

Data Mining

Data Mining- Tasks

Data mining

Data mining is a technique of extracting the hidden patterns of data from large data sets.

What are alternative names for data mining?

- Knowledge discovery in databases
- Data/pattern analysis
- knowledge extraction
- Data dredging
- Data archeology
- Business intelligence
- Information harvesting

Data Mining Applications

Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

Data mining deals with the **kind of patterns** that can be mined. On the basis of the kind of data to be mined, there are two **categories of functions** involved in Data Mining –

Descriptive

Classification and Prediction

Descriptive Function

The descriptive function deals with the 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/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales 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 the 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 the 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 a 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 item-sets or subsequences.

Mining of Association

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

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

It is a 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 the analysis of sets of training data. The derived model can be presented in the following forms

–

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

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.*

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*

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*



THANK YOU