Data Mining

Part 2. Data Preprocessing

2.1 Introduction

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- Why Data Preprocessing?
- Major Tasks in Data Preprocessing
- References

Why Data Preprocessing?

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=" ", martial status =" "
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing inconsistencies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., inconsistency between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - It was not considered important at the time of entry
  - Human/hardware/software problems

- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission

- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Data Preprocessing

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
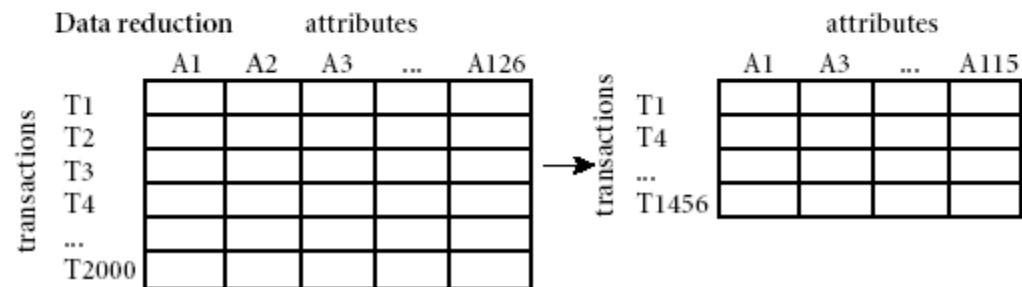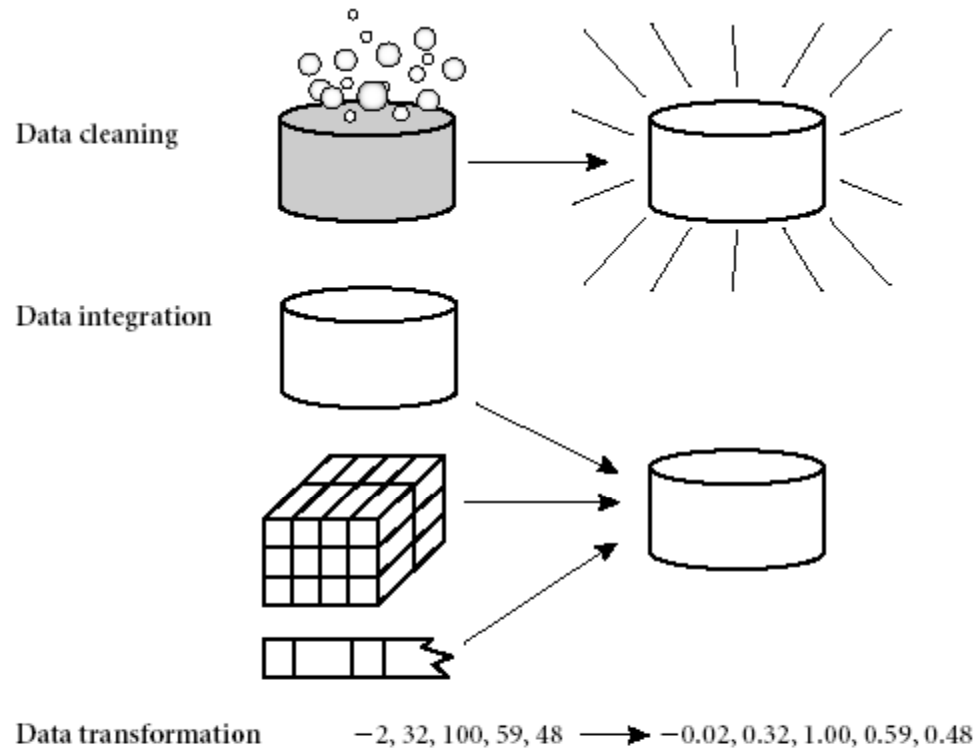
Major Tasks in Data Preprocessing

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values
  - Identify or remove outliers
  - Resolve inconsistencies
  - Schema integration
  - Handling redundancy
- Data transformation
  - Normalization and aggregation
  - For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.

Major Tasks in Data Preprocessing

- Data reduction
  - reduces the data size by aggregating
  - eliminating redundant features
  - obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing

Data cleaning

Data integration

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data reduction

	attributes						attributes			
	A1	A2	A3	...	A126		A1	A3	...	A115
T1						T1				
T2						T4				
T3						...				
T4						T1456				
...										
T2000										

transactions (left) / transactions (right)

Data Preprocessing

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end