Knowledge Discovery in Data Bases

Chien-Chung Chan
Department of CS
University of Akron
Akron, OH 44325-4003
Why KDD?

We are drowning in information, but starving for knowledge —— John Naisbett

Growing Gap between Data Generation and Data Understanding:

Automation of business activities:

Telephone calls, credit card charges, medical tests, etc.

Wal-Mart databases: over 20 millions of transactions per day [Babcock, 1994]

Scientific and government databases:

Earth observation satellites:
Estimated will generate one terabyte ($10^{15}$ bytes) of data per day. At a rate of one picture per second.

Biology: Human Genome database project has collected over gigabytes of data on the human genetic code [Fasman, Cuticchia, Kingsbury, 1994.

US Census data:
NASA databases:
…

World Wide Web:
What is KDD?

**KDD** refers to the overall process of finding and interpreting useful patterns from data.

**Data mining** refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.
Steps of the KDD Process [Fayyad et al, 1996]

1. Selection:
   Learning the application domain:
   Creating a target dataset:

2. PreProcessing:
   Data cleaning and preprocessing:

3. Transformation:
   Data reduction and projection:

4. Data Mining:
   Choosing the functions and algorithms of data mining:
   Association rules, classification rules, clustering rules

5. Interpretation and Evaluation:
   Validate and verify discovered patterns:

6. Using discovered knowledge:
Related Fields

Statistics
   Model Evaluation and Validation

Machine learning and pattern recognition
   Symbolic and neural network learning

Databases
   Data warehousing
   OLAP (online analytical processing)

Artificial intelligence
   Knowledge representation

Data visualization
   Report generation
**Typical Data Mining Tasks**

1. Finding Association Rules [Rakesh Agrawal et al, 1993]

   Each *transaction* is a set of items.

   Given a set of transactions, an association rule is of the form

   \[ X \Rightarrow Y \]

   where \( X \) and \( Y \) are sets of items.

   Intuitive meaning: transactions of the database which contain \( X \) tends to contain \( Y \).

   e.g.: 30% of transactions that contain beer also contain diapers;
   2% of all transactions contain both of these items.

   Here 30% is called the *confidence* of the rule,
   2% is called the *support* of the rule.

   **Problem:** Find all association rules that satisfied user-specified minimum support and minimum confidence constraints.

   **Potential Problem:**
   Too many rules may be generated.
   Using an “interesting” measure to prune redundant rules.

   **Applications:**
   Market basket analysis and cross-marketing
   Catalog design
   Store layout
   Buying patterns
2. Finding Sequential Patterns

Each data sequence is a list of transactions.

Find all sequential patterns with a user-specified minimum support.

e.g.: Consider a book-club database
A sequential pattern might be
5% of customers bought “Foundation”, then “Foundation and Empire”, and then “Second Foundation”.

Applications:
Add-on sales
Customer satisfaction
Identify symptoms/diseases that precede certain diseases

3. Finding Classification Rules
Finding discriminant rules for objects of different classes.

Approaches:
Finding Decision Trees
Finding Production Rules

Applications:
Process loans and credit cards applications
Model identification
Information Systems
An information system is a pair \((U, A)\)

Data Tables
KDD Products

**IBM Intelligent Miner**
To support data-intensive decision-support applications.

The Quest Data Mining System [Rakesh Agrawal et al, 1996], IBM Almaden Research Center:

**Tasks:**
- Association rules
- Sequential patterns
- Time-series clustering
- Classification
- Incremental mining

**Goals:**
- Discover patterns in very large databases, rather than simply verify that a pattern exists
- Have a completeness property that guarantees that all patterns of certain types have been discovered
- Have high performance and near-linear scaling on very large (multiple gigabytes) real-life databases
**DBMiner** [Han et al, Simon Fraser University]
Interactive mining of multiple-level knowledge in relational databases.

Tasks:
- Generalization
- Characterization
- Association
- Classification
- Prediction
Some Challenges in KDD

Limitations of collected data [G. Wiederhold, 1996]
1. Limited breadth and coverage:
   e.g.: they cover only the population that has a certain credit card.

2. Limited depth: essential variables may be missing.
   e.g.: in most clinical studies no data on diet is available, although
   for many medical problems food and drink are significant factors.

3. Data mining based on correlation alone can lead to nonsensical
   results.
   e.g.: “Drinking diet drinks leads to obesity.” [P. Bonissone]

Heterogeneous Data Formats
   DataBases
      Relational
      Object-Oriented

   Web Sites and Pages

   Flat Files (File Systems)

Combining Multiple Data Sources
   Reasoning with uncertainty

KDDMS (Knowledge and Data Discovery Management System) [Imielinski, 1996]

   KDD Query Languages