Data mining: = process that extracts information from a data set and transforms it into an understandable structure for further use;

- step of the Knowledge Discovery in Databases sub-field of Computer Science:

pre-processing → data mining → results validation

Common classes of tasks:
- Anomaly detection;
- Association rule learning;
- **Clustering**;
- Classification;
- Regression;
- Summarization.

**Classification**: is the supervised learning method that places a set of objects into predefined classes.

- **Given**:
  - number of classes;
  - class characteristics or a training set of labeled objects;
- **Good**:
  - when knowing the structure we are looking for in the data;
- **Bad**:
  - information on the classes is costly;
  - finding the 'natural' structure in the data.
Clustering: is the \textit{unsupervised} learning method that groups a set of objects into \textit{clusters}.

- \textbf{Given:}
  - no information on the clusters;

- \textbf{Good:}
  - exploratory analysis;
  - data compression;
  - reducing the complexity of data;

- \textbf{Bad:}
  - different fields with different structures in the data;
  - not easy finding the ‘right’ clustering algorithm.

\textbf{Applications and properties}

\textbf{Astronomy}

\textbf{Gene Expression}

Gene expression in salivary glands of Sjögren’s syndrome
Arthritis Rheum 2005; Vol.52, 5 Pages: 1534-1544

\textbf{Image Segmentation}

http://people.cs.uchicago.edu/~pff/segment/
Applications and properties

- Scalability
tenability to process very large datasets
- Attribute Types
tenability to use different types scales (numerical and categorical)
- Cluster shapes
impose no restriction on cluster shape
- Minimal requirements for domain knowledge
some algorithms require different parameters
- Noisy data
data can be missing or inaccurate
- Input order
insensibility to input order of data
- High dimensionality
ability to work with objects defined on many attributes
- Interpretability and usability
be able to read the results easily

Overview

Data characteristics

Objects: $X = \{x, y, z, \ldots\}, |X| = n$
Attributes: $F = \{i, j, k, \ldots\}, |F| = m \quad x = (x_i, x_j, x_k, \ldots)$

Measurement scales:

- Qualitative:
  - Nominal;
    (e.g. \{True, False\}, \{White, Red, \ldots\})
  - Ordinal;
    (e.g. \{Low, Medium, High\})

- Quantitative:
  - Interval;
    (e.g. Celsius temperature scale)
  - Ratio.
    (e.g. Kelvin temperature scale, Price, Distance)

Intuitive definition of clustering:
- grouping objects that are similar;
- separating objects that are dissimilar;
Similarity Modelling

Measures on quantitative scales:

- **Euclidean** (special case of Minkowski):
  - finds hyperspherical clusters;
  - attributes with large values and variances dominate others;
  - normally used in K-Means;

- **Mahalanobis**:
  - finds hyperellipsoidal clusters;

- **Pearson correlation**:
  - emphasises the number of differences, but not their magnitude;
  - used on microarray gene expression data;

- **Cosine similarity**:
  - no information on the magnitude of the differences;
  - most frequently used in document clustering.

Clustering and Multiple Criteria Decision Aid

Types of clustering approaches

Main classes of clustering algorithms:

- **Partitioning**
  - K-means,
  - Expectation
  - Maximization
  - DBSCAN,
  - GDBSCAN,
  - OPTICS,
  - DENCLUE

- **Overlapping**
  - Fuzzy C-means,
  - STING,
  - CLIQUE,
  - WaveCluster

- **Hierarchical**
  - BIRCH, CURE, ROCK,
  - CHAMELEON

- **Density-based**
  - K-means,
  - Expectation
  - Maximization

Dealing with missing values:

- Discard records with missing values;
- Compute distance on existing attributes;
- Replace missing values on each attribute with the average of the rest.

Outline:

1. Choose \( k \) points at random as cluster centres (centroids);
2. Assign each object to one centroid;
3. Calculate the new centroid for each cluster as the mean of the objects in that cluster;
4. Go back to Step 2, until there is no change in the cluster centres.

- Time complexity for each iteration of \( O(Nk) \), but the number of iterations is unknown;
- Aims at minimizing the squared error:

\[
E = \sum_{i \in 1..k} \sum_{x \in C_i} ||x - c||^2
\]
Limitations:

- requires \( k \) to be given;
  - OK when this is known a priori;
  - normally estimated from data or using heuristics (Gap, Silhouette, ...);
- doesn’t always find the optimal result (get’s stuck in local optima);
- sensitive to the initialization of the centres;
  - requires several runs of the algorithm;
  - different initialisation strategies (generate each initial centre as far away as possible from the centres already generated);
- needs objects on which the mean can be defined;
  - K-Medoids uses the median object instead of the mean;
- assumes spherical clusters (works poorly on non-convex clusters);
- sensitive to outliers.
Hierarchical Clustering

- Builds a hierarchy of nested clusters;
- 2 approaches:
  - agglomerative (bottom-up)
    - Starts with each object in a different cluster and iteratively merge two together;
  - divisive (top-down)
    - Starts with all objects in one cluster and iteratively splits one in two;
- doesn’t require the number of clusters $k$ to be given;

Dendrogram

Hierarchical Clustering

Computing distances between clusters:

- Single link;
- Complete Link;
- Centroid;
- ...

Clustering and Multiple Criteria Decision Aid

33/86

Hierarchical Clustering

Clustering and Multiple Criteria Decision Aid

34/86

Hierarchical Clustering

Clustering and Multiple Criteria Decision Aid

35/86

Hierarchical Clustering

Clustering and Multiple Criteria Decision Aid

36/86
Hierarchical Clustering

Strengths:

- good for detailed data analysis;
- provides more information than partitioning algorithms;
- deterministic (always the same result on the same data);
- can give out any number of clusters without restarting the algorithm.

Weaknesses:

- more difficult;
- no backtracking.

Overview

MCDA = aims at modelling the preferences of decision-makers and aiding them in reaching certain decisions;

objects → alternatives attributes → criteria

Preference models:

- value functions;
- outranking relations;

Main typologies of problems:

Best Choice:

- looking for the best of a given set of alternatives;
- finding the best compromise between the different objectives of each criterion;
- ex: buying a car, selecting a holiday location ...
### Overview

**Main typologies of problems:**

#### Ranking:
- Creating a complete or partial **order** between given alternatives;
- Ex: evaluating students based on grades from different subjects ...

#### Sorting:
- Given a set of predefined categories the aim is to place each alternative into one of them;
- Ex: placing students into “Good”, “Average”, “Bad” categories ...

---

#### Clustering in Data mining

= process of grouping objects that are **similar** and separating those that are **dissimilar**;

**Context:**
- Objects are defined on features, or attributes;
- Objects with similar evaluations cannot be distinguished → are grouped together;

**Properties:**
- Exploratory approach;
- Related to classification (which is a confirmatory approach);
- Many approaches: Partitioning, Hierarchical, Fuzzy, Density-based ...
- Many similarity and distance measures

---

<table>
<thead>
<tr>
<th>Situating the problem</th>
<th>Formulating the problem</th>
<th>Selecting the evaluation model</th>
<th>Constructing the final recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Alternatives</td>
<td>Decision Problem</td>
<td>Decision Objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The model</td>
<td>Value Functions</td>
</tr>
<tr>
<td>Stakes</td>
<td>Criteria</td>
<td>Ranking</td>
<td>Sorting</td>
</tr>
<tr>
<td>Resources</td>
<td></td>
<td></td>
<td>Outranking Relations</td>
</tr>
</tbody>
</table>

- directly
- indirectly
Clustering in MCDA = process of grouping alternatives that are indifferent and separating those that are strictly preferred or incomparable.

Data mining → MCDA
- objects → alternatives;
- attributes → criteria;
- similarity → indifference - $U(x) = U(y)$ or $x \ S y \land y \ S x$;
- strict preference - $U(x) > U(y)$ or $x \ S y \land y \ not S x$;
- incomparability - $x \ not S y \land y \ not S x$;
- alternatives that cannot be distinguished are similar;
- AND
- alternatives that cannot be distinguished are indifferent.

Clustering and Multiple Criteria Decision Aid
57/86

Similarity measures:
- $S^{L_1}$ - from the Manhattan distance;
- $S^{L_2}$ - from the Euclidian distance;
- $S^{thr}$ - [Bisdorff, Meyer, Olteanu 2011];
  - similarity and dissimilarity thr.;
  - step profile;
Indifference measure:
- $I(x, y) = \min \{ S(x, y), S(y, x) \}$;
  - indifference and preference thr.;
  - step profile.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^{L_1}$</td>
<td>46.5</td>
<td>73.2</td>
<td>74.5</td>
</tr>
<tr>
<td>$S^{L_2}$</td>
<td>32.8</td>
<td>67.4</td>
<td>72.0</td>
</tr>
<tr>
<td>$S^{thr}$</td>
<td>72.8</td>
<td>63.9</td>
<td>70.8</td>
</tr>
</tbody>
</table>

Clustering and Multiple Criteria Decision Aid
60/86

Types of clustering approaches

MCDA clustering approaches
- approaches that use similarity measures:
  - [Bisdorff 2002] - clusters similar criteria
  - [Valls 2000] - ordered clusters of similar alternatives
  - [Bisdorff et.al. 2011] - using a similarity measure with thresholds
  - [Rocha et.al. 2012] - partial or complete orders over groups of similar alternatives
- approaches that use preferential measures:
  - [Fernandez et.al. 2010] - clusters of indifferent alternatives
  - [de Smet, Guzman 2004], [de Smet, Eppe 2009] - relational clustering
  - [Nemery, de Smet 2005] - ordered clustering
  - [Meyer, Olteanu 2012] - ordered groups of indifferent alternatives
  - [Meyer, Olteanu 2013] - groups of indifferent alternatives ± relations between them ± relations for orders

Clustering and Multiple Criteria Decision Aid
62/86
**Types of clustering approaches**

**MCDA clustering**

- **non-relational**:
  - grouping alternatives that are indifferent;
  - separating alternatives that are not indifferent;

- **relational**:
  - grouping alternatives that are indifferent;
  - separating alternatives that are strictly preferred or incomparable;
  - one type of relation between any pair of clusters;

- **ordered**:
  - special case of relational clustering where the relations of strict preference between clusters form an order.

**Preferential relations between sets A and B:**

- using representatives;
- using all alternatives.

---

**Clustering in MCDA**

- Overview
- Similarity and Indifference
- Types of clustering approaches
- K-means on indifferences and SOMs
- CLIP
- Descriptive measures
Clustering and Multiple Criteria Decision Aid

K-means on indifferences and SOMs

The model
- $X = \{x, y, z, \ldots\}$ - set of $n$ alternatives;
- $F = \{1, \ldots, p\}$ - set of $p$ criteria;
- $x_i, \forall x \in X, i \in F$ - evaluations;
- $S$ - bipolar valued outranking relation $(r(x S y) \in [-1, 1])$;

$$C(x, y) = \begin{cases} 1, & \text{if } x_i - y_i \geq -q_i; \\ -1, & \text{if } x_i - y_i \leq -p_i; \\ 0, & \text{otherwise.} \end{cases}$$

$$r(x S y) = \min(r(x S y), r(y S x));$$

$$f_{0, K} = \sum_{l=1}^k \sum_{x \in K_l} r(x 1 c_l)$$

$\rightarrow X$ is reduced to $k$ alternatives, $C$;

- ordered clustering:
  - $K$ - set of $k$ clusters of $X$;
  - $C$ - set of $k$ representatives for $K$;
  - $K$ and $R$ are ordered from 1 to $k$;

$$f_0 = \sum_{l=1}^k \sum_{x \in K_l} r(x 1 c_l) + \sum_{l=1}^{k-1} \sum_{m=l+1}^k \sum_{x \in K_m} r(x P c_m) + \sum_{l=2}^k \sum_{m=1}^{l-1} \sum_{x \in K_l} r(c_m P x)$$

$\rightarrow X$ is reduced to $k$ alternatives, $C$, and an order btw. them.

Self-organizing maps (SOM)

- contain a set of nodes and a topology;
- objects are mapped to the closest node → clustering;
- nodes are updated from the objects mapped to them and their neighbours;
- trade-off between enforcing a model (map topology) and finding the structure in the data;
- effect of map topology decreases over time;
- useful for visualizing in few dimensions a high dimensional feature space;
- SOM with no topology = $K$-MEANS.

self-organizing maps for clustering in MCDA

Disconnected self-organizing map (DSOM)
- maps alternatives to closest node w.r.t. $I$;
- nodes are updated from the average evaluations:

$$c_l = \frac{1}{|M|} \sum_{x \in M} x_i, \forall i \in F$$

Chained self-organizing map (CSOM)
- neighbouring nodes are strictly preferred, in the order from the first to the last;
- nodes are updated as with DSOM and using the evaluations of the neighbouring alternatives:

$$c_l = \frac{|K_l| \cdot c_l + \sum_{x \in K_{l-1}} x_i + \sum_{x \in K_{l+1}} \cdot x_i}{|K_l| + |\{x \in K_{l-1}: x_i > c_l\}| + |\{x \in K_{l+1}: x_i < c_l\}|}$$

Ordered self-organizing map (OSOM)
- all nodes form an order on strict preferences.
CLIP (CLustering using Indifferences and Preferences)

1. Grouping on indifferences (internal):
   - finding an initial partition;
   - high concentration of indifference relations inside clusters;
   - low concentration of indifference relations between clusters;
   - graph theoretic inspired method using cluster cores;

2. Refining on preferences (external):
   - searching for the optimal result;
   - strengthen relations between clusters;
   - meta-heuristic approach.

Clustering in MCDA

- Overview
- Similarity and Indifference
- Types of clustering approaches
- K-means on indifferences and SOMs
- CLIP
- Descriptive measures
Descriptive measures

Central profiles \((r_C, \forall C \in K)\)
- \(r_C\) replaces the alternatives in \(C\);
- \(r_C\) is indifferent to alternatives in \(C\);

Bounding profiles \((r_C^+, r_C^-, \forall C \in K)\)
- \(r_C^+, r_C^-\) bound \(C\) from above and below;
  1) \(r_C^+\) outranks the alternatives in \(C\);
  2) \(r_C^-\) is outranked by the alternatives in \(C\);
- \(r_C^-\) first ensures 2) then 1);

Separating profiles \((r_{l^+}, \forall l \in 1..|K|)\)
- an order \(K_1 \succ K_2 \succ \ldots\) is considered;
- \(r_{l^+}\) outranks the alternatives in \(K_l\);
- \(r_{l^-}\) outranks the alternatives in \(K_{l+1}\);
- fitness measures based on \(r^*, \text{average } r, \text{ or } \min r\).

Toxic Chemical Releases

Toxics Release Inventory
- public data w.r.t. handling practices of toxic chemicals;
- over 53,000 facilities;
- records since 25 years ago;

The problem
- explore the data on the release practices
- considering intuitive preferential information on the quality of practices:
  - prefer the handling of less toxic chemicals;
  - prefer the release of fewer amounts of any chemical;
  - prefer the handling larger amounts of any chemical in mitigation processes.

Chemical toxicity
- acute diseases in humans;
- chronic diseases in humans;
- cancer diseases in humans;
- environmental hazard;

Chemical handling
- releases to air;
- releases to water;
- releases to landfills;
- underground injection;
- releases off-site;
- amounts recycled;
- amounts transferred;
- amounts burned;
- amounts treated.
Selecting the data

- 242 chemicals reported in 2010 on which TRI has complete information w.r.t. the 4 indicators;
- 22,676 reports were made in 2010 (without invalid entries);

Structuring the problem

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Criteria</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toxicity (t)</td>
<td>Toxicity to Humans (TH)</td>
<td>Cancer Indicator (CAN)</td>
</tr>
<tr>
<td>Toxicity to Environment (TE)</td>
<td>Environmental Indicator (ENV)</td>
<td></td>
</tr>
<tr>
<td>Releases (r)</td>
<td>Release Amounts (RA)</td>
<td>Air Release Amounts (AIR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Water Release Amounts (WAT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Land Release Amounts (LAN)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Underground Injection Amounts (UND)</td>
</tr>
<tr>
<td>Mitigation (m)</td>
<td>Mitigated Amounts (MA)</td>
<td>Transferred to POTW Amounts (POT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Converted to Energy Amounts (ENE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treated Amounts (TRE)</td>
</tr>
</tbody>
</table>

Considering a fictive DM

- $F = \{\text{TH, TE, RA, MA}\}$ are considered independent;
- $\text{CAN} \succ \text{CHR} \succ \text{ACU}$
- $\text{AIR} \succ \text{WAT} \succ \text{LAN} \succ \text{UND} \succ \text{OFF}$
- $\text{REC} \succ \text{POT} \succ \text{ENE} \succ \text{TRE}$

- $\tilde{S}_{LDP}$ (Bisdorff 2012);
- $w_{\text{TH}} = 0.2$, $w_{\text{TE}} = 0.1$, $w_{\text{RA}} = 0.3$ and $w_{\text{MA}} = 0.3$;
- indifference thresholds set at $10 + 0.1 \cdot x_i$, $\forall x \in X, \forall i \in F$; (0 in the case of CAN, CHR, ACU, ENV)
- preference thresholds set at $20 + 0.2 \cdot x_i$, $\forall x \in X, \forall i \in F$; (1 in the case of CAN, CHR, ACU, ENV)
- no large performance difference thresholds;
- median cut (for extracting $r^*$);

Non-relational clustering

<table>
<thead>
<tr>
<th>Fitness (%)</th>
<th>$f_{\text{NR}}^*$</th>
<th>$f_{\text{NR}}$</th>
<th>$f_{\text{NR}}^\min$</th>
<th>$f_{\text{R}}^*$</th>
<th>$f_{\text{R}}$</th>
<th>$f_{\text{R}}^\min$</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0</td>
<td>62.7</td>
<td>0.0</td>
<td>64.0</td>
<td>55.2</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

Cluster sizes

- $K_1$ 296
- $K_2$ 1,429
- $K_3$ 1,632
- $K_4$ 6,237
- $K_5$ 2,973
- $K_6$ 167
- $K_7$ 1,316
- $K_8$ 2,615
- $K_9$ 1,688
- $K_{10}$ 540
- $K_{11}$ 3,518
- $K_{12}$ 356