Data Mining Functionalities

Data mining tasks can be classified into two categories

- Descriptive mining – Characterize the general properties of the data in the database.
- Predictive mining – Perform inference on the current data in order to make prediction.

Concepts/Class Description: Characterization and Discrimination

- Data can be associated with classes or concepts
- Describe individual classes and concepts in summarized, concise, and precise terms
- Such descriptions of a class or concept are called class/concept description

Data characterization is a summarization of the general characteristics or features of target class of data. The data corresponding to the user-specified class are typically collected by a database query.

- There are several methods for effective data summarization and characterization. Simple data summaries based on statistical measures.
- An *attribute-oriented induction technique* can be used to perform data generalization and characterization without step-by-step user interaction.
- The output of data characterization can be presented in various formats. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional table, including crosstabs.
- ***For example***, the user may like to study the characteristics of software products whose sales increased by 10% in the last year.

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 are retrieved through database queries.
- The forms of output presentation are similar to those for characteristic descriptions.
- Discrimination descriptions expressed in rule form are referred to as **discriminant rules.**
- ***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, are patterns that occur frequently in data. There many kinds of frequent patterns, including **itemsets**, **subsequences**, and **substructures**.

A **frequent itemset** typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.

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, is a (frequent) sequential pattern.

A **substructure** can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets ro subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern.

Mining frequent patterns leads to discovery of interesting associations and correlations within data.

Association Analysis

- Multi-dimensional association:
  - age(X, “20..29”) $\wedge$ income(X, “20..29K”) $\Rightarrow$ buys(X, “PC”)

    [support = 2%, confidence = 60%]
- Single-dimensional association:
  - buys(T, “computer”) $\Rightarrow$ buys(T, “software”)

    [support = 1%, confidence = 75%]

Association rules are discarded as uninteresting if they do not satisfy both a min support threshold and min confidence threshold

Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, and 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).

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.

Decision trees can easily be converted to classification rules.

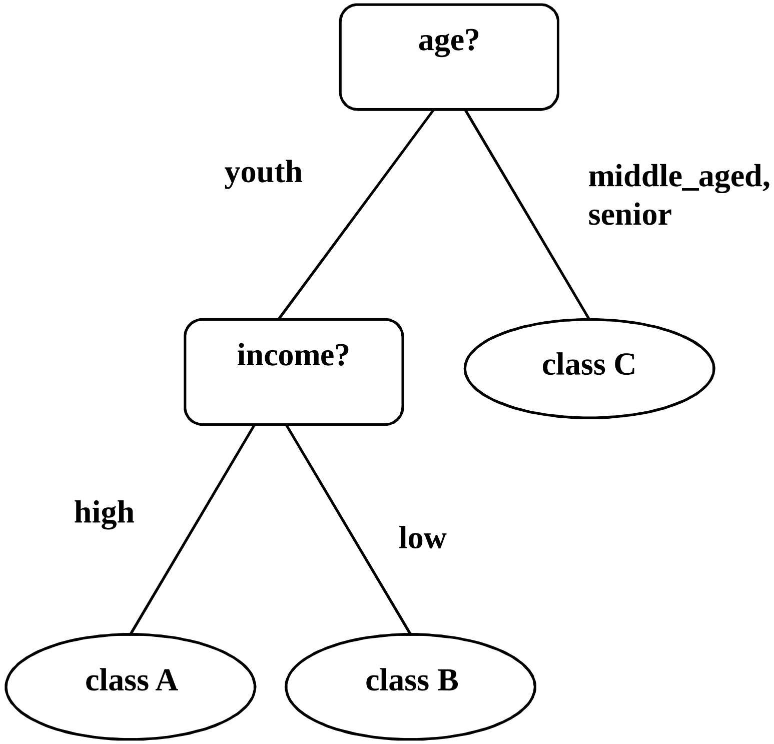
Prediction is used to predict missing or unavailable numerical data values rather than class labels. Regression analysis is a statistical methodology that is most often used for numeric prediction.

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”)



Cluster Analysis

Clustering analyzes data objects without consulting a known class label.

In general, the data labels are not present in the training data because they are not known to begin with. Clustering can be used to generate such labels.

The objects are clustered or grouped based 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. These data objects are **outliers**.

Most data mining methods discard outliers as noise or exceptions.

However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.

The analysis of outlier data is referred to as **outlier mining.**

Example : Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

Evolution Analysis

Data **evolution analysis** describes and models regularities or trends for objects whose behavior changes over time.

Example: A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies.

Interestingness of Patterns

A data mining system has the potential to generate thousands of patterns, or rules. But only a small fraction of the patterns potentially generated would actually be of interest to any giver user.

An interesting pattern represents **knowledge**.

- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective versus subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

- An objective measure for association rules of the form S ==> Y is rule **support**
- Another objective measure of association rules is **confidence**

support(X=> Y) = P(XUY)

confidence(X=> Y) = P(Y/X)

$$\text{support (X=> Y)} = \frac{\text{No. of tuples containing both X and Y}}{\text{total number of tuples}}$$

$$\text{confidence (X=> Y)} = \frac{\text{No. of tuples\_ containing both X and Y}}{\text{Number of tuples containing X}}$$

Classification of Data Mining Systems

Data mining is an interdisciplinary field, including database systems, statistics, machine learning, visualization, and information science

Data mining systems can be categorized according to various criteria

Classification according to the kinds of databases mined:

If classifying according to the special types of data handled, we may have time-series, text stream data, multimedia data mining systems, or World Wide Web mining system.

Classification according to the kinds of knowledge mined:

Based on data mining functionalities such as characterization, discrimination, association and correlation analysis, classification, clustering, prediction, outlier and evolution analysis.

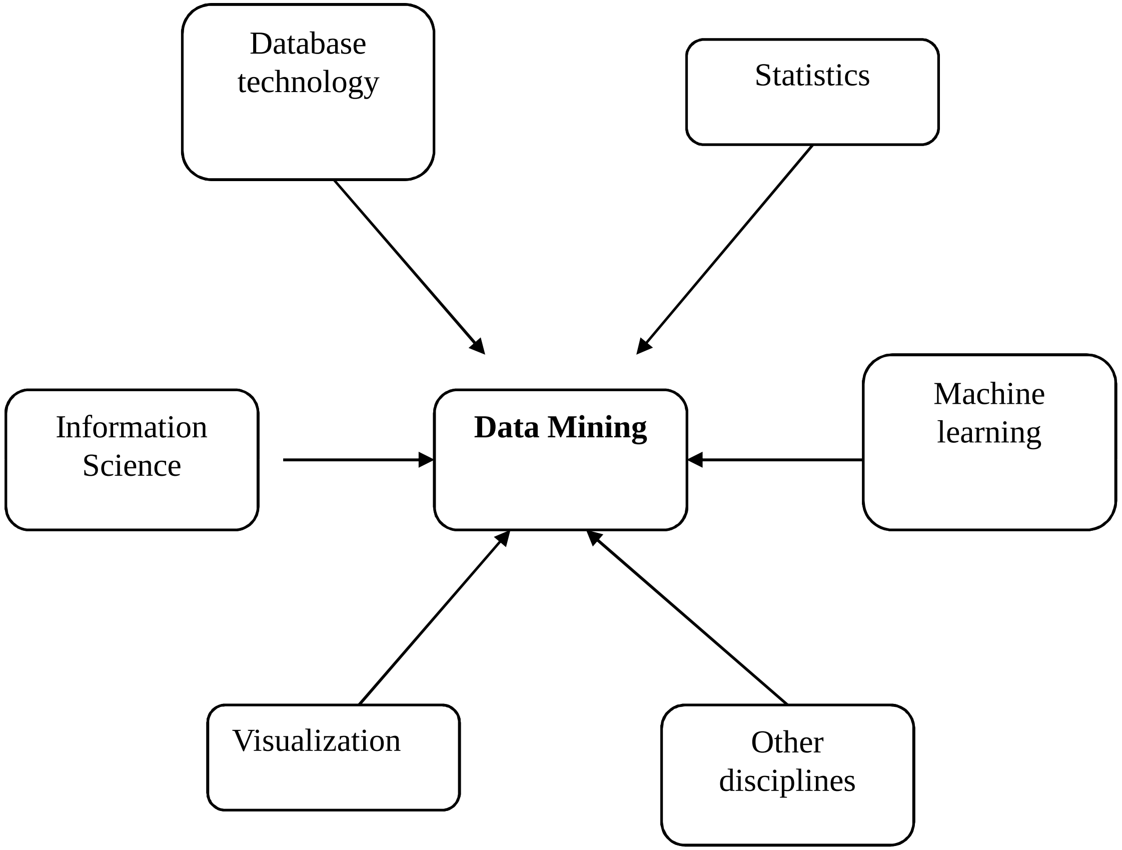
Based on levels of abstraction including generalized knowledge (high level of abstraction), primitive-level knowledge (raw data level), knowledge at multiple levels (several levels of abstraction)

Classification according to the kinds of techniques utilized:

Data mining systems can be categorized according to the underlying data mining techniques employed or degree of user interaction.

Classification according to the applications adapted:

For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on.



Database technology
Statistics
Information Science
Data Mining
Machine learning
Visualization
Other disciplines

Data Mining Task Primitives

A data mining query is defined in terms of **data mining task primitives**. These primitives allow the user interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

The data mining primitives specify the following.

***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. This includes the database attributes or data warehouse dimensions of interest.

***The kind of knowledge to be mined*:** This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

The ***background knowledge*** to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery

process and for evaluating the patterns found.

The ***interestingness measures and thresholds*** for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.

The expected ***representation for visualizing*** the discovered patterns: This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

Integration of a Data Mining System with a Database or Data Warehouse System

The possible integration schemes are as follows.

No coupling:

Data mining system will not utilize any function of a Database or Data warehouse system. It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then the mining results in another file.

Loose coupling:

Data mining system will use some facilities of a Database or Data warehouse system fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse.

Semitight coupling:

Besides linking a Data mining system to Database / Data warehouse system, efficient implementations of a few essential data mining primitives can be provided in the Database/Data warehouse system.

These primitives can include sorting, indexing, aggregation, histogram analysis, multi-way join, and pre-computation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on.

Tight coupling:

Data mining system is smoothly integrated into the Database/Data warehouse system. The data mining subsystem is treated as one functional component of an information system.

Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of DB/DW system.

Major Issues in Data Mining

The issues in data mining regarding mining methodology are given below.

***Mining methodology and user interaction issues*:** These reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization.

Mining different kinds of knowledge in databases: Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

Interactive mining of knowledge at multiple levels of abstraction:

The data mining process should be interactive.

Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. Specifically, knowledge should be drilling down, rolling up, and pivoting through the data space and knowledge space interactively.

Incorporation of background knowledge:

Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

***Data mining query languages and ad hoc data mining*:**

Data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns.

Presentation and visualization of data mining results:

Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.

Handling noisy or incomplete data:

The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to overfit the data.

***Pattern evaluation-the interestingness problem*:**

A data mining system can uncover thousands of patterns.