C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

INF-KDDM: Knowledge Discovery and Data Mining

Winter Term 2020/21

Lecture 2: Data preprocessing and feature spaces

Lectures: Prof. Dr. Matthias Renz

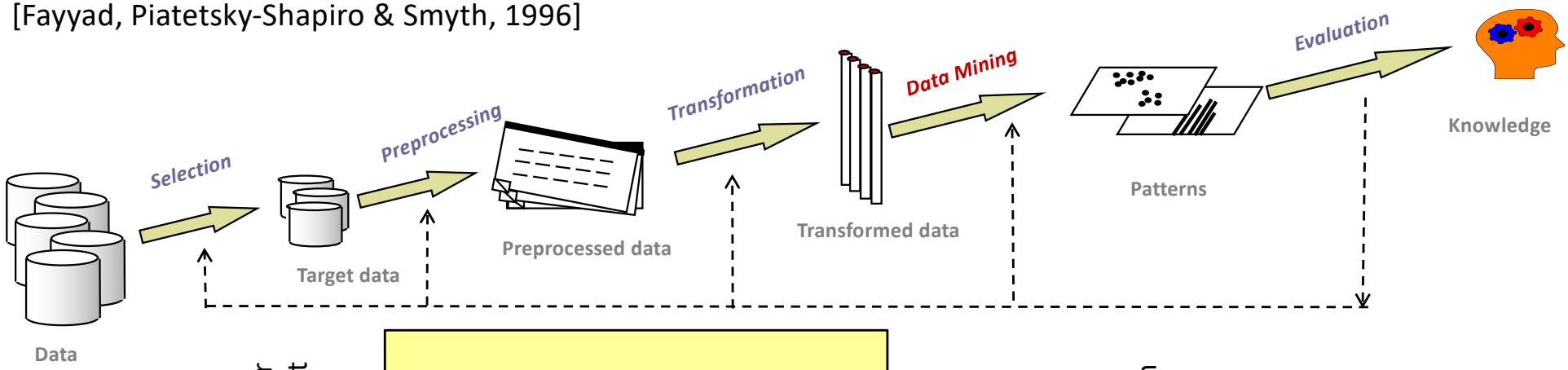
Exercises: Niko Amann

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data

Recap: The KDD process

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Data Mining:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

Why data preprocessing and transformation?

- Real world data are noisy, incomplete and inconsistent:
  - Noisy: errors/ outliers
    - erroneous values : e.g. salary = -10K
    - unexpected values: e.g. salary=100K when the rest dataset lies in [30K-50K]
  - Incomplete: missing data
    - missing values: e.g., occupation=“ ”
    - missing attributes of interest: e.g. no information on occupation
  - Inconsistent: discrepancies in the data
    - e.g. student grade ranges between different universities might differ, in DE [1-5], in GR [0-10]
- “Dirty” data → poor mining results
- Data preprocessing is necessary for improving the quality of the mining results !
- Data preprocessing techniques are Not a focus of this class!

Know your data!

Major tasks in data preprocessing and transformation

- Data cleaning:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

    → Some of these data cleaning tasks itself can be supported by Data Mining tasks
- Data integration:
  - Integration of multiple databases, data cubes, or files (Entity Resolution / Value Resolution)
- Data transformation:
  - Normalization in a given range, e.g., [0-1]
  - Generalization through some concept hierarchy, e.g. "*milk 1.5% brand x*"→"*milk 1.5%*" or "*milk*"
- Data reduction:
  - Aggregation, e.g., from 12 monthly salaries to month's average salary.
  - Dimensionality reduction, through e.g., PCA
  - Duplicate elimination

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data
- Homework/ Tutorial
- Things you should know from this lecture