Chapter 4:
Data Mining Knowledge Representation

—

Based on lecture notes of Zdravko Markov, PhD

Data Mining tasks

1. Task relevant data
2. Background knowledge
3. Interestingness measures
4. Representing input data and output knowledge
5. Visualization techniques

Task relevant data

Where to find the data?

- Database or data warehouse name
- Database tables or data warehouse cubes

How to retrieve the data?

- Using conditions for data selection: relevant attributes or dimensions and data grouping criteria

Background knowledge

Induced by a partial order over the values of a given attribute.

Examples:

- street $<$ city $<$ state $<$ country
- $\{13, ..., 39\}$ = young; $\{13, ..., 19\}$ = teenage; $\{13, ..., 19\} \subseteq \{13, ..., 39\} \Rightarrow$ teenage $<$ young

Interestingness measures

- Confidence of association “if A then B”
- Confidence of association “if A then B”
- Classification Accuracy

Representing input data and output knowledge

Things to be mined/learned (Concepts)

- Classification mining/learning:
  Predicting a discrete class
- Association mining/learning:
  Detecting associations between attributes
- Clustering:
  Grouping similar instances into clusters
- Numeric prediction:
  Predicting a numeric quantity

Representing input data and output knowledge

Input

- Set of instances (dataset), represented as a single relation (table)
  - Individual, independent examples of the concept to be learned
  - Described by predetermined set of attributes
- Attributes:
  - Predefined set of features to describe an instance
  - Nominal (categorical, enumerated, discrete) attributes:
    - Values are distinct symbols.
    - No relation among nominal values.
  - Ordinal attributes:
    - Partial order among nominal values
  - Numeric attributes:
    - Integer/real number

Representing input data and output knowledge

Output knowledge representation

- Association rules
- Decision trees
- Classification rules
- Rules with relations
- Prediction schemes:
  - Nearest neighbor
  - Bayesian classification
  - Neural networks
  - Regression
- Clusters:
  - Type of grouping: partitions/hierarchical
  - Grouping or describing: agglomerative/conceptual

Visualization techniques

Why?

- Better understanding the data
- Identifying problems, e.g., detecting outliers
- Identifying dependencies
- Checking the assumptions
- Consulting domain experts
- If data are too much, take a sample

References

Markov, Zdravko. (n.d.). Lecture on Data mining knowledge representation. Central Connecticut State University.