Chapter DM:II

II. Cluster Analysis

- Cluster Analysis Basics
- Hierarchical Cluster Analysis
- Iterative Cluster Analysis
- Density-Based Cluster Analysis
- Cluster Evaluation
- Constrained Cluster Analysis

Cluster Analysis Basics

Cluster analysis is the unsupervised classification of a set of objects in groups, pursuing the following objectives:

1. maximize the similarities within the groups (intra groups)
2. minimize the similarities between the groups (inter groups)

Cluster Analysis Basics

Cluster analysis is the unsupervised classification of a set of objects in groups, pursuing the following objectives:

1. maximize the similarities within the groups (intra groups)
2. minimize the similarities between the groups (inter groups)

Applications:

- identification of similar groups of buyers
- “higher-level” image processing: object recognition
- search of similar gene profiles
- specification of syndromes
- analysis of traffic data in computer networks
- visualization of complex graphs
- text categorization in information retrieval

Remarks:

- The setting of a cluster analysis is reverse to the setting of a variance analysis:
  - A variance analysis verifies whether a nominal feature defines groups such that the members of the different groups differ significantly with regard to a numerical feature. I.e., the nominal feature is in the role of the independent variable, while the numerical feature(s) is (are) in role of dependent variable(s).
    Example: The type of a product packaging (the independent variable) may define the number of customers (the dependent variable) in a supermarket who look at the product.
  - A cluster analysis, in turn, can be used to identify such a nominal feature, namely by constructing a suited feature domain for the nominal variable: each cluster corresponds implicitly to a value of the domain.
    Example: Equivalent but differently presented products in a supermarket are clustered (= the impact of product packaging is identified) with regard to the number of customers who buy the products.
- Cluster analysis is a tool for structure *generation*. Nearly nothing is known about the nominal variable that is to be identified. In particular, there is no knowledge about the number of domain values (the number of clusters).
- Variance analysis is a tool for structure *verification*.

Cluster Analysis Basics

Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ denote the p-dimensional feature vectors of n objects:

	Feature 1	Feature 2	...	Feature p
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	...	x_{1_p}
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	...	x_{2_p}
$\vdots$				
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	...	x_{n_p}

Cluster Analysis Basics

Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ denote the p-dimensional feature vectors of n objects:

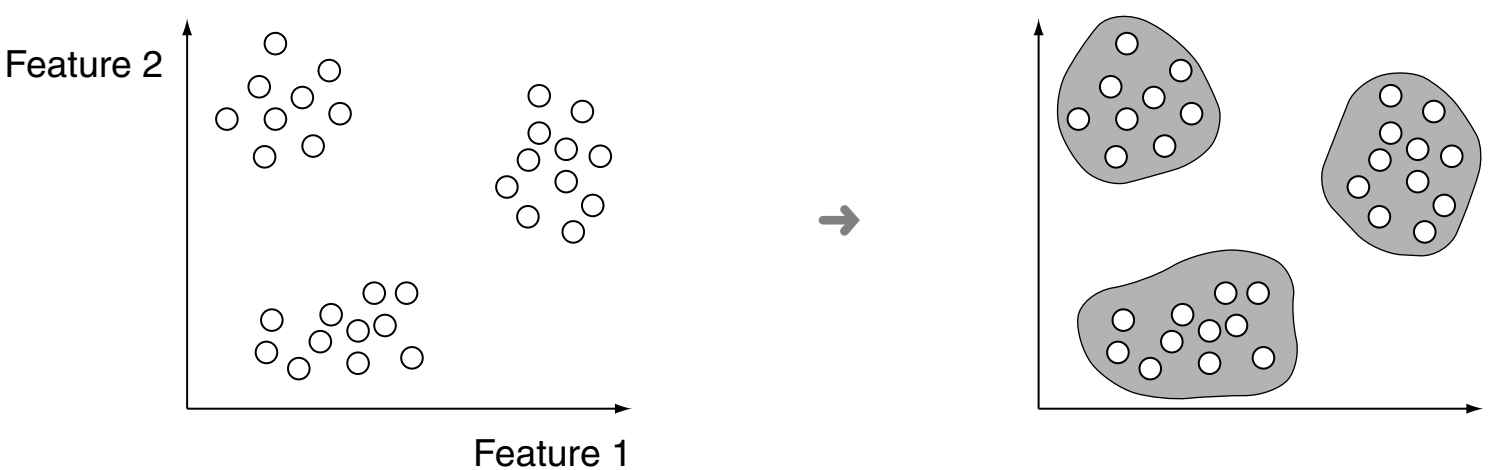
	Feature 1	Feature 2	...	Feature p	no Target concept
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	...	x_{1_p}	c_1
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	...	x_{2_p}	c_2
$\vdots$					$\vdots$
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	...	x_{n_p}	c_n

Cluster Analysis Basics

Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ denote the p-dimensional feature vectors of n objects:

	Feature 1	Feature 2	...	Feature p	no Target concept
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	...	x_{1_p}	c_1
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	...	x_{2_p}	c_2
$\vdots$					$\vdots$
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	...	x_{n_p}	c_n

30 two-dimensional feature vectors ($n = 30$, $p = 2$):



Cluster Analysis Basics

Definition 3 (Exclusive Clustering [splitting]**)**

Let X be a set of feature vectors. An exclusive clustering $\mathcal{C}$ of X, $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, $C_i \subseteq X$, is a partitioning of X into non-empty, mutually exclusive subsets C_i with $\bigcup_{C_i \in \mathcal{C}} C_i = X$.

Cluster Analysis Basics

Definition 3 (Exclusive Clustering [splitting]**)**

Let X be a set of feature vectors. An exclusive clustering $\mathcal{C}$ of X, $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, $C_i \subseteq X$, is a partitioning of X into non-empty, mutually exclusive subsets C_i with $\bigcup_{C_i \in \mathcal{C}} C_i = X$.

Algorithms for cluster analysis are unsupervised learning methods:

- the learning process is self-organized
- there is no (external) teacher
- the optimization criterion is task- and domain-independent

Cluster Analysis Basics

Definition 3 (Exclusive Clustering [splitting]**)**

Let X be a set of feature vectors. An exclusive clustering $\mathcal{C}$ of X, $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, $C_i \subseteq X$, is a partitioning of X into non-empty, mutually exclusive subsets C_i with $\bigcup_{C_i \in \mathcal{C}} C_i = X$.

Algorithms for cluster analysis are unsupervised learning methods:

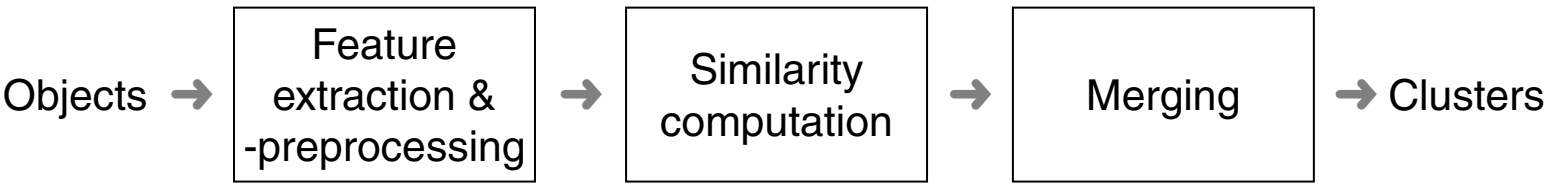
- the learning process is self-organized
- there is no (external) teacher
- the optimization criterion is task- and domain-independent

Supervised learning:

- a learning objective such as the target concept is provided
- the optimization criterion is defined by the task or the domain
- information is provided about *how* the optimization criterion can be maximized. Keyword: instructive feedback

Cluster Analysis Basics

Main Stages of a Cluster Analysis

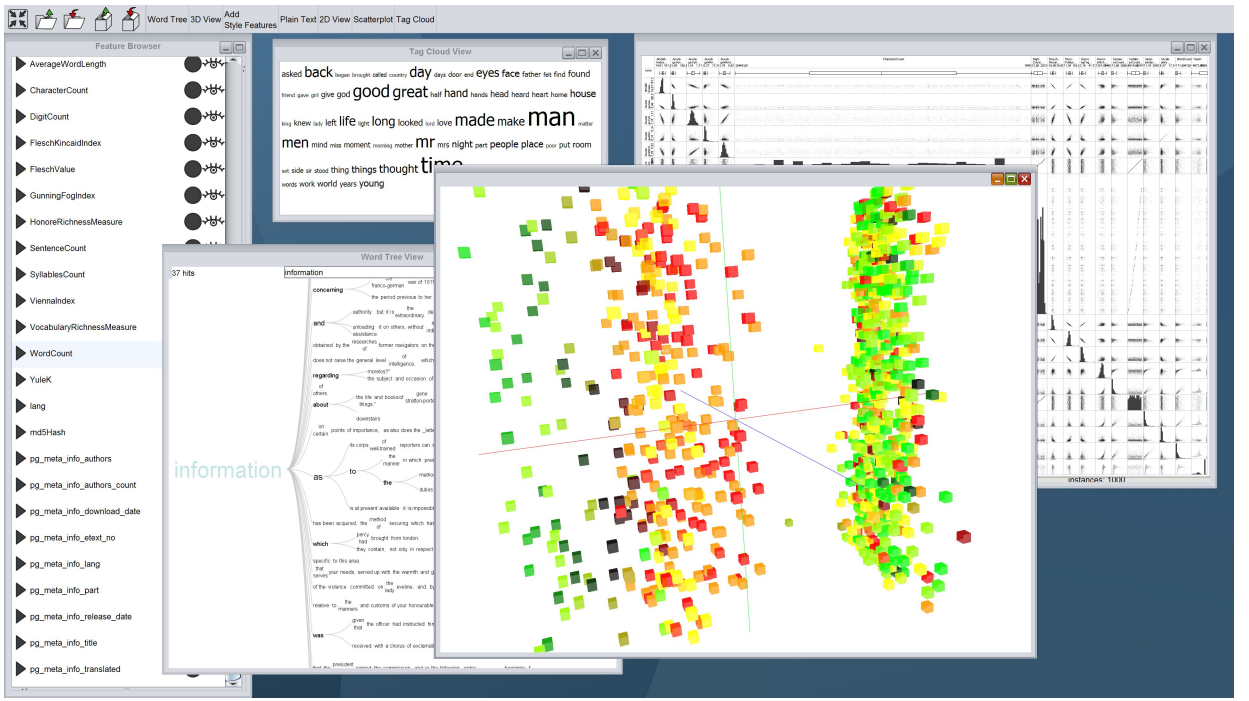


Cluster Analysis Basics

Feature Extraction and Preprocessing [cluster analysis stages]

Required are (possibly new) features of high variance. Approaches:

- analysis of spreading parameters
- dimension reduction: PCA, factor analysis, MDS
- visual inspection: scatter plots, box plots



[Webis 2012, VDM tool]

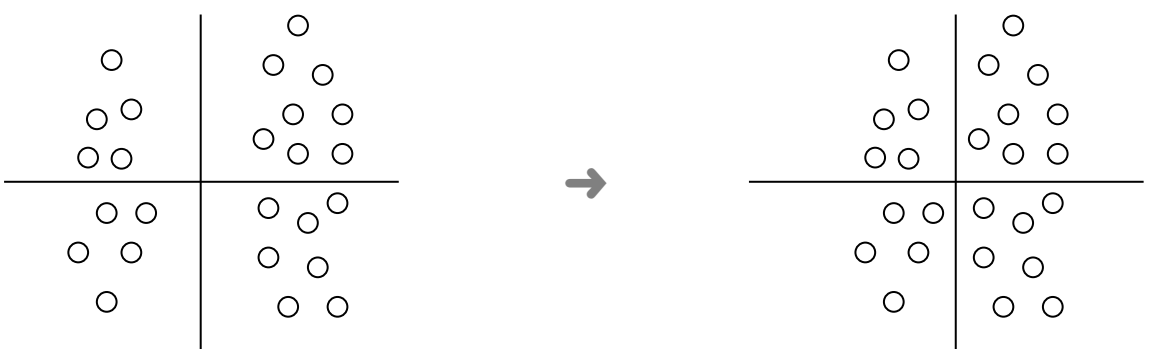
Cluster Analysis Basics

Feature Extraction and Preprocessing [cluster analysis stages]

Required are (possibly new) features of high variance. Approaches:

- analysis of spreading parameters
- dimension reduction: PCA, factor analysis, MDS
- visual inspection: scatter plots, box plots

Feature standardization can dampen the structure and make things worse:



Cluster Analysis Basics

Computation of Distances or Similarities [cluster analysis stages]

Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ denote the p-dimensional feature vectors of n objects:

	Feature 1	Feature 2	...	Feature p
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	...	x_{1_p}
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	...	x_{2_p}
$\vdots$				
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	...	x_{n_p}

	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_n$
$\mathbf{x}_1$	0	$d(\mathbf{x}_1, \mathbf{x}_2)$	...	$d(\mathbf{x}_1, \mathbf{x}_n)$
$\mathbf{x}_2$	-	0	...	$d(\mathbf{x}_2, \mathbf{x}_n)$
$\vdots$				
$\mathbf{x}_n$	-	-	...	0

Remarks:

- Usually, the distance matrix is defined implicitly by a metric on the feature space.
- The distance matrix can be understood as the adjacency matrix of a weighted, undirected graph G, $G = \langle V, E, w \rangle$. The set X of feature vectors is mapped one-to-one (bijection) onto a set of nodes V. The distance $d(\mathbf{x}_i, \mathbf{x}_j)$ corresponds to the weight $w(\{u, v\})$ of edge $\{u, v\} \in E$ between those nodes u and v that are associated with $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively.

Cluster Analysis Basics

Computation of Distances or Similarities (continued)

Properties of a distance function:

1. $d(\mathbf{x}_1, \mathbf{x}_2) \geq 0$
2. $d(\mathbf{x}_1, \mathbf{x}_1) = 0$
3. $d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_2, \mathbf{x}_1)$
4. $d(\mathbf{x}_1, \mathbf{x}_3) \leq d(\mathbf{x}_1, \mathbf{x}_2) + d(\mathbf{x}_2, \mathbf{x}_3)$

Cluster Analysis Basics

Computation of Distances or Similarities (continued)

Properties of a distance function:

1. $d(\mathbf{x}_1, \mathbf{x}_2) \geq 0$
2. $d(\mathbf{x}_1, \mathbf{x}_1) = 0$
3. $d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_2, \mathbf{x}_1)$
4. $d(\mathbf{x}_1, \mathbf{x}_3) \leq d(\mathbf{x}_1, \mathbf{x}_2) + d(\mathbf{x}_2, \mathbf{x}_3)$

Minkowsky metric for features with interval-based measurement scales:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{i=1}^{p} |x_{1_i} - x_{2_i}|^r \right)^{1/r}$$

where

- $r = 1$. Manhattan or Hamming distance, L_1 norm
- $r = 2$. Euclidean distance, L_2 norm
- $r = \infty$. Maximum distance, L_∞ norm or L_{max} norm

Cluster Analysis Basics

Computation of Distances or Similarities (continued)

Cluster analysis does not presume a particular measurement scale.

➜ Generalization of the distance function towards a (dis)similarity function by omitting the triangle inequality. (Dis)similarities can be quantified between all kinds of features—irrespective of the given levels of measurement.

Cluster Analysis Basics

Computation of Distances or Similarities (continued)

Cluster analysis does not presume a particular measurement scale.

➜ Generalization of the distance function towards a (dis)similarity function by omitting the triangle inequality. (Dis)similarities can be quantified between all kinds of features—irrespective of the given levels of measurement.

Similarity coefficients for two feature vectors, $\mathbf{x}_1$, $\mathbf{x}_2$, with binary features:

$$\text{Simple Matching Coefficient (SMC)} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}}$$

$$\text{Jaccard Coefficient (J)} = \frac{f_{11}}{f_{11} + f_{01} + f_{10}}$$

where

f_{11} = number of features with a value of 1 in both $\mathbf{x}_1$ and $\mathbf{x}_2$
f_{00} = number of features with a value of 0 in both $\mathbf{x}_1$ and $\mathbf{x}_2$
f_{01} = number of features with value 0 in $\mathbf{x}_1$ and value 1 in $\mathbf{x}_2$
f_{10} = number of features with value 1 in $\mathbf{x}_1$ and value 0 in $\mathbf{x}_2$

Remarks:

- The definitions for the above similarity coefficients can be extended towards features with a nominal measurement scale.
- Particular heterogeneous metrics have been developed, such as HEOM and HVDM, which allow the combined computation of feature values from different measurement scales.
- The computation of the correlation between all features of two feature vectors (not: between two features over all feature vectors) allows to compare feature profiles.
  Example: Q correlation coefficient
- The development of a suited, realistic, and expressive similarity measure is the biggest challenge within a cluster analysis tasks. Typical problems:
  - (unwanted) structure damping due to normalization
  - (unwanted) sensitivity concerning outliers
  - (unrecognized) feature correlations
  - (neglected) varying feature importance
- Similarity measures can be transformed straightforwardly into dissimilarity measures—and vice versa.

Cluster Analysis Basics

Merging Principles [cluster analysis stages]