Data Mining - Tasks

Introduction

Data Mining deals with what kind of patterns can be mined. On the basis of kind of data to be mined there are two kind of functions involved in Data Mining, that are listed below:

- Descriptive
- Classification and Prediction

Descriptive

The descriptive function deals with general properties of data in the database. Here is the list of descriptive functions:

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Class/Concept Description

Class/Concepts refers the data to be associated with classes or concepts. For example, in a company classes of items for sale include computer and printers, and concepts of customers include big spenders and budget spenders.Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by following two ways:

- **Data Characterization** - This refers to summarizing data of class under study. This class under study is called as Target Class.
- **Data Discrimination** - It refers to mapping or classification of a class with some predefined group or class.

Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns:

- **Frequent Item Set** - It refers to set of items that frequently appear together for example milk and bread.
- **Frequent Subsequence**- A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** - Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets or subsequences.

Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to process of uncovering the relationship among data and determining association rules.

For example A retailer generates association rule that show that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

It is kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item Sets to analyze that if they have positive, negative or no effect on each other.

Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on analysis of set of training data. The derived model can be presented in the following forms:

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

Here is the list of functions involved in this:

- **Classification** - It predicts the class of objects whose class label is unknown.Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on analysis set of training data i.e the data object whose class label is well known.
- **Prediction** - It is used to predict missing or unavailable numerical data values rather than class labels.Regression Analysis is generally used for prediction.Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** - The Outliers may be defined as the data objects that do not comply with general behaviour or model of the data available.
- **Evolution Analysis** - Evolution Analysis refers to description and model regularities or trends for objects whose behaviour changes over time.

Data Mining Task Primitives

- We can specify the data mining task in form of data mining query.
- This query is input to the system.
- The data mining query is defined in terms of data mining task primitives.

Note: Using these primitives allow us to communicate in interactive manner with the data mining system. Here is the list of Data Mining Task Primitives:

- Set of task relevant data to be mined
- Kind of knowledge to be mined
- Background knowledge to be used in discovery process

- Interestingness measures and thresholds for pattern evaluation
- Representation for visualizing the discovered patterns

SET OF TASK RELEVANT DATA TO BE MINED

This is the portion of database in which the user is interested. This portion includes the following:

- Database Attributes
- Data Warehouse dimensions of interest

KIND OF KNOWLEDGE TO BE MINED

It refers to the kind of functions to be performed. These functions are:

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

BACKGROUND KNOWLEDGE TO BE USED IN DISCOVERY PROCESS

The background knowledge allow data to be mined at multiple level of abstraction. For example the Concept hierarchies are one of the background knowledge that allow data to be mined at multiple level of abstraction.

INTERESTINGNESS MEASURES AND THRESHOLDS FOR PATTERN EVALUATION

This is used to evaluate the patterns that are discovers by the process of knowledge discovery. There are different interestingness measures for different kind of knowledge.

REPRESENTATION FOR VISUALIZING THE DISCOVERED PATTERNS

This refers to the form in which discovered patterns are to be displayed. These representations may include the following:

- Rules

- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

Source:

http://www.tutorialspoint.com/data_mining/dm_tasks.htm