SNS COLLEGE OF ENGINEERING
Coimbatore-107

Unit-III
DATA MINING - DM FUNCTIONALITIES

By
T.R.Lekhaa
AP-IT
SNSCE

Data Mining Functionalities

- Characterization and discrimination
- Mining frequent patterns, associations, correlations
- Classification and regression
- Clustering analysis
- Outlier analysis

DM Functionalities

It is used to specify the kind of patterns to be found in DM tasks.

DM tasks can be classified into two categories;

Descriptive mining tasks – characterize the *general properties of data* in DB

Predictive mining tasks – perform induction on the current data in order to make *predictions*

concept/class description

- Data can be associated with classes or concepts
- Ex: in the electronics store, classes of items for sale includes computers and printers and concept of customers include big spenders and budget spenders.
- Can be useful to describe individual classes and concepts in summarized.
- Such descriptions are called concept/class descriptions

concept/class description

These descriptions can be derived via.

- Data characterization
  - By summarizing the data of the class under study
- Data discrimination
  - By comparison of the target class with one or a set of comparative classes
- Or both
  - Both characterization & discrimination

Data characterization

- Data characterization
  - It is a summarization of general characteristics or features of a target class of data.
  - Several methods like OLAP roll up operation technique are used for data summarization.
  - Eg: to study the characteristics of software products with sales that increased by 10% in previous year
  - summarize the characteristics of customers who spend more than $5000 a year. The result -> profile of customer with 40to50years old, employed.
  - The output can be presented in various forms like;

    Pie charts,, bar charts, curves, multidimensional cubes, multidimensional tables etc

Data discrimination

- Comparison of general features of target class data objects with general features of objects from one or multiple contrasting classes.
- Eg: comparing general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during same period.
- Compare two group of customer -> those who shop computer products regularly , who rarely shop for products
- Result: 80% of customers -> buy products are between 20 and 40 years old, have university education
- 60% -> seniors or youths have no university degree
- The output of data discrimination can be presented in the same manner as data characterization

Mining frequent patterns, associations and correlations

Frequent patterns – patterns that occur frequently in data. The different types of patterns are;

- Frequent item sets – refers to a *set of items that frequently appear together* in a transaction data set, such as milk and bread.
- Frequent Subsequences or sequential pattern - pattern that customers tend to purchase : first a PC , followed by camera, then a memory card.
- Frequent Substructures – refers different structural forms (eg graphs, trees) that are combined with itemsets or subsequences.

Association analysis

- Conditions that *occur frequently together* in a given set of data.
- Association rule expression X=>Y, implies that the transaction of DB which contains X tends to contain Y.

  where X and Y are the set of items.
- This rule should satisfy 2 measures;

  Support and confidence

Classification and Regression for predictive analysis

- Classification

➢ Process of **finding a model or function** that describes and distinguishes data classes or concepts

➢ Involves finding rules that partition the data into disjoint groups

➢ The input for the classification is the training data set, whose class labels are already known

➢ Classification analyzes the training data set and constructs a model based on the class label.

Regression for Predictive Analysis

- Classification -> predicts categorical i.e discrete, unordered labels
- Regression -> predicts missing or unavailable numerical data values.

Cluster analysis

- Classification and regression -> analyze class-labeled data sets
- Clustering -> data objects without consulting class-labels
- Clustering is a method of grouping data into different groups, so that data in each group share similar trends and patterns.
- Principle to cluster:
  - Maximize intraclass similarity
  - Minimize interclass similarity
- The objectives of clustering are;
  - To uncover(find out) natural groupings
  - To initiate hypothesis about the data
  - To find consistent and valid organization of the data.

Outlier analysis

- Data objects which differ significantly from the remaining data objects are referred to as outliers. The analysis of outlier data is referred as **outlier analysis or anomaly mining**
- Most of the data mining methods discard outliers as noise or exceptions
- Some of the techniques for detecting outliers are statistical test, distance measures and deviation based method.

Thank You