# Chapter VII: Cluster analysis

## Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

# Chapter VII: Cluster analysis

## What is cluster analysis?

**Cluster: A collection of data objects within a larger set that are.**

Similar (or related) to one another within the same group and,
dissimilar (or unrelated) to the objects outside the group.

**Cluster analysis (or clustering, data segmentation, . . .).**

Define similarities among data based on the characteristics found in the data (input from user!).
Group similar data objects into clusters.

**Unsupervised learning:**

No predefined classes.
I.e., learning by observation (vs. learning by examples: supervised).

**Typical applications:**

As a stand-alone tool to get insight into data distribution.
As a preprocessing step for other algorithms.

## Clustering for data understanding and applications

**Biology:**

Taxonomy of living things: kingdom, phylum, class, order, family, genus, and species.

**Information retrieval:**

Document clustering.

**Land use:**

Identification of areas of similar land use in an earth-observation database.

**Marketing:**

Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

**City planning:**

Identifying groups of houses according to their house type, value, and geographical location.

**Earthquake studies:**

Observed earthquake epicenters should be clustered along continent faults.

**Climate:**

Understanding earth climate, find patterns of atmosphere and ocean.

# Quality: what is good clustering?

**A good clustering method will produce high-quality clusters.**

**High intra-class similarity:**

Cohesive within clusters.

**Low inter-class similarity:**

Distinctive between clusters.

**The quality of a clustering method depends on:**

the **similarity measure** used by the method,
its implementation, and
its ability to discover some or all of the hidden patterns.

## Measure the quality of clustering

**Dissimilarity/similarity metric:**

Similarity is expressed in terms of a distance function, typically a metric: $d(x, y)$.
The definitions of distance functions are usually rather different for interval-scaled, Boolean, categorical, ordinal, ratio, and vector variables (see chapter 2).
**Weights** should be associated with different variables
based on applications and data semantics.

**Quality of clustering:**

There is usually a separate **"quality" function** that measures the "goodness" of a cluster.
It is hard to define "similar enough" or "good enough."
The answer is typically highly subjective.

## Considerations for cluster analysis

**Partitioning criteria:**

Single level vs. hierarchical partitioning.
Often, multi-level hierarchical partitioning is desirable.

**Separation of clusters:**

Exclusive (e.g., one customer belongs to only one region) vs.
Non-exclusive (e.g., one document may belong to more than one class).

**Similarity measure:**

Distance-based (e.g., Euclidian, road network, vector) vs.
Connectivity-based (e.g., density or contiguity).

**Clustering space:**

Full space (often when low-dimensional) vs.
Subspaces (often in high-dimensional clustering).

## Requirements and challenges

**Scalability:**

Clustering all the data instead of only on samples.

**Ability to deal with different types of attributes:**

Numerical, binary, categorical, ordinal, linked, and mixture of these.

**Constraint-based clustering:**

User may give inputs on constraints.
Use domain knowledge to determine input parameters.

**Interpretability and usability.**

**Others:**

Discovery of clusters with arbitrary shape.
Ability to deal with noisy data.
Incremental clustering and insensitivity to input order.
High dimensionality.

## Major clustering approaches

**Partitioning approach:**

Construct various partitions and then evaluate them by some criterion.
E.g., minimizing the sum of square errors.
Typical methods: k-means, k-medoids, CLARA, CLARANS.

**Hierarchical approach:**

Create a hierarchical decomposition of the set of data (or objects) using some criterion.
Typical methods: AGNES, DIANA, BIRCH, CHAMELEON.

**Density-based approach:**

Based on connectivity and density functions.
Typical methods: DBSCAN, OPTICS, DENCLUE.

**Grid-based approach:**

Based on a multiple-level granularity structure.
Typical methods: STING, WaveCluster, CLIQUE.

## Major clustering approaches (II)

**Model-based approach:**

A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other.
Typical methods: EM, SOM, COBWEB.

**Frequent-pattern-based approach:**

Based on the analysis of frequent patterns.
Typical methods: p-Cluster.

**User-guided or constraint-based approach:**

Clustering by considering user-specified or application-specific constraints.
Typical methods: COD (obstacles), constrained clustering.

**Link-based clustering:**

Objects are often linked together in various ways.
Massive links can be used to cluster objects: SimRank, LinkClus.

# Chapter VII: Cluster analysis

## Partitioning algorithms: basic concept

**Partitioning method:**

Partition a database $D$ of $n$ objects $o_j, j \in \{1, \ldots, n\}$ into a set of $k$-clusters $C_i$, $1 \leq i \leq k$ such that the sum of squared distances to $c_i$ is minimized
(where $c_i$ is the **centroid** or **medoid** of cluster $C_i$):

$$\min \sum_{i=1}^{k} \sum_{o \in C_i} d(o, c_i)^2. \tag{1}$$

**Given $k$, find a partition of $k$ clusters that optimizes the chosen partitioning criterion.**

Globally optimal: exhaustively enumerate all partitions.
Heuristic methods: k-means and k-medoids algorithms.

**k-means** (MacQueen'67, Lloyd'57/'82):

Each cluster is represented by the center of the cluster.

**k-medoids** or PAM (Partition around medoids) (Kaufman & Rousseeuw'87):

Each cluster is represented by one of the objects in the cluster.

Thank you for your attention.

**Any questions about the seventh chapter?**

Ask them now, or again, drop me a line:

luciano.melodia@fau.de.