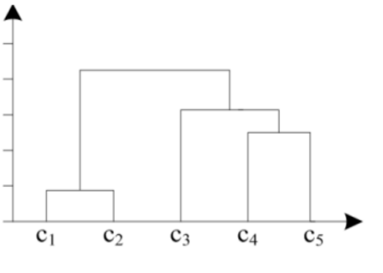


3

Dendogram

- Decompose data objects into a several levels of nested partitioning (<u>tree</u> of clusters), called a <u>dendrogram</u>.





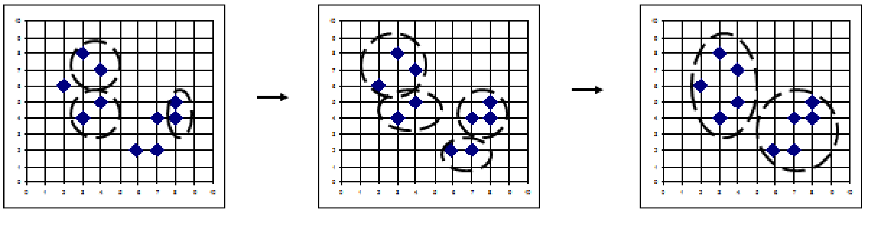
4

Dendogram

- The root node of the dendrogram represents the whole data set, each leaf node is regarded as a data point.
- The intermediate nodes describe the extent to which the objects are proximal to each other.
- The height of the dendrogram expresses the distance between each pair of data points or clusters, or a data point and a cluster.
- The clustering results can be obtained by cutting the dendrogram at different levels

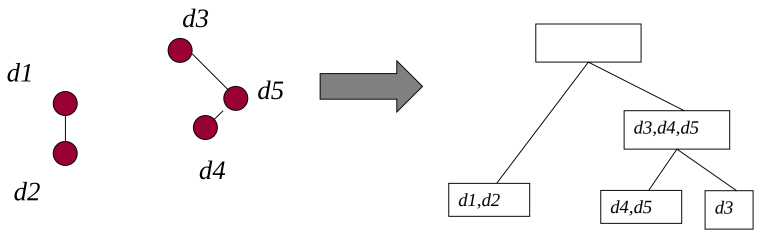
AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Start with each document being a single cluster.
- Merge nodes that have the least dissimilarity
- Eventually all documents belong to the same cluster.



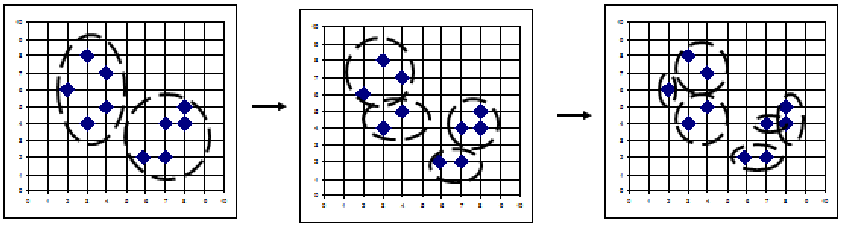
AGNES (Agglomerative Nesting)

- As clusters *agglomerate*, docs likely to fall into a hierarchy of "topics" or concepts.



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Inverse order of AGNES
- Start with all documents belong to the same cluster.
- Eventually each node forms a cluster on its own.



DIANA (Divisive Analysis)

- For a data set with N objects, a divisive hierarchical algorithm would start by considering $2^{n-1} - 1$ possible divisions of the data into two nonempty subsets, which is computationally expensive even for small-scale data sets.
- Therefore, divisive clustering is not a common choice in practice.
- It is less likely to suffer from the accumulated erroneous decisions, which cannot be corrected by the successive process.

DIANA (Divisive Analysis)

- Suppose that cluster C_l is going to be split into clusters C_i and C_j :
  - Start with $C_i = C_l$ and C_j empty
  - For each data object x_m in C_i :
    - (1) For the first iteration, compute its average distance to all the other objects
    - (2) For the remaining iterations, compute the difference between the average distance to C_i and the average distance to C_j

DIANA (Divisive Analysis)

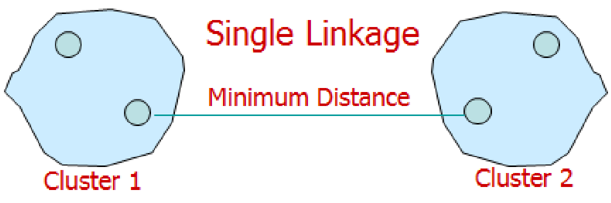
- Suppose that cluster C_l is going to be split into clusters C_i and C_j :
  - Decide whether to move element x_m to cluster C_j or keep it in cluster C_i :
    - For the first iteration, move the object with the maximum value to C_j (that is, the element farther from every other element is separated from the others)
    - For the remaining iterations, if the maximum difference value is greater than zero, move the data object with the maximum difference into C_j j , then repeat steps 1 and 2. If the maximum value is less than zero, stop.

Distance Between Clusters

- **Single-link:** Similarity of the *most* cosine-similar.
- **Complete-link:** Similarity of the "furthest" points, the *least* cosine-similar.
- **Average-link:** Average cosine between pairs of elements.
- **Centroid:** Clusters whose centroids (centers of gravity) are the most cosine-similar.
- **Ward's procedure:** For each cluster, the sum of squares is calculated.

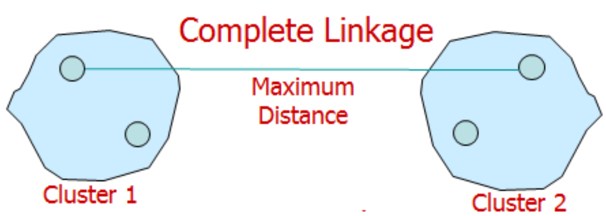
Single Linkage

- The **single linkage** method is based on minimum distance, or the nearest neighbor rule.
- $min\ d(i,j)$
- $d_{k(i,j)} = \min(d_{ki}, d_{kj})$



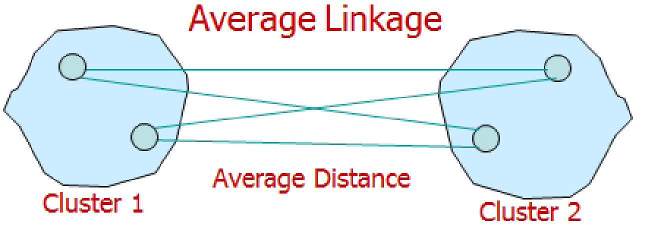
Complete Linkage

- The **complete linkage** method is based on the maximum distance or the furthest neighbor approach.
- $min\ d(i,j)$
- $d_{k(i,j)} = \max\,(d_{ki}, d_{kj})$



Average Linkage

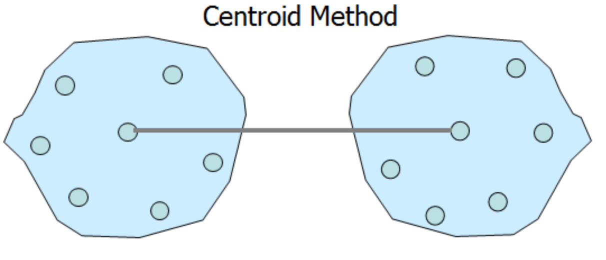
- The **average linkage** method the distance between two clusters is defined as the average of the distances between all pairs of objects.



15

Centroid Linkage

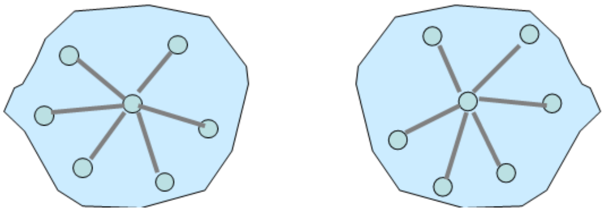
- In the **centroid methods**, the distance between two clusters is the distance between their centroids (means for all the variables).



16

Ward’s Procedure

- **Ward's procedure** is commonly used. For each cluster, the sum of squares is calculated. The two clusters with the smallest increase in the overall sum of squares within cluster distances are combined.
- $\sum_{i=1}^{n} W_k$



17

Example

- Assume that there are 5 objects in the data set and each object has 2 features.

No	X_1	X_2
1	4	2
2	6	4
3	5	1
4	10	6
5	11	8

18

Example

- Using Euclidean distance, distances between points are calculated and a "distance table" is created.

No	X_1	X_2
1	4	2
2	6	4
3	5	1
4	10	6
5	11	8

$$d(1,2) = \sqrt{(4-6)^2 + (2-4)^2} = 2.83$$

$$d(1,3) = \sqrt{(4-5)^2 + (2-1)^2} = 1.41$$

$$d(1,4) = \sqrt{(4-10)^2 + (2-6)^2} = 7.21$$

19

Example

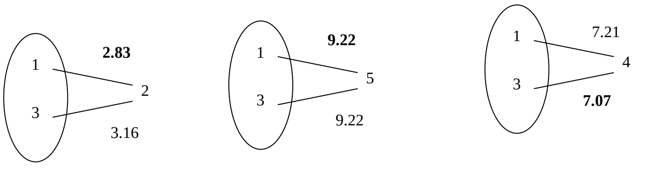
No	1	2	3	4	5
1					
2	2.83				
3	1.41	3.16			
4	7.21	4.47	7.07		
5	9.22	7.20	9.22	2.24	

$$Min\ d(i,j) = 1.41$$

- The re-distances are calculated according to the (1,3) set obtained in the next step.
- Some cells did not filled because the matrix is symmetric.

20

Example

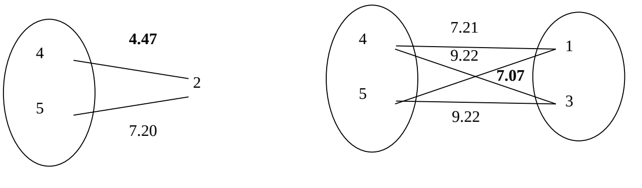


No	(1,3)	2	4	5
(1,3)				
2	2.83			
4	7.07	4.47		
5	9.22	7.20	2.24	

$$d(3,4) = \sqrt{(5-10)^2+(1-6)^2}$$

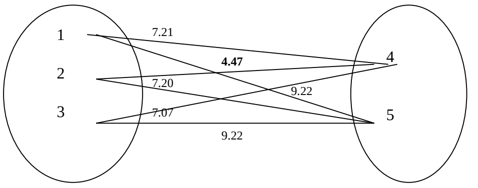
$$d(1,4) = \sqrt{(4-10)^2+(2-6)^2}$$

Example

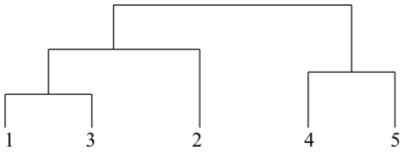


No	(1,3)	2	(4,5)
(1,3)			
2	2.83		
(4,5)	7.07	4.47	

Example



No	(1,2,3)	(4,5)
(1,2,3)		
(4,5)	4.47	



Disadvantages

- It is **difficult** to select merge or split points.
- Hierarchical clustering has **no backtracking.**
- If a particular merge or split turns out to be poor choice, it **cannot be corrected**

Extensions to Hierarchical Clustering

- Integration of hierarchical & distance-based clustering
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling