Chapter DM:I

I. Introduction

- Data Mining Overview
- On Data
Data Mining Overview

Definition 1 (Knowledge Discovery in Databases, KDD [Fayyad 1996, Wrobel 1998])

Knowledge Discovery in Databases is the process of identifying valid, new, relevant, and interpretable patterns in huge data sets.
Definition 1 (Knowledge Discovery in Databases, KDD) [Fayyad 1996, Wrobel 1998)
Knowledge Discovery in Databases is the process of identifying valid, new, relevant, and interpretable patterns in huge data sets.

Definition 2 (Data Mining)
Data mining is the systematic, usually automated or semi-automated discovery and extraction of so far unknown relations from huge data sets.
Data mining technology belongs to the field of *explorative data analysis*. Explorative data analysis deals with both data presentation and search for structures, peculiarities, and anomalies. It is applied if the research question is fuzzy or if the choice of the statistical model is unclear.

The data mining definition does not use the notion of “information”: under the viewpoint of semiotics, data mining operates on the sigmatic layer only. The *interpretation* of discovered patterns, i.e., the examination of information with regard to new findings and a subjective knowledge gain, which happens on the pragmatic layer, belongs to the field of KDD.

In the business world, the terms data mining and knowledge discovery in databases, KDD, are used synonymously. Note however, that data mining is only a *single step* within a KDD process, namely the analysis step for pattern recognition.

Web data mining is the transfer and usage of data mining technology for information extraction on the Internet and especially the World Wide Web. Text mining is the identification of relevant information in text.
## Data Mining Overview

### KDD, Knowledge Discovery in Databases

<table>
<thead>
<tr>
<th>Data mining</th>
<th>Web mining, Text mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario:</td>
<td>gigabytes, databases, on the (semantic) Web, in unstructured text</td>
</tr>
</tbody>
</table>

### Machine learning

<table>
<thead>
<tr>
<th>Statistic analysis</th>
<th>clean data, hypothesis evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario:</td>
<td>in main memory, specific deduction model</td>
</tr>
</tbody>
</table>

### Explorative data analysis

### OLAP, Online Analytical Processing
## Data Mining Overview

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Information visualization</th>
<th>Data aggregation...</th>
<th>Descriptive data analysis</th>
<th>Explorative data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLAP, Online Analytical Processing</td>
<td>KDD, Knowledge Discovery in Databases</td>
<td>Data mining, Web mining, Text mining</td>
<td>Explorative data analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text</td>
<td>Data mining, Web mining, Text mining</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scenario: in main memory, specific deduction model</td>
<td>Explorative data analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Statistic analysis</td>
<td>Explorative data analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scenario: clean data, hypothesis evaluation</td>
<td>Explorative data analysis</td>
</tr>
</tbody>
</table>

**Data mining**, Web mining, Text mining

**Machine learning**

**Statistic analysis**

**Explorative data analysis**

**KDD, Knowledge Discovery in Databases**

**Online Analytical Processing**

**Information visualization**

**Data aggregation**

**Descriptive data analysis**

**Explorative data analysis**

**Statistic analysis**

**Scenario:**

- Gigabytes, databases, on the (semantic) Web, in unstructured text
- In main memory, specific deduction model
- Clean data, hypothesis evaluation
Data Mining Overview

- **KDD, Knowledge Discovery in Databases**
- **Data mining**, Web mining, Text mining
- **Machine learning**
  - Scenario: in main memory, specific deduction model
- **Statistic analysis**
  - Scenario: clean data, hypothesis evaluation
- **OLAP, Online Analytical Processing**
- **Explorative data analysis**
- **Descriptive data analysis**
- **Information visualization**
- **Data aggregation**

**Analysis**

**Pragmatics**

- Semantics ➔ knowledge

**Sigmatics**

- data

**Semiotics layer**

**Syntax**
Data Mining Overview

Retrieval

Information retrieval, Information extraction
Structured query processing

Analysis

Information visualization
Data aggregation...

OLAP, Online Analytical Processing
KDD, Knowledge Discovery in Databases

Data mining, Web mining, Text mining
Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text

Machine learning
Scenario: in main memory, specific deduction model

Statistic analysis
Scenario: clean data, hypothesis evaluation

Descriptive data analysis
Explorative data analysis

Information need

Pragmatics
Semantics ➔ knowledge

Sigmatics ➔ data

Syntax

Semiotics layer
A clear separation between machine learning and data mining is not always possible. A key difference, however, results from the sizes of the analyzed data sets: machine learning applications may be executed in main memory. The field of data mining arose from the necessity to apply analysis methods to large data bases.

The foci of machine learning are the processes and theories of learning and deduction, such as analogical reasoning, learning from examples, or reinforcement learning. The major driving force behind data mining is the business world with their large data bases.

The following count to relevant data mining problems:

- undirected association analysis to identify dependencies between consumer products (market basket analysis)
- grouping and categorization
- filtering of process data
- forecasting and prediction
Data Mining Overview

Relevant Data Mining Methods

- Cluster analysis

- Learning of propositional or description-logical rules – example:

  IF status=married AND house_owner=true THEN creditor=good

- Learning of association rules – example:

  “75% of the buyers of product A will buy the products B, C, and D as well.”

- Principal component analysis (PCA), factor analysis

- Multi-dimensional scaling (MDS)