Data Mining Functionalities

- Multidimensional concept description: Characterization and discrimination

  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

- Frequent patterns, association, correlation vs. causality

  - Diaper → Beer [0.5%, 75%] (Correlation or causality?)

- Classification and prediction

  - Construct models (functions) that describe and distinguish classes or concepts for future prediction

    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)

  - Predict some unknown or missing numerical values

Data Mining Functionalities

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

GALGOTIAS UNIVERSITY

Data Mining Function: Generalization

- Information integration and data warehouse construction

    - Data cleaning, transformation, integration, and multidimensional data model

- Data cube technology

    - Scalable methods for computing (i.e., materializing) multidimensional aggregates

    - OLAP (online analytical processing)

- Multidimensional concept description: Characterization and discrimination

    - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Function: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)

    - What items are frequently purchased together in your Walmart?

- Association, correlation vs. causality

    - A typical association rule

        - Diaper → Beer [0.5%, 75%] (support, confidence)

    - Are strongly associated items also strongly correlated?

- How to mine such patterns and rules efficiently in large datasets?

- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: Classification

- Classification and label prediction
    - Construct models (functions) based on some training examples
    - Describe and distinguish classes or concepts for future prediction
        - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
    - Predict some unknown class labels
- Typical methods
    - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
    - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)

- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns

- Principle: Maximizing intra-class similarity & minimizing interclass similarity

- Many methods and applications

Data Mining Function: Outlier Analysis

- Outlier analysis

  - Outlier: A data object that does not comply with the general behavior of the data

  - Noise or exception? — One person's garbage could be another person's treasure

  - Methods: by product of clustering or regression analysis, …

  - Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- Are all mined knowledge interesting?

  - One can mine tremendous amount of "patterns" and knowledge

  - Some may fit only certain dimension space (time, location, …)

  - Some may not be representative, may be transient, …

- Evaluation of mined knowledge → directly mine only interesting knowledge?

  - Descriptive vs. predictive

  - Coverage

  - Typicality vs. novelty

  - Accuracy

  - Timeliness

  - …