Data Warehousing and Data Mining

M.Tech. Part-I 2nd Semester, 2011

CSE-205E

Dr. Anirban Mukhopadhyay
Assistant Professor
Department of Computer Science and Engineering
University of Kalyani

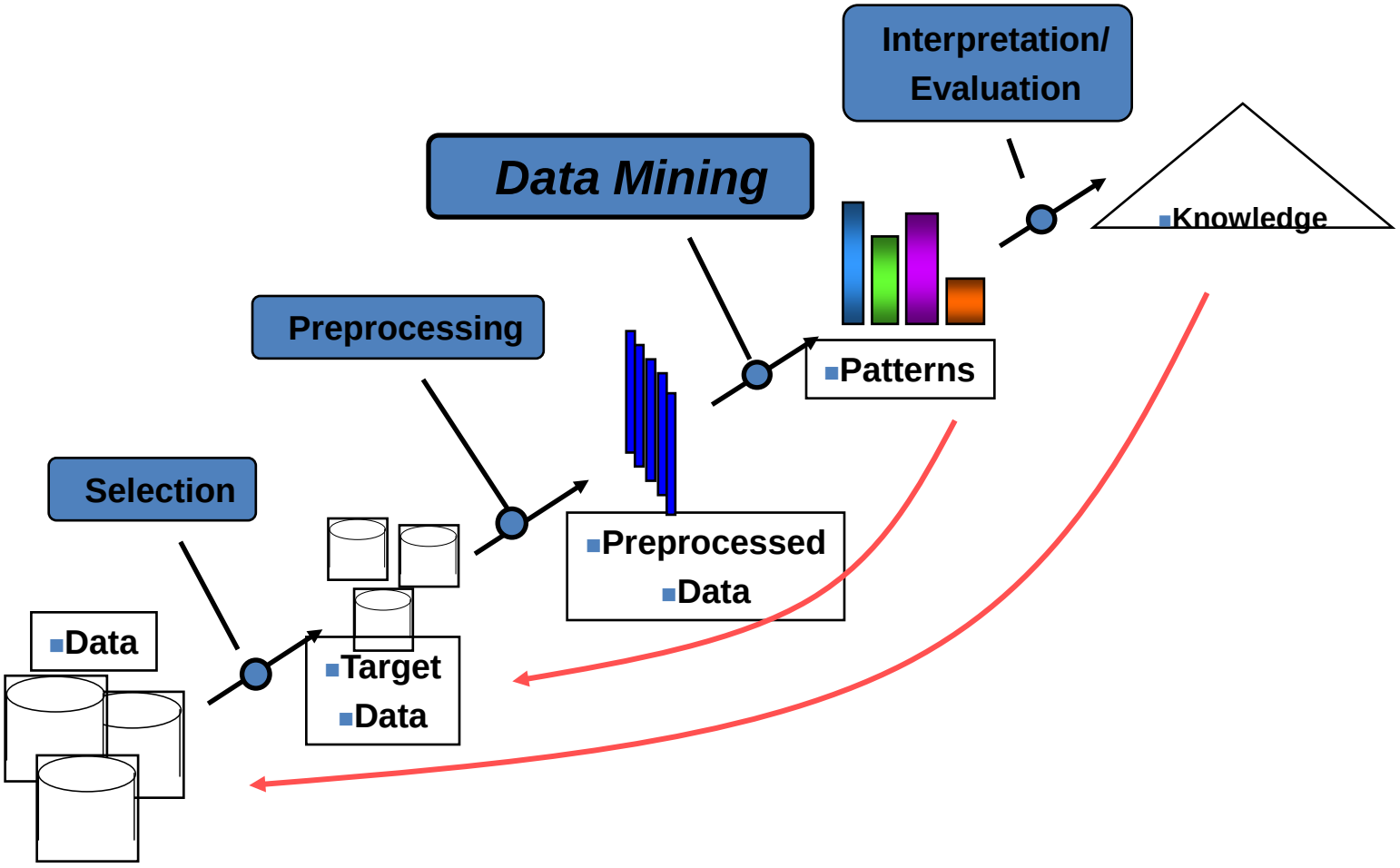
Topics

- Introduction
- Data pre-processing
- Association rules and sequential patterns
- Clustering
- Regression
- Classification
- Deviation detection
- Feature selection
- Post-processing of data mining results
- Applications

What is data mining?

- Data mining is also called *knowledge discovery from Databases* (KDD)
- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, images, etc.
- Patterns must be:
  - valid, novel, potentially useful, understandable

KDD steps


Interpretation/
Evaluation
Data Mining
Knowledge
Preprocessing
Patterns
Selection
Preprocessed
Data
Data
Target
Data

Classic data mining tasks

- **Regression:**
  investigation of relationships between different sets of variables.
- **Classification:**
  mining patterns that can classify future (new) data into known classes.
- **Association rule mining**
  mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items. E.g.,
  Cheese, Milk$\rightarrow$ Bread [sup =5%, confid=80%]
- **Clustering**
  identifying a set of similarity groups in the data

Classic data mining tasks (contd)

- Sequential pattern mining:

  A sequential rule: $A \rightarrow B$, says that event *A* will be immediately followed by event *B* with a certain confidence

- Deviation detection:

  discovering the most significant changes in data

- Data visualization: using graphical methods to show patterns in data.

Why is data mining important?

- Computerization of businesses produce huge amount of data
  - How to make best use of data?
  - Knowledge discovered from data can be used for competitive advantage.
- Online e-businesses are generate even larger data sets
  - Online retailers (e.g., amazon.com) are largely driving by data mining.
  - Web search engines are information retrieval (text mining) and data mining companies

Why is data mining necessary?

- Make use of your data assets
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Many interesting things that one wants to find cannot be found using database queries

  "find people likely to buy my products"

  "Who are likely to respond to my promotion"

  "Which movies should be recommended to each customer?"

Why data mining?

- The data is abundant.
- The computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every company is doing (or has to do) it

Related fields

- Data mining is an multi-disciplinary field:
  - Machine learning
  - Statistics
  - Databases
  - Information retrieval
  - Visualization
  - Natural language processing
  - etc.

Data mining (KDD) process

- Understand the application domain
- Identify data sources and select target data
- Pre-processing: cleaning, attribute selection, etc
- Data mining to extract patterns or models
- Post-processing: identifying interesting or useful patterns/knowledge
- Incorporate patterns/knowledge in real world tasks

Data mining applications

- **Marketing:** customer profiling and retention, identifying potential customers, market segmentation.
- **Engineering:** identify causes of problems in products.
- **Scientific data analysis**, e.g., bioinformatics
- **Fraud detection:** identifying credit card fraud, intrusion detection.
- **Text and web:** a huge number of applications …
- Any application that involves a large amount of data …