Data mining tasks classified into two categories

→ Clustering, summarization, Association rules, sequence discovery.
Descriptive mining tasks

(1) Descriptive – Characterize the general properties of the data in the database.

(2) predictive – predictive mining tasks perform inference on the current data in order to make predictions.

Classification
Regression
Time series Analysis, prediction

Data mining functionalities

(I) Characterization and Discrimination.

Characterization → summarizing the data of the class under study (target class)

Discrimination → comparison of the target class with one or a set of comparative classes.

(II) Mining frequent patterns, Association and correlations.

Frequent patterns → patterns that occur frequently in data.

Mining frequent patterns help to discover the associations and correlation within data.

Association analysis

buys (X, "computer") ⇒ buys (X, "s/w") [support = 1% confidence = 50%]

confidence of 50% → if a customer buys a computer, there is a 50% chance that he/she will buy s/w as well.

1% support → 1% of all the transactions under analysis showed that pc and s/w were purchased together.

buy → is the attribute

age (X, "20 ... 29") ∧ income (X, "20K ... 29K") ⇒ buys (X, "CD player") [support 2%, confidence = 60%]

out of all customers under study 2% are 20 to 29 yrs. of age with an income of 20K – 29K and have purchased a CD player.

→ There is 60% probability that a customer in this age and income group will purchase a CDPlayer.

→ Association between more than one attributes
ex→ age, income, and buys.

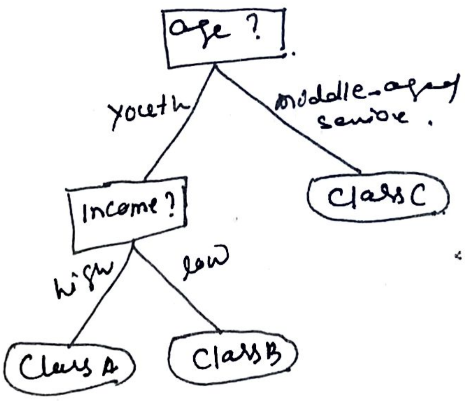
Classification and prediction.

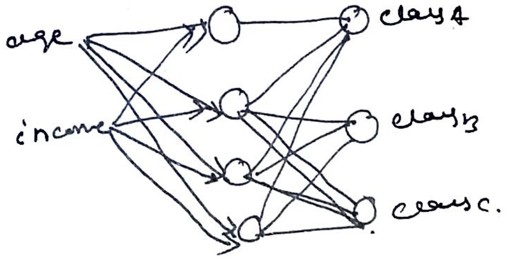
Classification is the process of finding a model that describes and distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

The derived model is based on the analysis of a set of training data.

Classification can be done by If-Then rules, decision trees or neural network

(a) age (X, "youth") AND income (X, "high") → class(X, "A")
age (X, "youth") AND income (X, "low") → class(X, "B")
age (X, "youth") AND income (X, "low") → class(X, "C")
age (X, "middle_aged) → class(X, "C")
age (X, "senior")





prediction → It is used to predict missing or unavailable numerical data values

statistical methodology regression analysis is most often used for numeric prediction.

Cluster analysis Here the objects are clustered or grouped on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

Outlier analysis

A database may contain data objects that do not comply with the general behavior or model of the data. The analysis of outlier data is known as outlier mining.

Outliers may be detected using.

→ statistical tests that using probability model.

→ using distance measures

→ deviation based methods - identify outliers by examining differences in the main characteristics of objects in a group.

Ex It may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account no. in comparison to regular charges incurred by the same account.

Evolution analysis.

Dat evolution analysis describes and models trends for objects whose behavior changes over time.

Ex stock market forecasting.

Classification on Data mining system.

(1) Based on the kind of databases mined.

Ex → relational, transactional or data warehouse mining.

(2) According to the kinds of knowledge mined.

→ they use the data mining functionalities.

(3) According to the kinds of techniques utilized.

Ex → machine learning, satistics, pattern recognition or NN.

(4) According to the applications adapted.

Ex → finance, telecommunication, DNA, stock market etc.

Data mining task primitives

A data mining task can be specified in the form of a data mining query, which is input to the datamining system.

A data mining query is defined in terms of data mining task primitives.

The data mining primitives specify the followings:-

The data mining task can be specified by 5 primitives:

(1) The set of task-relevant data to be mined.

→ this specifies the portions of the database or the set of data in which the user is interested. Ex:- relevant attributes or dimensions.

(2) The kind of knowledge to be mined.

→ This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification etc.

Data mining primitives define a data mining job, which can be specified in the form of a data mining query.

(3) The background knowledge to be used in the discovery process.

→ The knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns.

(4) The interestingness measures and thresholds for pattern evaluation.

① simplicity
② certainty (validity)
③ utility (usefulness)
④ Novelty.

They may be used to guide the mining process or after discovery, to evaluate the discovered patterns.

(5) The expected representation for visualizing the discovered patterns.

→ refers to the form in which discovered patterns are to be presented, for example in the form of rules, tables, charts etc.

Integration of Data mining system with a database or data warehouse system.

(1) NO coupling → DM system will not use any function of a DB/DW system.
It uses flat files as data sources.

(2) Loose coupling → DM system will use some facilities of a DB or DW system.

→ Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing and other system facilities.

→ It is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semitight coupling.

It means besides linking a DM system to a DB/DW system, efficient implementation of a few essential data mining primitives can be provided in the DB/DW system.

→ These primitives include sorting, indexing, aggregation, histogram analysis ~~each~~ etc.

→ Because these intermediate mining results are either precomputed or can be computed efficiently, this design will enhance the performance of a DM system.

Tight coupling

It means a DM system is integrated into the DB/DW system.

Major issues in Data mining.

(1) Mining different kinds of knowledge in database.

→ Because different users are interested in different kinds of knowledge, data mining should cover a wide range of data analysis and knowledge discovery tasks.

→ The data maining tasks may use the same database in different ways and require the development of numerous data mining techniques.