CHAPTER 12 – CLUSTERING ALGORITHMS II

- They produce a hierarchy of (**hard**) clusterings instead of a single clustering.
- Applications in:
  - Social sciences
  - Biological taxonomy
  - Modern biology
  - Medicine
  - Archaeology
  - Computer science and engineering

- Let $X=\{\underline{x}_1,\ldots,\underline{x}_N\}$, $\underline{x}_i=[x_{i1},\ldots,x_{il}]^T$. Recall that:
  - In hard clustering each vector belongs exclusively to a single cluster.
  - An m-(hard) clustering of X, $\Re$, is a partition of X into m sets (clusters) $C_1,\ldots,C_m$, so that:
    - $C_i \neq \emptyset, i=1,2,\ldots,m$
    - $\overset{m}{\underset{i=1}{U}} C_i = X$
    - $C_i \cap C = \emptyset, \quad i \neq j, \quad i, j = 1,2,\ldots, \quad m$

    By the definition: $\Re=\{C_j, j=1,\ldots m\}$

  - <u>Definition:</u> A clustering $\Re_1$ containing k clusters is said to be nested in the clustering $\Re_2$ containing r $(<k)$ clusters, if each cluster in $\Re_1$ is a subset of a cluster in $\Re_2$.

    We write $\Re_1 \angle \Re_2$

- Example: Let $\Re_1=\{\{\underline{x}_1,\underline{x}_3\},\{\underline{x}_4\},\{\underline{x}_2,\underline{x}_5\}\}$, $\Re_2=\{\{\underline{x}_1,\underline{x}_3,\underline{x}_4\},\{\underline{x}_2,\underline{x}_5\}\}$,

  $\Re_3=\{\{\underline{x}_1,\underline{x}_4\},\{\underline{x}_3\},\{\underline{x}_2,\underline{x}_5\}\}$, $\Re_4=\{\{\underline{x}_1,\underline{x}_2,\underline{x}_4\},\{\underline{x}_3,\underline{x}_5\}\}$.

  It is $\Re_1 \angle\ \Re_2$, but not $\Re_1 \angle\ \Re_3$, $\Re_1 \angle\ \Re_4$.

- Remarks:
  - Hierarchical clustering algorithms produce a hierarchy of nested clusterings.
  - They involve N steps at the most.
  - At each step t, the clustering $\Re_t$ is produced by $\Re_{t\text{-}1}$.

- Main categories:
  - Agglomerative clustering algorithms: Here $\Re_0=\{\{\underline{x}_1\},\ldots,\{\underline{x}_N\}\}$, $\Re_{N\text{-}1}=\{\{\underline{x}_1,\ldots,\underline{x}_N\}\}$ and $\Re_0 \angle \ldots \angle\ \Re_{N\text{-}1}$.
  - Divisive clustering algorithms: Here $\Re_0=\{\{\underline{x}_1,\ldots,\underline{x}_N\}\}$, $\Re_{N\text{-}1}= \{\{\underline{x}_1\},\ldots,\{\underline{x}_N\}\}$ and $\Re_{N\text{-}1} \angle \ldots \angle\ \Re_0$.

AGGLOMERATIVE ALGORITHMS

- Let $g(C_i, C_j)$ a proximity function between two clusters of X.

- *Generalized Agglomerative Scheme (GAS)*
  - Initialization
    - Choose $\Re_0 = \{\{\underline{x}_1\}, \dots, \{\underline{x}_N\}\}$
    - $t=0$
  - Repeat
    - $t=t+1$
    - Choose (C_i, C_j) in $\Re_{t-1}$ such that

$$g(C_i, C_j) = \begin{cases} \min_{r,s} g(C_r, C_s), \; \textit{if} \;\; g \; \textit{is a disim. function} \\ \max_{r,s} g(C_r, C_s), \; \textit{if} \;\; g \; \textit{is a sim. function} \end{cases}$$

    - Define $C_q = C_i \cup C_j$ and produce $\Re_t = (\Re_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$
  - Until all vectors lie in a single cluster.

➢Remarks:

- If two vectors come together into a single cluster at level t of the hierarchy, they will remain in the same cluster for all subsequent clusterings. As a consequence, there is no way to recover a "poor" clustering that may have occurred in an earlier level of hierarchy.
- Number of operations: $O(N^3)$

- Definitions of some useful quantities:

  Let $X=\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N\}$, with $\underline{x}_i=[x_{i1},x_{i2},\ldots,x_{il}]^T$.

  - Pattern matrix $(D(X))$: An Nxl matrix whose i-th row is $\underline{x}_i$ (transposed).
  - Proximity (similarity or dissimilarity) matrix $(P(X))$: An NxN matrix whose (i,j) element equals the proximity $\wp(\underline{x}_i,\underline{x}_j)$ (similarity $s(\underline{x}_i,\underline{x}_j)$, dissimilarity $d(\underline{x}_i,\underline{x}_j)$).

  - Example 1: Let $X=\{\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{x}_4, \underline{x}_5\}$, with $\underline{x}_1=[1, 1]^T$, $\underline{x}_2=[2, 1]^T$, $\underline{x}_3=[5, 4]^T$, $\underline{x}_4=[6, 5]^T$, $\underline{x}_5=[6.5, 6]^T$.

Euclidean distance

Tanimoto similarity

$$D(X)=\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}$$

$$P(X)=\begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0 \end{bmatrix}$$

$$P'(X)=\begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

- **Threshold dendrogram** (or **dendrorgram**): It is an effective way of representing the sequence of clusterings which are produced by an agglomerative algorithm.

In the previous example, if $d_{\min}^{ss}(C_i, C_j)$ is employed as the distance measure between two sets and the Euclidean one as the distance measure between two vectors, the following series of clusterings are produced:

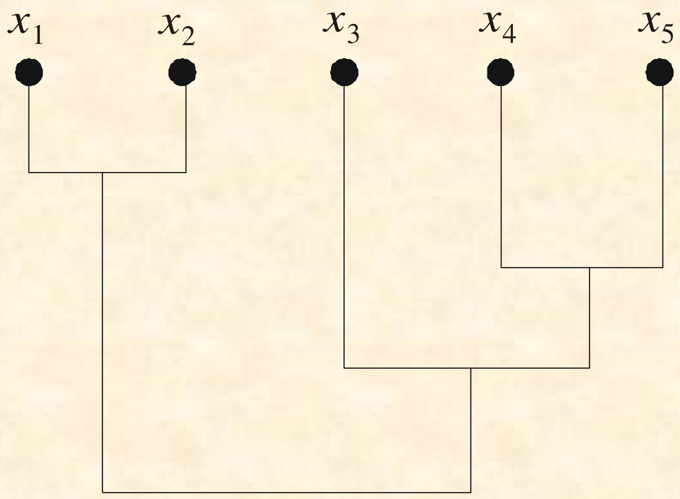
$\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}\}$

$\{\{x_1,x_2\},\{x_3\},\{x_4\},\{x_5\}\}$

$\{\{x_1,x_2\},\{x_3\},\{x_4,x_5\}\}$

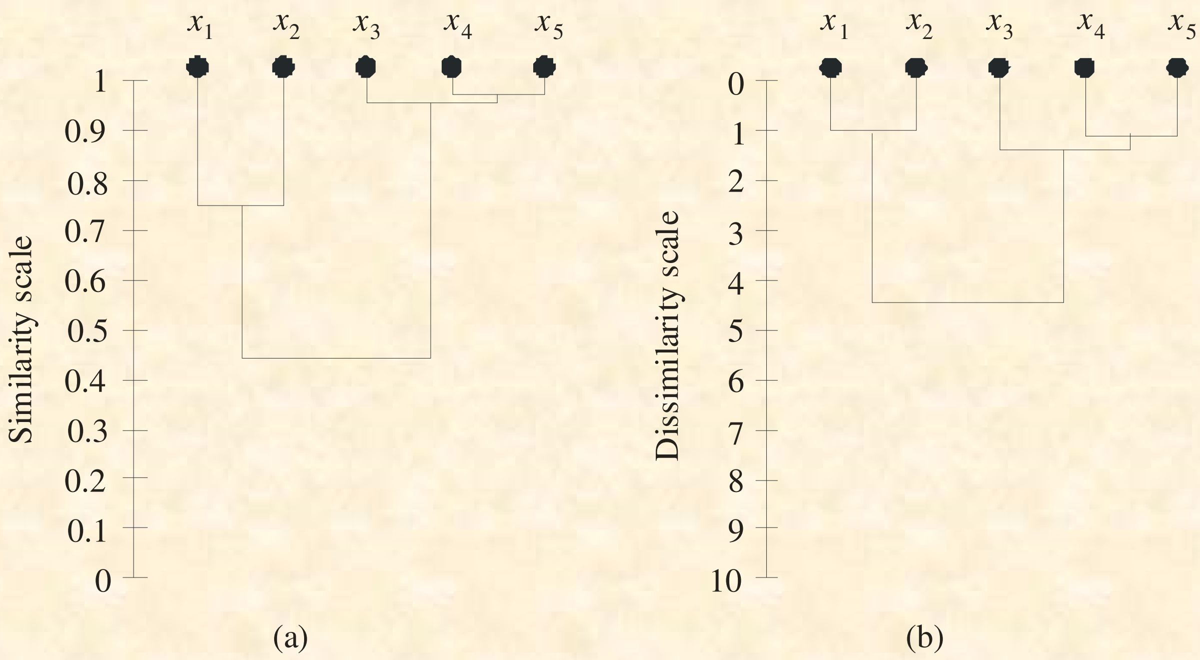
$\{\{x_1,x_2\},\{x_3,x_4,x_5\}\}$

$\{\{x_1,x_2,x_3,x_4,x_5\}\}$



- Proximity (dissimilarity or dissimilarity) dendrogram: A dendrogram that takes into account the level of proximity (dissimilarity or similarity) where two clusters are merged for the first time.

- Example 2: In terms of the previous example, the proximity dendrograms that correspond to $P'(X)$ and $P(X)$ are



(a) (b)

- Remark: One can readily observe the level in which a cluster is formed and the level in which it is absorbed in a larger cluster (indication of the natural clustering).

- Agglomerative algorithms are divided into:
  - Algorithms based on matrix theory.
  - Algorithms based on graph theory.

In the sequel we focus only on dissimilarity measures.

- Algorithms based on matrix theory.
  - They take as input the NxN dissimilarity matrix $P_0=P(X)$.
  - At each level t where two clusters C_i and C_j are merged to C_q, the dissimilarity matrix P_t is extracted from P_{t-1} by:
    - Deleting the two rows and columns of P_t that correspond to C_i and C_j.
    - Adding a new row and a new column that contain the distances of newly formed $C_q=C_i \cup C_j$ from the remaining clusters C_s, via a relation of the form

$$d(C_q,C_s)=f(d(C_i,C_s),d(C_j,C_s),d(C_i,C_j))$$

- A number of distance functions comply with the following update equation

$$d(C_q,C_s)=a_i d(C_i,C_s)+a_j(d(C_j,C_s)+bd(C_i,C_j)+c|d(C_i,C_s)-d(C_j,C_s)|$$

Algorithms that follow the above equation are:

- Single link (SL) algorithm ($a_i=1/2$, $a_j=1/2$, $b=0$, $c=-1/2$). In this case

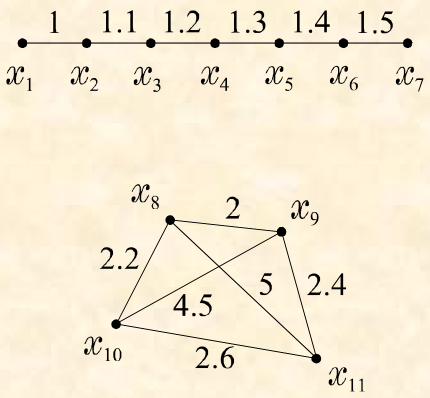
$$d(C_q,C_s)=min\{d(C_i,C_s),\ d(C_j,C_s)\}$$

- Complete link (CL) algorithm ($a_i=1/2$, $a_j=1/2$, $b=0$, $c=1/2$). In this case

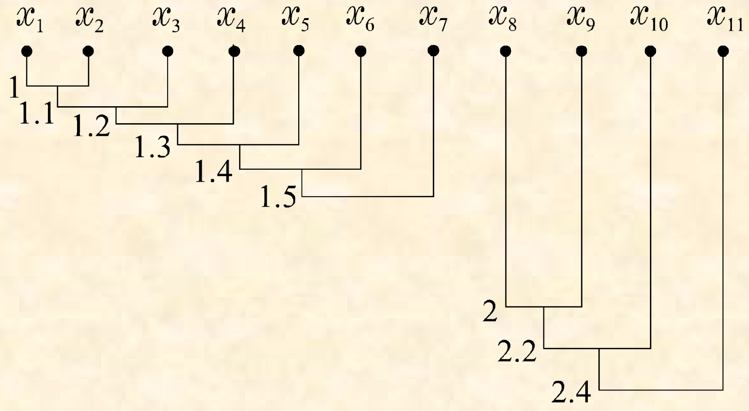
$$d(C_q,C_s)=max\{d(C_i,C_s),\ d(C_j,C_s)\}$$

- Remarks:
  - Single link forms clusters at low dissimilarities while complete link forms clusters at high dissimilarities.
  - Single link tends to form elongated clusters (*chaining effect*) while complete link tends to form compact clusters.
  - The rest algorithms are compromises between these two extremes.

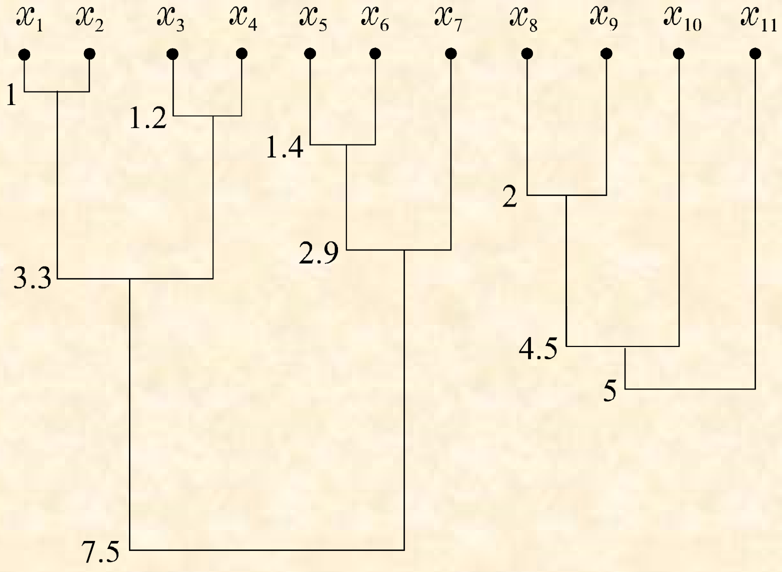
➢ Example:



(a)



(b)



(c)

(a) The data set X.

(b) The single link algorithm dissimilarity dendrogram.

(c) The complete link algorithm dissimilarity dendrogram

- Weighted Pair Group Method Average (WPGMA) (a_i=1/2, a_j=1/2, b=0, c=0). In this case:

$$d(C_q,C_s)=(d(C_i,C_s)+d(C_j,C_s))/2$$

- Unweighted Pair Group Method Average (UPGMA) ($a_i=n_i/(n_i+n_j)$, $a_j=n_j/(n_i+n_j)$, b=0, c=0, where n_i is the cardinality of C_i). In this case:

$$d(C_q,C_s)=(n_i\, d(C_i,C_s)+n_j\, d(C_j,C_s))/(n_i+n_j)$$

- Unweighted Pair Group Method Centroid (UPGMC) ($a_i=n_i/(n_i+n_j)$, $a_j=n_j/(n_i+n_j)$, $b=-n_i\, n_j/(n_i+n_j)^2$, c=0). In this case:

$$d_{qs}=\frac{n_i}{n_i+n_j}d_{is}+\frac{n_j}{n_i+n_j}d_{js}-\frac{n_i n_j}{(n_i+n_j)^2}d_{ij}$$

For the UPGMC, it is true that $d_{qs}=\|\underline{m}_q-\underline{m}_s\|^2$, where $\underline{m}_q$ is the mean of C_q.

- Weighted Pair Group Method Centroid (WPGMC) (a_i=1/2, a_j=1/2, b=-1/4, c=0). In this case

$$d_{qs}=(d_{is}+d_{js})/2-d_{ij}/4$$

For WPGMC there are cases where $d_{qs}\leq max\{d_{is}, d_{js}\}$ (crossover)

- Ward or minimum variance algorithm. Here the distance d'_{ij} between C_i and C_j is defined as

$$d'_{ij}=(n_i\, n_j/(n_i+n_j))\, \|\underline{m}_i-\underline{m}_j\|^2$$

d'_{qs} can also be written as

$$d'_{qs}=((n_i+n_j)d'_{is}+(n_i+n_j)d'_{js}-n_s d'_{ij})/(n_i+n_j+n_s)$$

- Remark: Ward’s algorithm forms $\Re_{t+1}$ by merging the two clusters that lead to the smallest possible increase of the total variance, i.e.,

$$E_t=\sum_{r=1}^{N-t}\sum_{\underline{x}\in C_r}\|\underline{x}-\underline{m}_r\|^2$$

- Example 3: Consider the following dissimilarity matrix (Euclidean distance)

$$P_0 = \begin{bmatrix} 0 & 1 & 2 & 26 & 37 \\ 1 & 0 & 3 & 25 & 36 \\ 2 & 3 & 0 & 16 & 25 \\ 26 & 25 & 16 & 0 & 1.5 \\ 37 & 36 & 25 & 1.5 & 0 \end{bmatrix}$$

$\Re_0=\{\{\underline{x}_1\}, \{\underline{x}_2\}, \{\underline{x}_3\}, \{\underline{x}_4\}, \{\underline{x}_5\}\}$,
$\Re_1=\{\{\underline{x}_1, \underline{x}_2\}, \{\underline{x}_3\}, \{\underline{x}_4\}, \{\underline{x}_5\}\}$,
$\Re_2=\{\{\underline{x}_1, \underline{x}_2\}, \{\underline{x}_3\}, \{\underline{x}_4, \underline{x}_5\}\}$,
$\Re_3=\{\{\underline{x}_1, \underline{x}_2, \underline{x}_3\}, \{\underline{x}_4, \underline{x}_5\}\}$,
$\Re_4=\{\{\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{x}_4, \underline{x}_5\}\}$

All the algorithms produce the above sequence of clusterings at different proximity levels:

	SL	*CL*	*WPGMA*	*UPGMA*	*WPGMC*	*UPGMC*	*Ward*
$\Re_0$	0	0	0	0	0	0	0
$\Re_1$	1	1	1	1	1	1	0.5
$\Re_2$	1.5	1.5	1.5	1.5	1.5	1.5	0.75
$\Re_3$	2	3	2.5	2.5	2.25	2.25	1.5
$\Re_4$	16	37	25.75	27.5	24.69	26.46	31.75

➢ Complexity issues:

- GAS requires, in general, $O(N^3)$ operations.
- More efficient implementations require $O(N^2 logN)$ computational time.
- For a class of widely used algorithms, implementations that require $O(N^2)$ computational time and $O(N^2)$ or $O(N)$ storage have also been proposed.
- Parallel implementations on SIMD machines have also been considered.