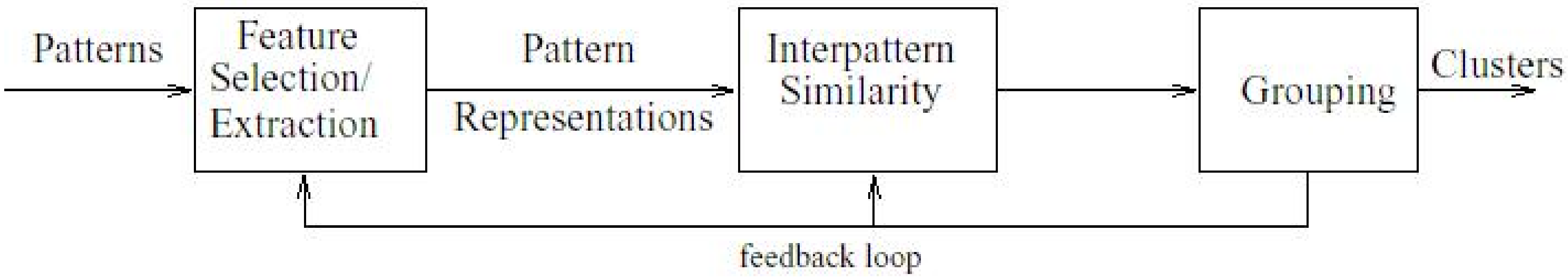
Introduction to Hierarchical Clustering Analysis

Introduction

- Data clustering concerns how to group a set of objects based on their similarity of attributes and/or their proximity in the vector space.
- Main methods
  - Partitioning : K-Means…
  - Hierarchical : BIRCH,ROCK,…
  - Density-based: DBSCAN,…
- A good clustering method will produce high quality clusters with
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters

Stages in Clustering



Clustering Algorithms

A. Distance and Similarity Measures

B. Hierarchical Clustering

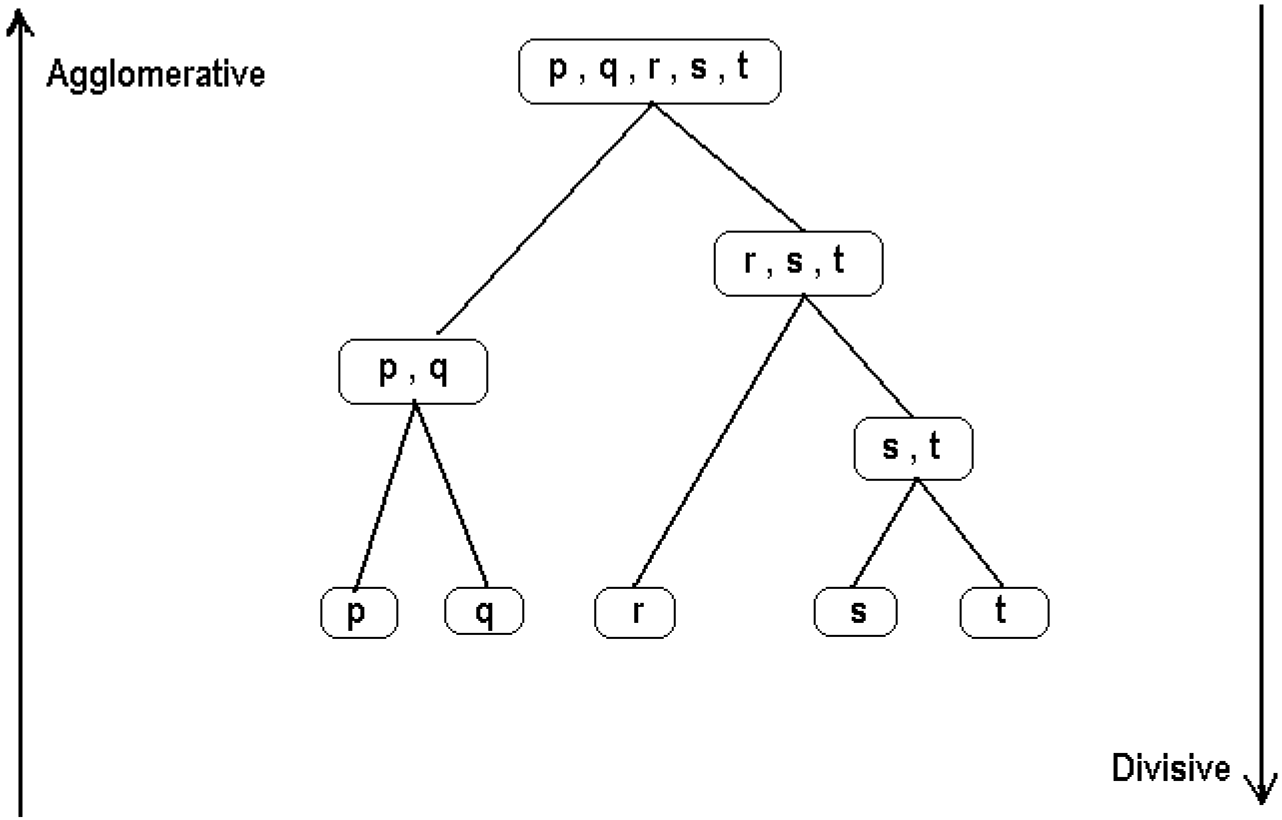
— Agglomerative

Single linkage, complete linkage, group average linkage, median linkage, centroid linkage, balanced iterative reducing and clustering using hierarchies (BIRCH), clustering using representatives (CURE), robust clustering using links (ROCK)

— Divisive

divisive analysis (DIANA), monothetic analysis (MONA)

Hierarchical Clustering



Agglomerative clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. Divisive clustering works the other way around.

Distance and Similarity Measures

SIMILARITY AND DISSIMILARITY MEASURE FOR QUANTITATIVE FEATURES

Measures	Forms	Comments	Examples and Applications
Minkowski distance	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/n} \right)^n$	Metric. Invariant to any translation and rotation only for $n=2$ (Euclidean distance). Features with large values and variances tend to dominate over other features.	Fuzzy *c*-means with measures based on Minkowski family [130].
Euclidean distance	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/2} \right)^2$	The most commonly used metric. Special case of Minkowski metric at $n=2$. Tend to form hyperspherical clusters.	*K*-means algorithm [191]
City-block distance	$D_{ij} = \sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|$	Special case of Minkowski metric at $n=1$. Tend to form hyperrecutangular clusters.	Fuzzy ART [57]
Sup distance	$D_{ij} = \max_{1 \le l \le d} \left\| x_{il} - x_{jl} \right\|$	Special case of Minkowski metric at $n \to \infty$.	Fuzzy *c*-means with sup norm [39].
Mahalanobis distance	$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{S}$ is the within-group covariance matrix.	Invariant to any nonsingular linear transformation. $\mathbf{S}$ is calculated based on all objects. Tend to form hyperellipsoidal clusters. When features are not correlated, squared Mahalanobis distance is equivalent to squared Euclidean distance. May cause some computational burden.	Ellipsoidal ART [13], Hyperellipsoidal clustering algorithm [194].
Pearson correlation	$D_{ij} = (1 - r_{ij})/2$, where $r_{ij} = \frac{\sum_{l=1}^{d} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{d} (x_{il} - \bar{x}_i)^2 \sum_{l=1}^{d} (x_{jl} - \bar{x}_j)^2}}$	Not a metric. Derived from correlation coefficient. Unable to detect the magnitude of differences of two variables.	Widely used as the measure for analyzing gene expression data [80].
Point symmetry distance	$D_{ir} = \min_{\substack{j=1,\ldots,N \\ \text{and } j \ne i}} \frac{\left\| (\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r) \right\|}{\left\| (\mathbf{x}_i - \mathbf{x}_r) \right\| + \left\| (\mathbf{x}_j - \mathbf{x}_r) \right\|}$	Not a metric. Compute the distance between an object $\mathbf{x}_i$ and a reference point $\mathbf{x}_r$. D_{ir} is minimized when a symmetric pattern exists.	SBKM (Symmetry-based *K*-means) [264].
Cosine similarity	$S_{ij} = \cos \alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\left\| \mathbf{x}_i \right\| \left\| \mathbf{x}_j \right\|}$	Independent of vector length. Invariant to rotation, but not to linear transformations.	The most commonly used measure in document clustering [261].

Similarity Measurements

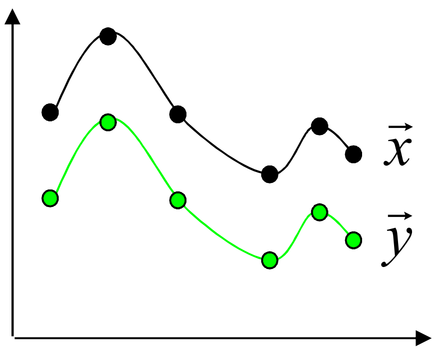
- Pearson Correlation

Two profiles (vectors) $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ *and* $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

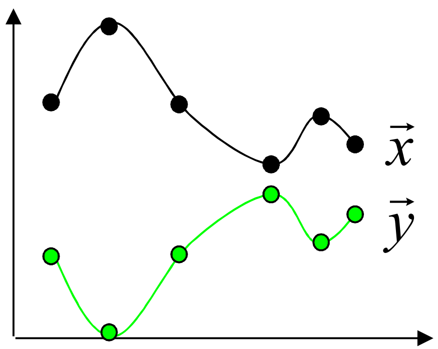
$$C_{pearson}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i - m_x)^2][\sum_{i=1}^{N}(y_i - m_y)^2]}}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$m_y = \frac{1}{N}\sum_{n=1}^{N} y_n$$

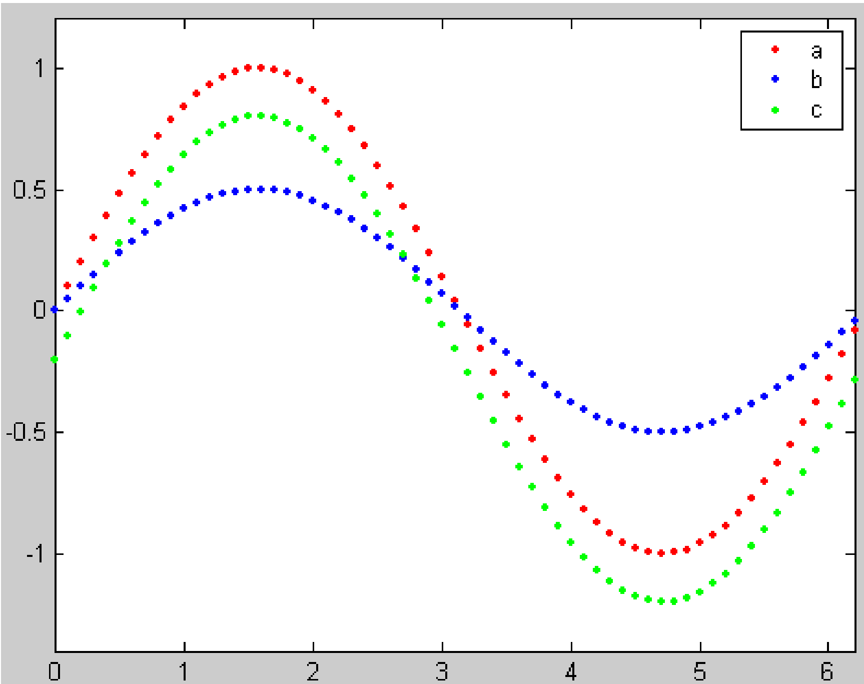


+1 ≥ Pearson Correlation ≥ – 1



Similarity Measurements

- Pearson Correlation: Trend Similarity



$$\vec{b} = 0.5\vec{a}$$

$$\vec{c} = \vec{a} - 0.2$$

$$C_{pearson}(\vec{a}, \vec{b}) = 1$$

$$C_{pearson}(\vec{a}, \vec{c}) = 1$$

$$C_{pearson}(\vec{b}, \vec{c}) = 1$$

Similarity Measurements

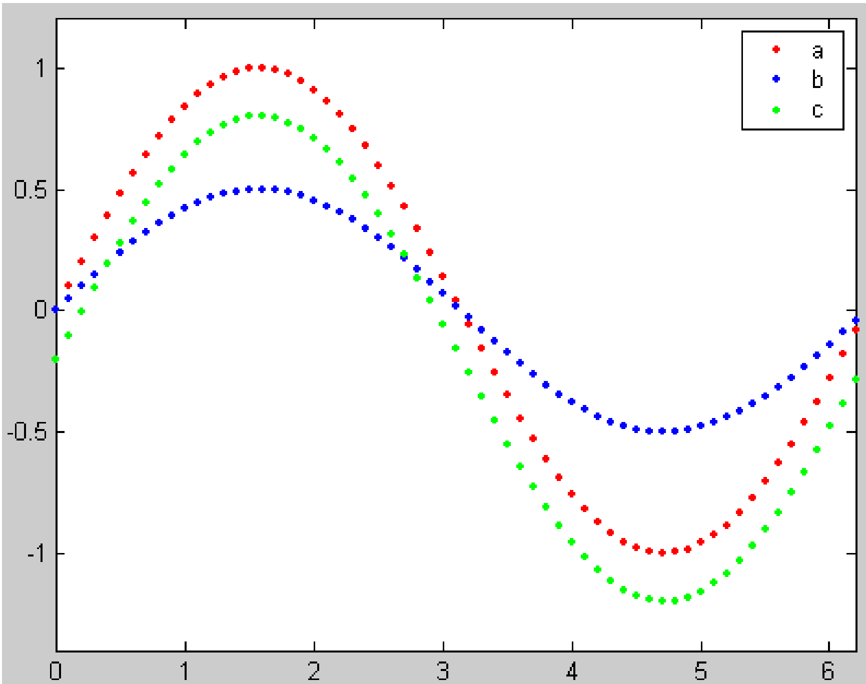
- Euclidean Distance

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

Similarity Measurements

- Euclidean Distance: Absolute difference



$$\vec{b} = 0.5\vec{a}$$

$$\vec{c} = \vec{a} - 0.2$$

$$d(\vec{a},\vec{b}) = 2.8025$$

$$d(\vec{a},\vec{c}) = 1.5875$$

$$d(\vec{b},\vec{c}) = 3.2211$$

Similarity Measurements

- Cosine Correlation

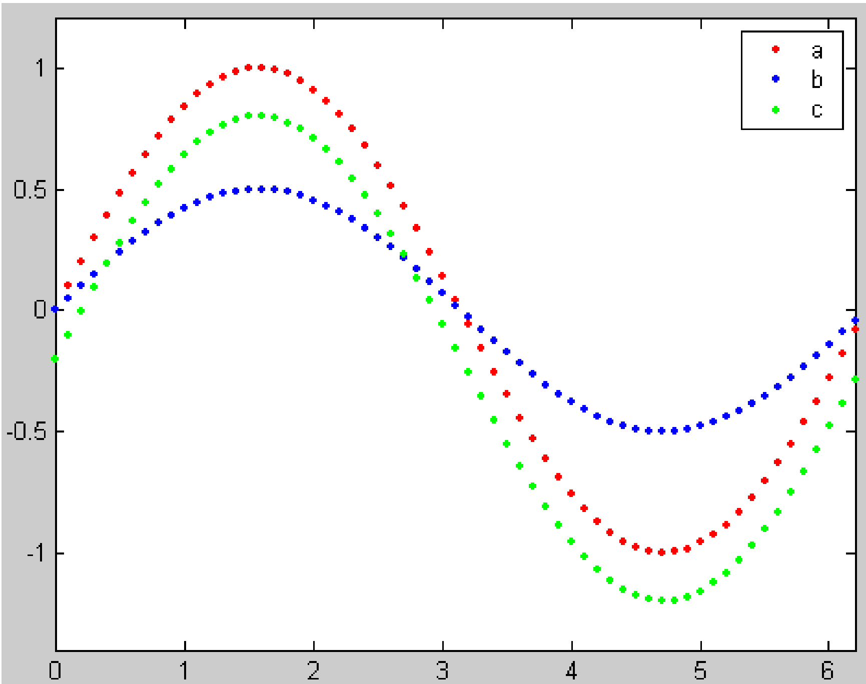
$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$C_{\text{cosine}}(\vec{x}, \vec{y}) = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i \times y_i}{\left\|\vec{x}\right\| \times \left\|\vec{y}\right\|}$$

$\vec{x} = \vec{y}$ *+1* $\geq$ *Cosine Correlation* $\geq$ *– 1* $\vec{x} = -\vec{y}$

Similarity Measurements

- Cosine Correlation: Trend + Mean Distance



$$\vec{b} = 0.5\vec{a}$$

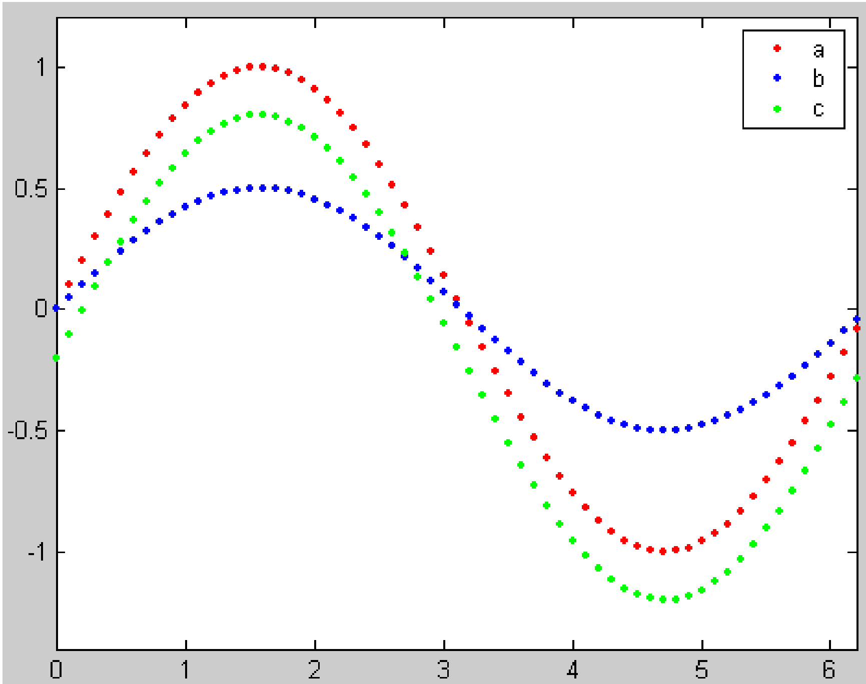
$$\vec{c} = \vec{a} - 0.2$$

$$C_{\cos ine}(\vec{a}, \vec{b}) = 1$$

$$C_{\cos ine}(\vec{a}, \vec{c}) = 0.9622$$

$$C_{\cos ine}(\vec{b}, \vec{c}) = 0.9622$$

Similarity Measurements



$$\vec{b} = 0.5\vec{a}$$

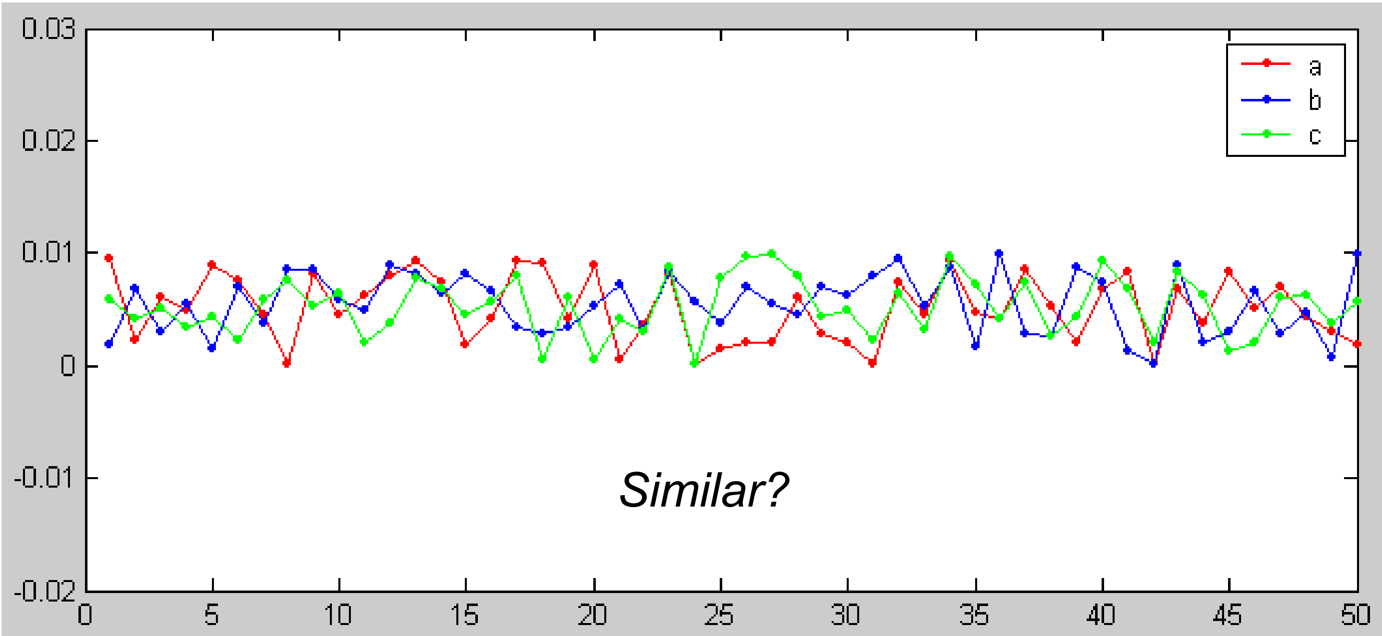
$$\vec{c} = \vec{a} - 0.2$$

$C_{pearson}(\vec{a},\vec{b}) = 1$ $\quad d(\vec{a},\vec{b}) = 2.8025$ $\quad C_{\cos ine}(\vec{a},\vec{b}) = 1$

$C_{pearson}(\vec{a},\vec{c}) = 1$ $\quad d(\vec{a},\vec{c}) = 1.5875$ $\quad C_{\cos ine}(\vec{a},\vec{c}) = 0.9622$

$C_{pearson}(\vec{b},\vec{c}) = 1$ $\quad d(\vec{b},\vec{c}) = 3.2211$ $\quad C_{\cos ine}(\vec{b},\vec{c}) = 0.9622$

Similarity Measurements

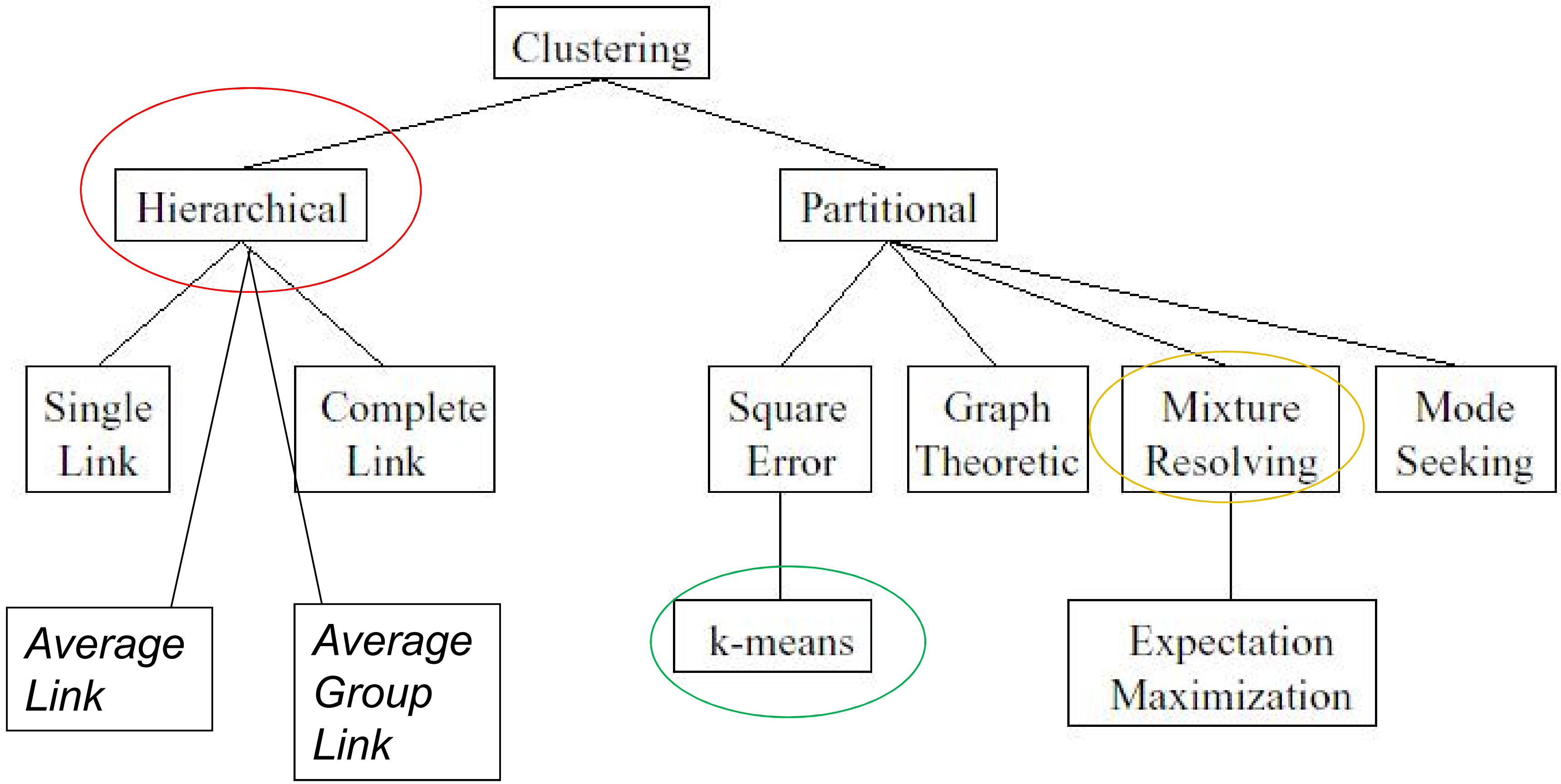


$$C_{pearson}(\vec{a},\vec{b}) = -0.1175 \qquad d(\vec{a},\vec{b}) = 0.0279 \qquad C_{\cos ine}(\vec{a},\vec{b}) = 0.7544$$

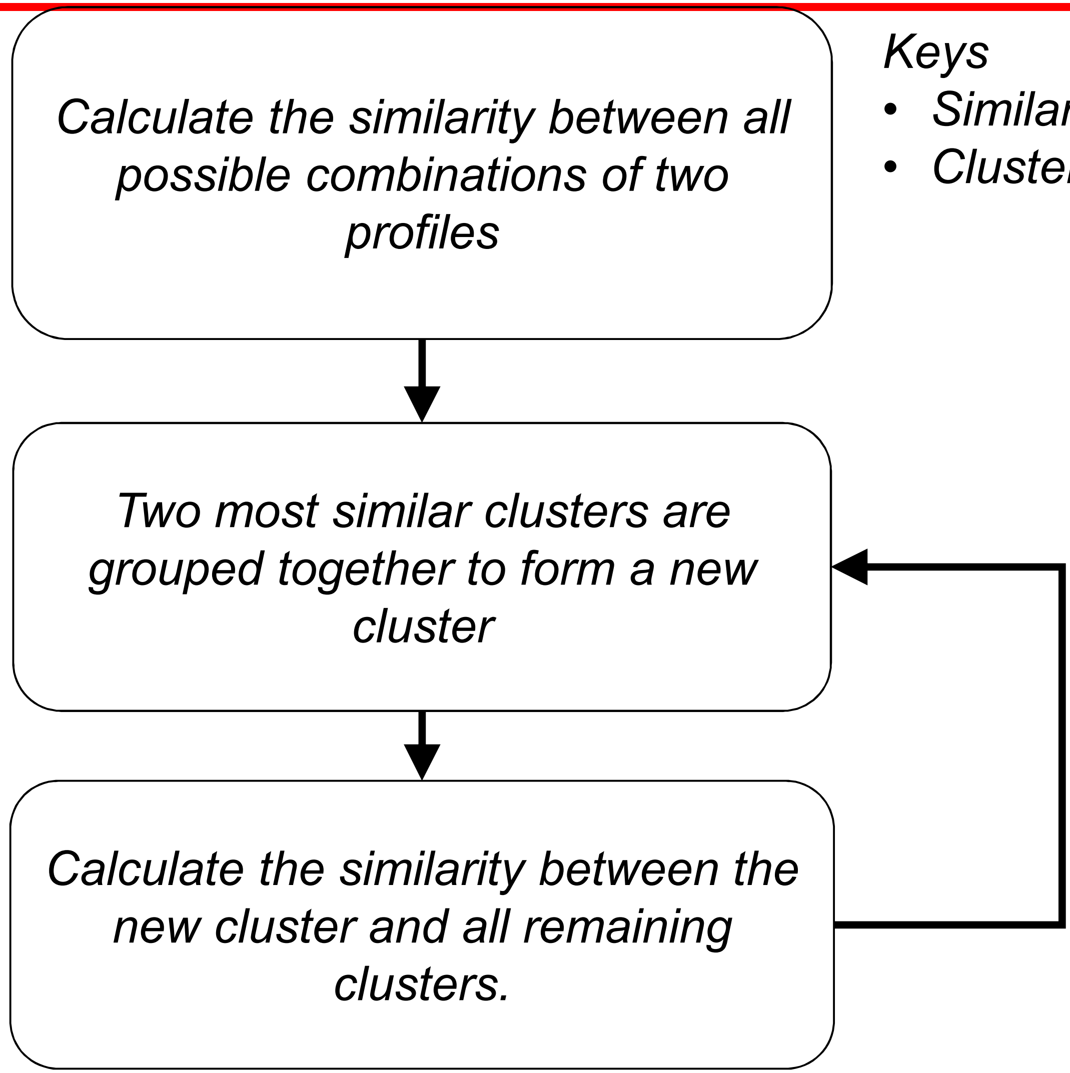
$$C_{pearson}(\vec{a},\vec{c}) = 0.1244 \qquad d(\vec{a},\vec{c}) = 0.0255 \qquad C_{\cos ine}(\vec{a},\vec{c}) = 0.8092$$

$$C_{pearson}(\vec{b},\vec{c}) = 0.1779 \qquad d(\vec{b},\vec{c}) = 0.0236 \qquad C_{\cos ine}(\vec{b},\vec{c}) = 0.844$$

Taxonomy of Clustering Approaches



Hierarchical Clustering



Keys

- *Similarity*
- *Clustering*

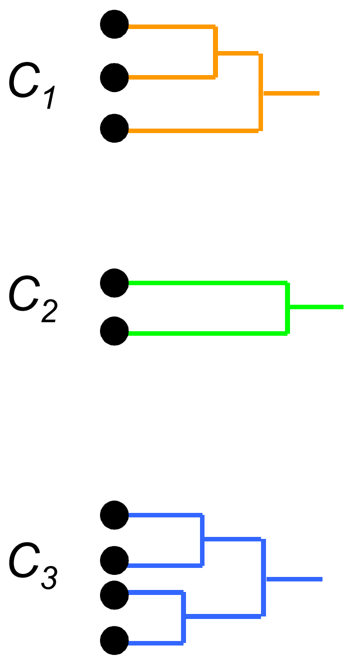
General **agglomerative** clustering

1) Start with N singleton clusters. Calculate the proximity matrix for the N clusters.
2) Search the minimal distance

$$D(C_i, C_j) = \min_{\substack{1 \le m,l \le N \\ m \ne l}} D(C_m, C_l)$$

   where $D(*,*)$ is the distance function discussed before, in the proximity matrix, and combine cluster C_i and C_j to form a new cluster.
3) Update the proximity matrix by computing the distances between the new cluster and the other clusters.
4) Repeat steps 2)–3) until all objects are in the same cluster.

Clustering

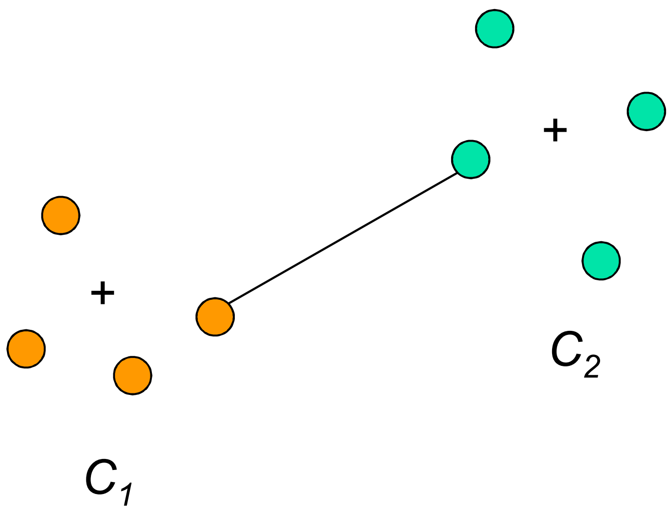


Merge which pair of clusters?

- 4 options or distance functions

Clustering

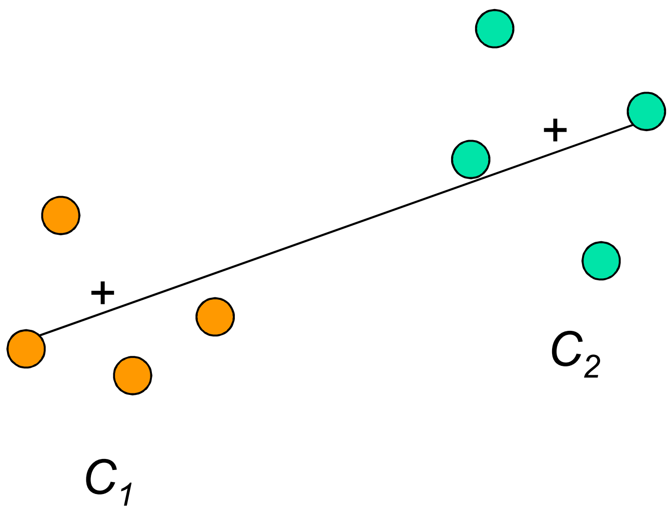
Single Linkage



Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters

Clustering

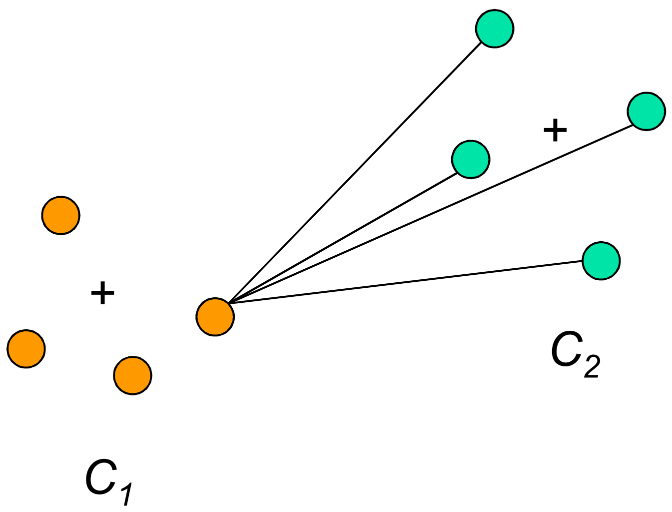
Complete Linkage



Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Clustering

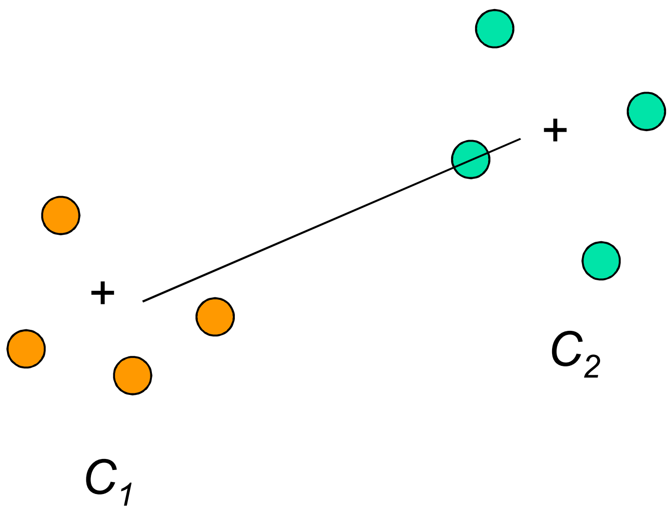
Average Linkage



Dissimilarity between two clusters = Averaged distances of all pairs of objects (one from each cluster).

Clustering

Average Group Linkage



Dissimilarity between two clusters = Distance between two cluster means.