Lecture-3

Data Mining

Introduction

(Data Mining Functionalities)

Department of Computer Science and Engineering
ITER, Siksha ‘O’ Anusandhan University.

Content

- Data Mining Functionalities
  - Concept/Class Description: Characterization and Discrimination
  - Mining Frequent Patterns, Associations, and Correlations
  - Classification and Prediction
  - Cluster Analysis
  - Outlier Analysis
  - Evolution Analysis
- Summary

Data Mining Functionalities

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- data mining tasks can be classified into two categories: descriptive and predictive.

  - Descriptive mining: tasks characterize the general properties of the data in the database.
  - Predictive mining: tasks perform inference on the current data in order to make predictions.

Data Mining Functionalities [Cont..]

➢Data mining functionalities includes:

- Concept/Class Description: Characterization and Discrimination
- Mining Frequent Patterns, Associations, and Correlations
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis

Concept/Class Description: Characterization and Discrimination

➢Data characterization

- Data characterization is a summarization of the general characteristics or features of a target class of data.
- The data corresponding to the user-specified class are typically collected by a database query.
- Example: to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected by executing an SQL query.

Concept/Class Description: Characterization and Discrimination[Cont..]

➢Data discrimination

- Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.
- The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.
- For example, the user may like to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.

Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns
  - Patterns that occur frequently in data
- Kinds of Frequent Pattern
  - ✓ Frequent itemset: refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.
  - ✓ Frequent sequential pattern: A frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card.
  - ✓ Frequent structured pattern: A substructure can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets or subsequences which occurs frequently.

Mining Frequent Patterns, Associations, and Correlations [Cont..]

- Association analysis
- Multi-dimensional vs. single-dimensional association
  - Single-dimensional association rules
  - Association rules that contain a single predicate

Example: *buys*(*X*; “*computer*”)→*buys*(*X*; “*software*”) [*support* = 1%, *confidence* = 50%]

- *Here X* is a variable representing a customer.
- A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.
- 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

Mining Frequent Patterns, Associations, and Correlations [Cont..]

➢Multidimensional association rule

- Association rules that contain multiple predicate.
- In multidimensional databases each attribute is referred to as a dimension.

- *Example: age*(*X*, “20….29”)^*income*(*X*, “20K…..29K”)→*buys*(*X*, “*CD player*”) [*support* = 2%, *confidence* = 60%]

Classification and Prediction

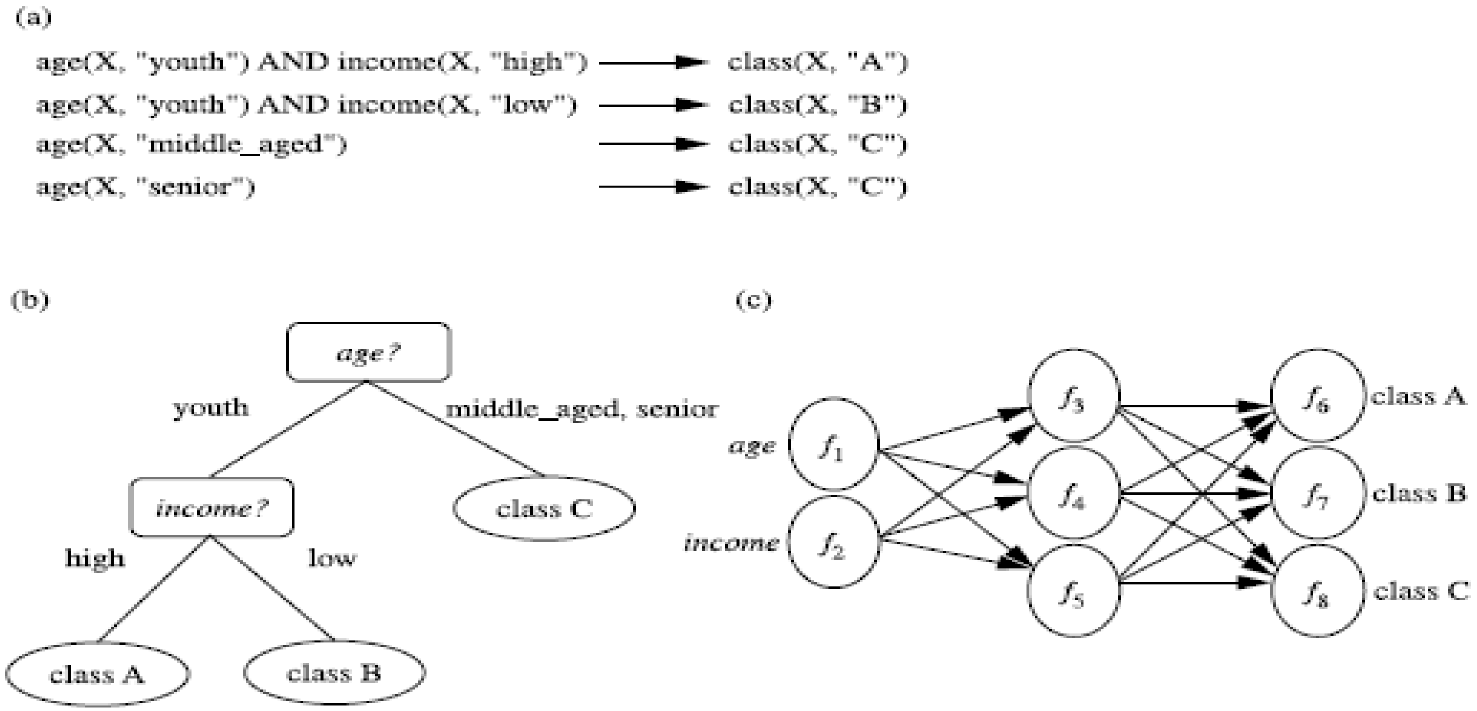
➢ Classification

- It is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

➢ How is the derived model presented

- Represented by Classification method such as:
  - (IF-THEN) rules
  - decision trees
  - mathematical formulae
  - neural networks

Classification and Prediction [Cont..]



(a)

age(X, "youth") AND income(X, "high") → class(X, "A")

age(X, "youth") AND income(X, "low") → class(X, "B")

age(X, "middle_aged") → class(X, "C")

age(X, "senior") → class(X, "C")

Figure 1. A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

Classification and Prediction [Cont..]

➢Prediction

- Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions.
- That is, it is used to predict missing or unavailable numerical data values rather than class labels.
- Regression analysis is a statistical methodology that is most often used for numeric prediction.

Cluster Analysis

➢Clustering

- Clustering analyzes data objects without consulting a known class label.
- The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity.
- Each cluster that is formed can be viewed as a class of objects, from which rules can be derived.

Outlier Analysis

➢ Outlier

- The data objects that do not comply with the general behavior or model of the data.
- The analysis of outlier data is referred to as outlier mining.
- Outliers may be detected by:
  - using statistical tests
  - using distance measures

Evolution Analysis

- Data evolution analysis
  - Describes and models regularities or trends for objects whose behavior changes over time.
  - This may include:
    - characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of *time related* data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Summary

- Data mining functionalities include the discovery of concept/class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis.

- Characterization and discrimination are forms of data summarization.

Thank You