CS5228 Knowledge Discovery and Data Mining

AY2024/25 Sem2 By Zhao Peiduo

Lecture 1

Common Data Mining Tasks Data mining encompasses various techniques to analyze and extract patterns from large datasets. Some of the most common data mining tasks include:

- **Association Rules:** This method analyzes *transactional data*, where a transaction is a data record consisting of a set of items from a fixed collection. The goal is to identify *association rules* that predict the occurrence of items based on the presence of other items in the dataset.
- **Clustering:** Clustering involves grouping data points based on a well-defined notion of similarity. The objective is to form *clusters*, ensuring that data points within the same cluster have high intra-cluster similarity while minimizing inter-cluster similarity with other clusters.
- **Classification:** This method uses datasets with multiple attributes to determine the *categorical value* of an attribute as a function of other attribute values. Popular classification techniques include K-Nearest Neighbor, Decision Trees, and Linear Classification.
- **Regression:** Similar to classification, regression also works with datasets having multiple attributes, but it predicts *numerical values* of an attribute as a function of other attributes. Common regression methods include K-Nearest Neighbor, Regression Trees, and Linear Regression.
- **Graph Mining:** This technique analyzes data represented as a graph, $G = (V, E)$, where V represents data points (vertices) and E represents relationships between them (edges). Typical patterns derived from graph mining include identifying communities of nodes and detecting important nodes within the network.
- **Recommender Systems:** Recommender systems work with *user-rated items* (such as movie ratings) to predict missing values and recommend items based on similarities. They exploit item features and user similarities to enhance recommendations.

Types of Attributes

- **Categorical (Qualitative):**
  - **Nominal:**
    - Values are only labels.
    - Operations: $=, \neq$
    - Examples: sex (m/f), eye color, zip code.
  - **Ordinal:**
    - Values are labels with a meaningful order.
    - Operations: $=, \neq, <, >$
    - Examples: street numbers, education level.
- **Numerical (Quantitative):**
  - **Interval:**
    - Values are measurements with a meaningful distance.
    - Operations: $=, \neq, <, >, +, -$
    - Examples: body temperature in °C, calendar dates.
  - **Ratio:**
    - Values are measurements with a meaningful ratio.
    - Operations: $=, \neq, <, >, +, -, \times, \div$
    - Examples: age, weight, income, blood pressure.

Types of Data

Data can be classified into three main types based on structure and organization:

- **(Well-)Structured Data:**
  - Highly organized: adheres to a predefined data model.
  - Each object has the same fixed set of attributes.
  - Easy to search, aggregate, manipulate, and analyze.
  - Examples: relational databases, spreadsheets.
- **Semi-Structured Data:**
  - No rigid data model: mix of structured and unstructured data.
  - Data exchange formats: XML, JSON, CSV.
  - Tagged unstructured data (e.g., photo with date/time, location, exposure, resolution, flash, etc.).
- **Unstructured Data:**
  - No fixed data model.
  - Requires more advanced data analysis techniques.
  - Examples: images, videos, audio, text, social media.

Data Quality

- **Noise**: Data can be defined as: true signal + **noise**. Sources of noise include:
  - Sensor readings from faulty devices (e.g., intrinsic noise or external influences).
  - Errors during data entry (by humans or machines).
  - Errors during data transmission.
  - Inconsistencies in data formats (e.g., ISO time vs. Unix time, DD/MM/YYYY vs. MM/DD/YYYY).
  - Inconsistencies in conventions (e.g., meters vs. miles, meters vs. centimeters).
- **Outliers**:An outlier is a data point with attribute values considerably different from other points.Outliers can be classified into:
  - **Outliers as noise:**
    - They negatively interfere with data analysis.
    - Removal of outliers or using robust methods is recommended.
  - **Outliers as targets:**
    - The goal is to detect rare or anomalous events such as credit card fraud detection and intrusion detection in security systems.
- **Missing Values**
  - Common causes of missing values:
    - Attribute values not collected (e.g., broken sensor, person refused to report age).
    - Attributes not applicable in all cases (e.g., no income data for children).
  - Handling missing values:
    - Remove data points with missing values.
    - Remove attributes with missing values (if not essential).
    - Try to fill in missing values (e.g., using average temperature from nearby sensors).
- **Duplicates**
  - Duplicates refer to data points representing the same object/entity.
    - **Exact duplicates:** Data points have identical attribute values.
    - **Near duplicates:** Data points slightly differ in their attribute values (e.g., same person with phone numbers in different formats).
  - Duplicate elimination:
    - Relatively easy for exact duplicates.
    - Challenging for near duplicates.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in data analysis to identify potential issues such as noise, outliers, missing values, and class distribution imbalances.

- **Identifying Noise**
  - Using histograms to inspect the distribution of data values.
- **Identifying Noise / Outliers**
  - Using box plots to inspect the distribution of attribute values.
    - Make outliers explicit.
  - Using scatter plots to inspect correlations.
    - Not always feasible in practice.
    - Requires good understanding of data.
- **Handling Missing Values**
  - Example: Default value (0) if people did not disclose weight.
    - Can negatively affect simple analysis such as calculating means/averages.
- **Distribution of Class Labels**
  - Classification tasks generally benefit from balanced datasets.
    - Balanced = all classes are (almost) equally represented.
    - Distribution of classes also affects the evaluation of found patterns.

Data Preprocessing

- **Main Purposes**
  - Improve data quality (*"Garbage in, garbage out!"*).
  - Generate valid input for data mining algorithms.
  - Remove complexity from data to ease analysis.
- **Core Preprocessing Tasks**
  - Data cleaning
  - Data reduction
  - Data transformation
  - Data discretization
- **Data Cleaning**
  - Remove or fill missing values.
  - Identify and remove outliers (if outliers are not the goal of the analysis).
  - Identify and remove/merge duplicates.
  - Correct errors and inconsistencies (e.g., convert inches to centimeters).
  - *Non-trivial tasks that are typically very application-specific.*
- **Data Reduction**
  - **Reducing the number of data points**
    - Sampling — selecting a subset of data points (random or stratified sampling).
    - Used for preliminary analysis or large datasets.
  - **Reducing the number of attributes**
    - Removing irrelevant attributes (e.g., IDs, sensitive attributes).
    - Dimensionality reduction (PCA, LDA, t-SNE).
  - **Reducing the number of attribute values**
    - Aggregation or generalization.
    - Binning with smoothing.
- **Data Transformation**
  - Some data reduction techniques also transform data (e.g., dimensionality reduction, aggregation, binning).
  - Attribute construction:
    - Add or replace attributes inferred from existing attributes.
    - Example: weight, volume $\rightarrow$ density.
  - Normalization:
    - Scaling attribute values to a specified range (e.g., [0,1]).
    - Standardization: scaling using mean and standard deviation.
- **Data Discretization**
  - Converting continuous attributes into ordinal attributes.
  - Some algorithms accept only categorical attributes.
  - Convert a regression task to a classification task.
- **One-Hot Encoding**
  - Converting categorical attributes into numerical attributes.
  - Transform categorical attributes into binary attributes (0/1).
  - Allows the application of numerical methods on categorical attributes.