Data Mining: Tasks, Techniques, and Applications

Yongjian Fu

Department of Computer Science

University of Missouri - Rolla

Rolla, MO 65409 - 0350

Tel: (573)341-4040

Fax: (573)341-4491

Dr. Yongjian Fu is an assistant professor in the Department of Computer Science at the University of Missouri - Rolla. His research interests include data mining, data warehousing, information systems, and distributed databases systems. He can be reached by email at yongjian@umr.edu.

1 Introduction

A great amount of data are now available in science, business, industry, and many other areas, due to the rapid advances in computerization and digitalization techniques. Such data may provide a rich resource for knowledge discovery and decision support. For example, when you shop at a supermarket, the casher scans the bar-codes of items and stores your shopping transaction into a database. The supermarket can find valuable information for marketing by analyzing the sales data in its shopping transaction database.

In order to understand, analyze, and eventually make use of the huge amount of data, a multidisciplinary approach, data mining, is proposed to the meet the challenge. Data mining is the process of identifying interesting patterns from large databases.

Data mining is the core part of the Knowledge Discovery in Database (KDD) process as shown in Figure 1 [2]. The KDD process may consist of the following steps: data selection, data cleaning, data transformation, pattern searching (data mining), finding presentation, finding interpretation, and finding evaluation. Data mining and KDD are often used interchangeably because data mining is the key part of the KDD process.

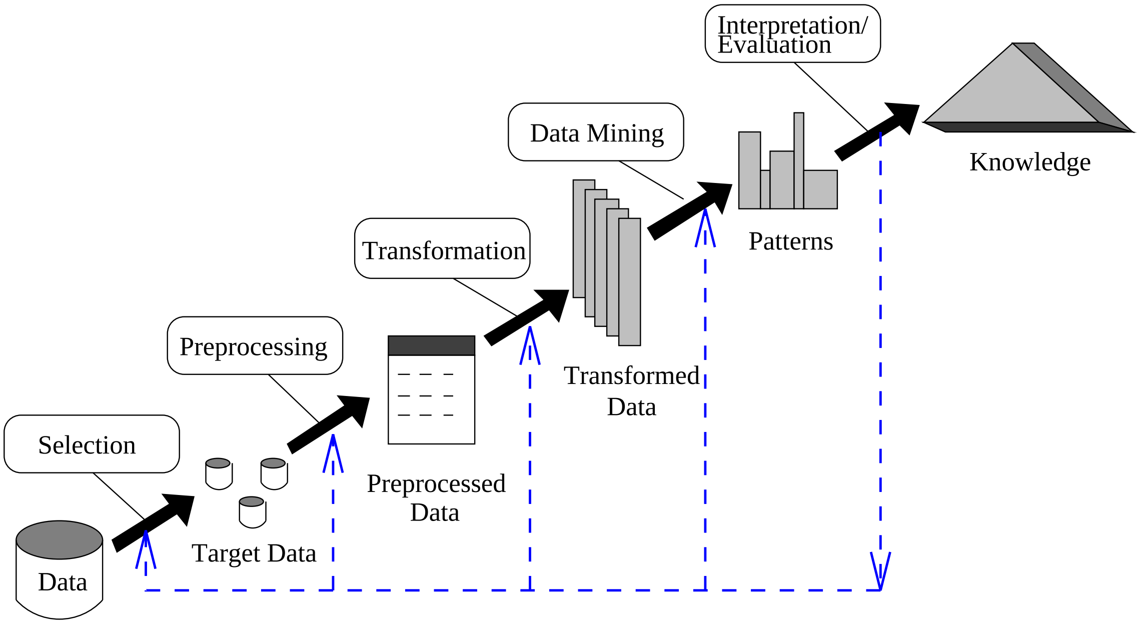


Figure 1: A typical knowledge discovery process [2].

2 Data Mining Tasks

The tasks of data mining are very diverse and distinct because there are many patterns in a large database. Different kinds of methods and techniques are needed to find different kinds of patterns. Based on the kinds of patterns we are looking for, tasks in data mining can be classified into summarization, classification, clustering, association, and trend analysis [2, 1].

- Summarization

  Summarization is the abstraction or generalization of data. A set of task-relevant data is summarized and abstracted, resulting a smaller set which gives a general overview of the data and usually with aggregation information. For example, the long distance calls of a costumer can be summarized into total_minutes, total_spending, total_calls, etc. Such high-level, summary information, instead of detailed calls, is presented to the sales managers for costumer analysis.

  The summarization can go up to different abstraction levels and can be viewed from different angles. For example, the calling minutes and spending can be totaled along the calling period in weeks, months, quarters, or years. Similarly, the calls can be summarized into in-state calls, state-to-state calls, Asia calls, Europe calls, etc., which can be further summarized into domestic calls and international calls. Different combinations of abstraction levels and dimensions reveal various kinds of patterns and regularities.

- Classification

  Classification is the derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database.

For example, from a set of diagnosed patients, who serve as the training set, a classification model can be built, which concludes a patient's disease from his/her diagnostic data. The classification model can be used to diagnose a new patient's disease based on the patient's diagnostic data, such as age, sex, weight, temperature, blood pressure, etc.

- Association

  Association is the discovery of togetherness or connection of objects. Such kind of togetherness or connection is termed as *association rule.* An association rule reveals the associative relationships among objects, i.e., the appearance of a set of objects in a database is strongly related to the appearance of another set of objects. For example, in a telecommunication database, an association rule that "call waiting" is associated with "call display", denoted as "call waiting → call display", says if a customer subscribes to the "call waiting" service, he or she very likely also has "call display".

  The association rules can be useful for marketing, commodity management, advertising, etc. For example, a retail store may discover that people tend to buy soft drinks together with potato chips, and then put the potato chips on sale to promote the sale of soft drinks.

- Clustering

  Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are unknown. The objected are so clustered that the intraclass similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. Once the clusters are decided, the objects are labeled with their corresponding clusters, and common features of the objects in a cluster are summarized to form the class description.

  For example, a bank may cluster its customers into several groups based on the similarities of their age, income, residence, etc., and the common characteristics of the customers in a group can be used to describe that group of customers. The clusters will help the bank to understand its costumers better and thus provide more suitable products and customized services.

- Trend analysis

  A lot of data available now are time series data that are accumulated over time. For example, a company's sales, a costumer's credit card transactions, and stock prices, are all time series data. Such kind of data can be viewed as objects with an attribute *time*, and the objects are the snapshots of entities with values that changes over time. It is interesting to find the patterns and regularities in the data evolutions along the dimension of time.

  Trend analysis discovers interesting patterns in the evolution history of the objects. One topic in trend analysis is the identification of patterns in an object's evolution, such as up, down, peak, valley, etc. A model or function is constructed to simulate the behaviors of the object, which can be used to predict the future behaviors. For example, we can estimate this year's profit of a company from its last year's profit and the estimated annual increasing rate.

  Another topic in trend analysis is the matching of the objects' changing trends, such as increasing streaks, decreasing streaks, etc. By comparing two or more objects' historical changing curves or tracks, similar and dissimilar trends can be discovered which will help us to understand the behaviors of the objects. For example, a company's sales and profit figures can be analyzed to find the disagreeing trends and search for the reasons behind such disagreements.

3 Data Mining Techniques

As a multi-disciplinary field, data mining adopted its techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization.

- Statistical approaches.

  Many statistical tools have been used for data mining, including Bayesian network, regression analysis, correlation analysis, and cluster analysis. Usually statistical models are built from a set of training data. An optimal model, based on a predefined statistical measure, is searched among the hypothesis space. Rules, patterns, and regularities are then drawn from the model.

A Bayesian network is a directed graph which represents the casual relationships among the variables, computed using the Bayesian probability theorem. Regression is the derivation of a function which maps a set of attributes of objects to an output variable. Correlation analysis studies the correspondence of variables to each other, such as the χ^2. Cluster analysis finds groups from a set of objects based on distance measures.

A simple Bayesian network for a medical problem is given in Figure 2. Nodes in a Bayesian network represent variables or states, while edges represent the dependencies between nodes, directed from the cause to the effect. From the figure, we can see that a patient's age, occupation, and diet affect the disease which in turn causes symptoms.

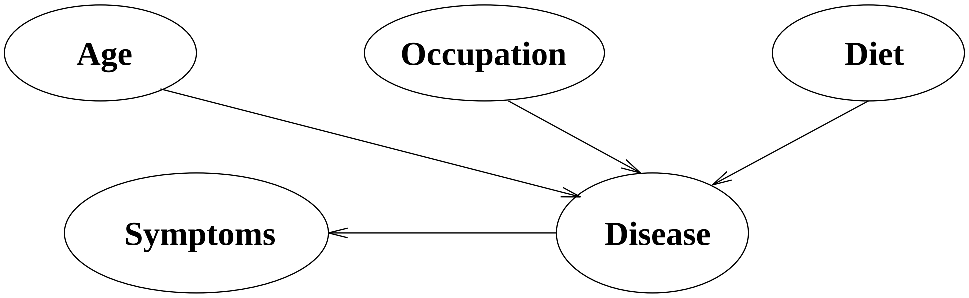


Figure 2: A simple Bayesian network.

- Machine learning approaches.

  Like statistical methods, machine learning methods search for a best model that matches the testing data. Unlike statistical methods, the searching space in machine learning methods is a cognitive space of n attributes instead of a vector space of n dimensions. Besides, most machine learning methods use heuristics in the searching.

  The most common machine learning methods used for data mining include decision tree induction, inductive concept learning, and conceptual clustering. A decision tree is a classification tree which determines an object's class by following the path from the root to a leaf node, choosing the branches according to the attribute values of the object. Decision trees are induced from the training set and classification rules can be extracted from the decision trees. Inductive concept learning derives a concise, logical description of a concept from a set of examples. Conceptual clustering finds groups or clusters in a set of objects, based on conceptual closeness among objects.

A simple decision tree is given in Figure 3. It determines a car's mileage from its size, transmission type, and weight. The leaf nodes are in square boxes which represents the three classes of mileages. From the decision tree, we can conclude, for example, a medium size, automatic car will have medium mileage.

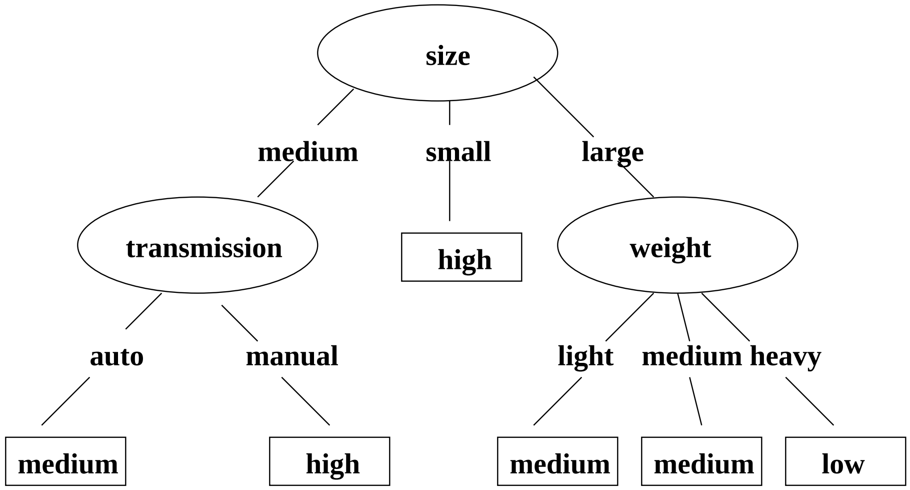


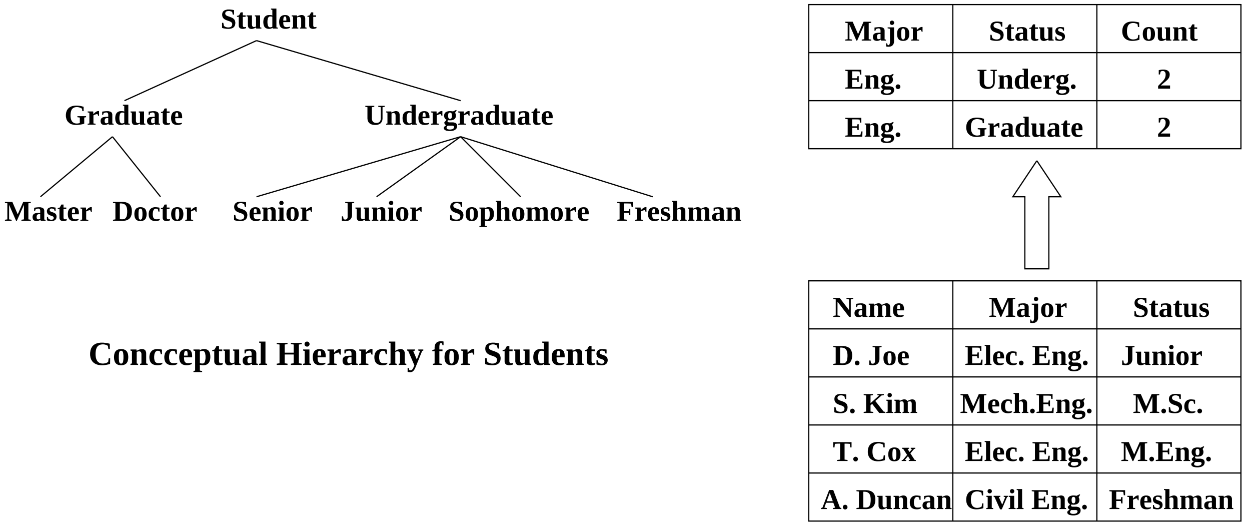
Figure 3: A simple decision tree.

- Database-oriented approaches.

Database-oriented methods do not search for a best model as the previous two kinds of methods. Instead, data model or database specific heuristics are used to exploit the characteristics of the data in hand. The attribute-oriented induction, the iterative database scanning for frequent item sets, and the attribute focusing, are representatives of the database-oriented methods.

In attribute-oriented induction, primitive, low-level data are generalized into high-level concepts using conceptual hierarchies. The iterative database scanning method is employed to search for frequent item sets in a transactional database. The association rules are then derived from these frequent item sets. The attribute focusing method looks for patterns with unusual probabilities by adding attributes selectively into patterns.

The left side of Figure 4 shows a simple conceptual hierarchy for students, and the right side shows a example of attribute-oriented induction. In the example, the students of a local IEEE chapter are summarized.



Major	Status	Count
Eng.	Underg.	2
Eng.	Graduate	2

Name	Major	Status
D. Joe	Elec. Eng.	Junior
S. Kim	Mech.Eng.	M.Sc.
T. Cox	Elec. Eng.	M.Eng.
A. Duncan	Civil Eng.	Freshman

Figure 4: Conceptual hierarchy and attribute-oriented induction.

- Other approaches.

Many other techniques have been adopted for data mining, including neural networks, rough sets, and visualization.

A neural network is a set of interlinked nodes, called neurons. A neuron is a simple computing device that computes a function of its inputs which can be outputs of other neurons or attribute values of an object. By adjusting the connection and the functional parameters of the neurons, a neural network can be trained to model the relationship between a set of input attributes and an output attribute. A neural network can be used, for example, in classification when the output attribute is the object's class. A rough set is a set whose membership is fuzzy. A set of objects can be arranged to form a group of rough sets which may be used, for example, in classification and clustering. Visual exploration is another interesting data mining technique. Data are transformed into visual objects such as dots, lines, areas etc, and displayed in a two or three dimensional space. Users can interactively explore the interesting spots by visual examination.

The above methods can be integrated or combined to deal with complicated problems or provide alternative solutions. For example, summarization of data is usually visually presented as charts, graphs, etc., to help the understanding of the results and further examination. Indeed, most data mining systems employ multiple methods to deal with different kinds of data, different data mining tasks, and different application areas.

4 Data Mining Applications

Data mining techniques have been applied successfully in many areas, from traditional areas such as business and science, to new areas such as sports.

- Business applications.

  A lot of organizations now employ data mining as a secret weapon to keep or gain competitive edge. Data mining has been used in database marketing, retail data analysis, stock selection, credit approval, etc.

  - Database marketing is one of the most successful and popular business applications of data mining. By mining historical customer databases, patterns and trends are extracted and customer profiles are built which can be used for more effective marketing.
  - Retail databases contain customer shopping transactions. Data mining can find customer shopping patterns which can be used, for example, in sales campaign.
  - Using data mining techniques, investors can build models which can be used to predict the performance of stocks. By searching trends and patterns in stocks data, data mining can also help investors to find stocks with good performance.
  - Applications for credit or loan are decided based on the applicants' information. A decision support model for credit or loan approval may be constructed from historical data using data mining tools.

- Science applications.

  Data mining techniques have been used in astronomy, molecular biology, medicine, geology, and many more. For example, Jet Propulsion Lab at the California Institute of Technology has developed a data mining system which can classify the sky objects, such as stars, in the satellite images.

- Other applications.

  Data mining techniques have also been used in many other areas, such as health care management, tax fraud detection, money laundering monitoring, even sports. For example, the Advanced Scout system developed by IBM has been used by coaches of more than a dozen teams in the National Basketball Association (NBA) to improve their games.

To summarize, data mining is the process of extracting interesting patterns from large databases. Data mining can be the solution to the data analysis problems faced by many organizations. As a young and emerging field, a lot more work is needed although a great deal of progress has been made in data mining research and development.

5 Learn more about it

- Journals and books.

  - Fayyad, et al, (Eds) Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
  - Journal of Data Mining and Knowledge Discovery, Kluwer Academic Publishers.
  - IEEE Transaction on Knowledge and Data Engineering, Special Issue on Data Mining, Vol. 8, No. 6, 1996.
  - IEEE Expert, Intelligent Systems and Their Applications, Special Issue on Data Mining, Vol. 1, No. 5, 1996.
  - IEEE Transaction on Knowledge and Data Engineering, Special Issue on Data Mining, Vol. 6, No. 5, 1993.

  - Piatetsky-Shapiro, G. and Frawley, W., (Eds) Knowledge Discovery in Databases, MIT Press, 1991.

- Conferences.
  - Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, 1997.
  - Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.
  - First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 1995.
  - 1997 ACM-SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Tucson, Arizona, 1997.
  - 1996 ACM-SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 1996.

- World Wide Web.
  - Knowledge Discovery Mine

    http://www.kdnuggets.com/
  - Data Mine

    http://www.cs.bham.ac.uk/ anp/TheDataMine.html

References

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8:866–883, 1996.

[2] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–35. AAAI/MIT Press, 1996.