7

Data Mining Tasks, Techniques, and Applications

Objectives:

- To compete effectively in today's marketplace, business managers must take timely advantage of high return opportunities.
- This chapter explores two applications of current data mining techniques: market basket analysis and customer segmentation.

Data mining takes two forms. *Verification-driven data mining* extracts information in the process of validating a hypothesis postulated by a user.

Discovery driven data mining uses tools such as symbolic and neural clustering, association discovery, and supervised induction to automatically extract information.

- Once a data warehouse has been developed, the data mining process falls into four basic steps: data selection, data transformation, data mining, and result interpretation.
- Seven operations are associated with data mining: three with verification-driven data mining and four with discovery-driven data mining.
- **Verification-driven data mining operations.** These include query and reporting, multidimensional analysis, and statistical analysis.
- **Discovery-driven data mining operations.** These include predictive modeling, link analysis, database segmentation, and deviation detection.
- Data mining adopted its techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization.

Abstract. In this chapter a reality check for data mining is proposed. *Data mining* – the process of extracting the valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions – currently performs this task for a growing range of businesses. After presenting an overview of current data mining techniques, this chapter explores two particularly noteworthy applications of those techniques: *market basket analysis* and *customer segmentation.*

S. Sumathi and S.N. Sivanandam: *Data Mining Tasks, Techniques, and Applications*, Studies in Computational Intelligence (SCI) **29**, 195–216 (2006)
www.springerlink.com

Mountainous amounts of data records are now available in science, business, industry, and many other areas. Such data can provide a rich resource of knowledge discovery and decision support. For example, when we shop at a supermarket, the cashier scans the bar codes of items and stores your shopping transaction into a database. The supermarket can find valuable information for product selection by analyzing the sales data in its transaction database. (The store can also sell this valuable information.). To understand, analyze, and eventually use this data, a multidisciplinary approach called data mining has been proposed. Data mining is the process of identifying interesting patterns from large databases.

7.1 Reality Check for Data Mining

To compete effectively in today's marketplace, business managers must take timely advantage of high return opportunities. Doing so requires that they be able to exploit the mountains of data their organizations generate and collect during daily operations. Yet the difficulty of discerning the value in that information – of separating the wheat from the chaff – prevents many companies from fully capitalizing on the wealth of data at their disposal.

For example, a bank account manager might want to identify a group of married, two-income, affluent customers and send them information about the bank's growth mutual funds, before a competing discount broker can lure them away. The information surely resides in the bank's computer system and has probably been there in some form for years. The trick, of course, is to find an efficient way to extract and apply it.

7.1.1 Data Mining Basics

Data mining takes two forms. *Verification-driven data mining* extracts information in the process of validating a hypothesis postulated by a user. It involves techniques such as statistical and multidimensional analysis. *Discovery-driven data mining* uses tools such as symbolic and neural clustering, association discovery, and supervised induction to automatically extract information. The extracted information from both approaches takes one of several forms:

Facts – for example, sales increased by 12%,
Trends – sales of the northeastern stores are decreasing,
Regression or classification models,
Relations between database records – such as when consumers buy lettuce and tomatoes, they also buy bacon 80% of the time, and
Deviations from norms, among others.

To be effective, a data mining application must do three things. First, it must have access to organization-wide views of data instead of department-specific ones. Frequently, the organization's data is supplemented with open-source or purchased data. The resulting database is called the *data warehouse.*

During data integration, the application often cleans the data – by removing duplicates, deriving missing value (when possible) – and establishing new derived attributes, for example. Second, the data mining application must mine the information in the warehouse. Finally, it must organize and present the mined information in a way that enables decision making.

Systems that can satisfy one or more of these requirements range from commercial decision-support systems such as Lotus 1-2-3 and Pilot Software's Lightship, to customized decision-support systems and executive information systems such as SAS/EIS. As one example, Lotus 1-2-3 can access data from a data warehouse, allow its users to perform variety of statistical operations, and present the results using simple business graphics.

The overall objective of each decision-making operation determines the type of information to be mined and the ways for organizing the mined information. For example, by establishing the objective of identifying good prospective customers for mutual funds, the bank account manager mentioned earlier implicitly indicates that she wants to segment the database of bank customers into groups of related customers – such as urban, married, two income, mid-thirties, low-risk, high-net-worth individuals – and establishes the vulnerability of each group to various types of promotional campaigns.

7.1.2 The Data Mining Process

Once a data warehouse has been developed, the data mining process falls into four basic steps: data selection, data transformation, data mining, and result interpretation (see Fig. 7.1).

Data selection. A data warehouse contains a variety of data, not at all of which is needed to achieve each data mining goal. The first step in the data

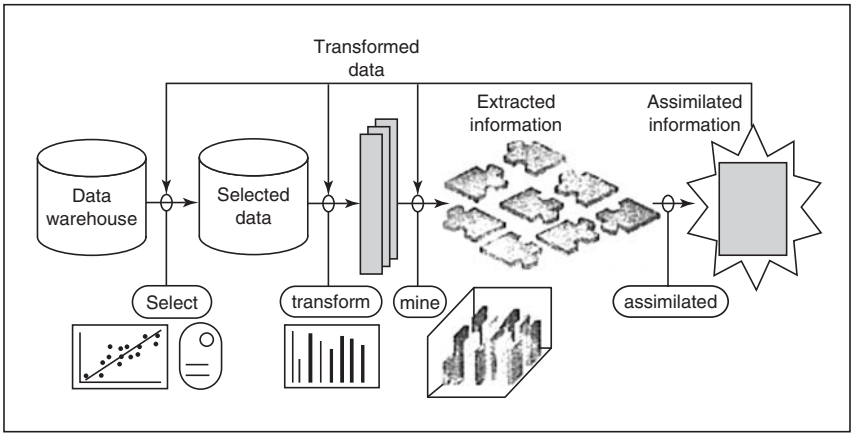

Fig. 7.1. The Data Mining Process

mining process is to select the target data. For example, marketing databases contain data describing customer purchases, demographics, and lifestyle preferences. To identify which items and quantities to purchase for a particular store, as well as how to organize the items on the store's shelves, a marketing executive might need only to combine customer purchase data with demographic data. The selected data types may be organized along multiple tables: during data selection, the user might need to perform table joints. Furthermore, even after selecting the desired database tables, mining the contents of the entire table is not always necessary for identifying useful information. Under certain conditions and for certain types of data mining operations (such as when creating a classification or regression. model), it is usually a less expensive operation to sample the appropriate table, which might have been created by joining other tables, and then mine only the sample.

Data transformation. After selecting the desired database tables and identifying the data to be mined, the user typically needs to perform certain transformations on the data. Three considerations dictate which transformation to use: the task (mailing-list creation, for example), the data mining operations (such as predictive modeling), and the data mining technique (such as neural networks) involved. Transformation methods include organizing data in desired ways (organization of individual consumer data by household), and converting one type of data to another (changing nominal values into numeric ones so that they can be processed by a neural network). Another transformation type, the definition of new attributes (derived attributes), involves applying mathematical or logical operators on the values of one or more data base attributes – for example, by defining the ratio of two attributes.

Data mining. The user subsequently mines the transformed data using one or more techniques to extract the desired type of information. For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulations manager might need to first use clustering to segment the subscriber database, and then apply rule induction to automatically create a classification model for each desired cluster.

Result interpretation. The user must finally analyze the mined information according to his decision-support task and goals. Such analysis identifies the best of the information. For example, if a classification model has been developed, during result interpretation, the data mining application will test the model's robustness, using established error-estimation methods such as cross validation. During this step, the user must also determine how best to present the selected mining-operation results to the decision maker, who will apply them in taking specific actions. (In certain domains, the user of the data-mining application – usually a business analyst – is not the decision- maker. The latter may make business decisions by capitalizing on the data mining results through a simple query and reporting tool.) For example, the user might

decide that the best way to present the classification model is logically in the form of if-then rules.

Three observations emerge from this four-step process:

Mining is only one step in the overall process. The quality of the mined information is a function of both the effectiveness of the data-mining technique used and the quality, and often sizes, of the data being mined. If users select the wrong data, choose inappropriate attributes, or transform the selected data inappropriately, the results will likely suffer.

The process is not linear but involves a variety of feedback loops. After selecting a particular data-mining technique, a user might determine that the selected data must be preprocessed in particular ways or that the applied technique, did not produce the results of the expected quality. The user then must repeat earlier steps, which might mean restarting the entire process from the beginning.

Visualization plays an important role in the various steps. In particular, during the selection and transformation steps, a user could use statistical visualizations – such as scatter plots and histograms – to display the results of exploratory data analysis. Such exploratory analyses often provide preliminary understanding of the data, which helps the user select certain data, subsets. During the mining step, the user employs domain-specific visualizations. Finally, visualizations – either special landscapes or business graphics – can present the results of a mining operation.

7.1.3 Data Mining Operations

Seven operations are associated with data mining: three with verification-driven data mining and four with discovery-driven data mining.

Verification-driven data-mining operations: These include query and reporting, multidimensional analysis, and statistical analysis.

Query and reporting. This operation constitutes the most basic form of decision support and data mining. Its goal is to validate a hypothesis expressed by the user, such as "sales of four-wheel-drive vehicles increase during the winter." Validating a hypothesis through a query-and-reporting operation entails creating a query, or set of queries, that best expresses the stated hypothesis, posing the query to the database, and analyzing the returned data to establish whether it supports or refutes the hypothesis. Each data interpretation or analysis step might lead to additional queries, either new ones or refinements of the initial one. Reports subsequently compiled for distribution throughout an organization contain selected analysis results, presented in graphical, tabular, and textual form, and include a subset of the queries. Because these reports include the queries, analyses can be automatically repeated at predefined times, such as once a month.

Multidimensional analysis. While traditional query and reporting suffices for several types of verification-driven data mining, effective data mining in certain domains requires the creation of very complex queries. These often contain an embedded temporal dimension and may also express change between two stated events. For example, the regional manager of the department store chain might say "Show me weekly sales during the first quarter of 1994 and 1995, for Midwestern stores, broken down by department." Multidimensional databases, often implemented as multidimensional arrays, organize data along predefined dimensions (time or department, for example), have facilities for taking advantage of sparsely populated portions of the multidimensional structure, and provide specialized languages that facilitate querying long dimensions while expediting query-processing performance. These databases also allow hierarchical organizations of the data along each dimension, with summaries on the higher levels of the hierarchy and the actual data at the lower levels. Quarterly sales might take one level of summarization and monthly sales a second level, with the actual daily sales taking the lowest level of the hierarchy.

Statistical analysis. Simple statistical analysis operations usually execute during both query and reporting, as well as during multidimensional analysis. Verifying more complex hypothesis, however, requires statistical operations (such as principal component analysis regression modeling) coupled with data visualization tools. Several statistical analysis tools (SAS, SPSS, S+) incorporate components that can be used for discovery-driven modeling (such as CHAID in SPSS and S+). To be effective statistical analysis must rest on a methodology, such as exploratory data analysis. A methodology might need to be business or domain dependent, so statistics tools such as SAS and SPS are open ended, providing function libraries that can be organized into larger analysis software systems.

Discovery-driven data mining operations: These include predictive modeling, link analysis, database segmentation, and deviation detection.

Predictive modeling. This is the most commonly used of the discovery-driven data mining operations, primarily because of the proliferation of automatic model development techniques. (These include symbolic-induction techniques such as CART or C4.5, and neural network techniques such as back propagation.) This operation uses the database records that reflect historical data information about past behavior – to automatically generate a model that can predict the future behavior. For example, an insurance underwriter might want to predict the likelihood that a customer will let her policy lapse. A marketing executive might want to predict whether a particular customer would switch brands for a specific product. The value of discovery-driven modeling techniques, especially symbolic induction techniques, is that the models are expressed as sets of if-then rules, and are therefore comprehensible and explainable.

Database segmentation. It is often necessary to automatically partition a database into collection of related records. Such collections enable users to summarize each database or select a smaller portion of the database on which to apply a different data mining operation such as model creation or link analysis. For example, by segmenting a department stores point off scale data, a user can automatically create segments containing transactions from specific periods, such as back to school days or after Christmas sales.

Link analysis. While the modeling and segmentation operations aim to create the generalized description that characterizes database's contents, link analysis seeks to establish relations between the records in the database. For example, a merchandise buyer must determine what items sell together men's shirts sell together with ties and men's fragrances – so that he can decide what items to buy for a store (shirts, ties, and fragrances), as well as where to place these items in relation to each other (ties and fragrances must be displayed close to the store's men's shirts section). Certain algorithms that fall in this category can also trace connections between records overtime.

Deviation detection. This operation attempts to identify points that cannot be fitted into the segment and then explain whether each such point is noise or should be examined in more detail. This operation usually operates in conjunction with database segmentation and, because "outliers" express deviation from expected norms, often leads to true discovery.

7.1.4 Discovery-Driven Data Mining Techniques:

While there are only four basic discovery-driven mining operations, numerous data mining techniques support these operations. For example, supervised induction techniques support predictive model creation; association discovery and sequence discovery techniques support link analysis; and statistical techniques support deviation detection.

Supervised induction. This process involves automatically creating a classification model from a set of records (examples) called *the training set*. The training set might either be a sample of the database or a warehouse being mined, the entire database, or a data warehouse. The records in the training set must belong to a small set of classes that the analyst has predefined. The induced model consists of patterns – essentially generalizations over the records—that are useful for distinguishing the classes. Once induced, a model can help automatically predict the class of other unclassified records. Supervised induction methods can be either neural or symbolic. Neural methods such as back propagation represent the model as architecture of nodes and weighted links. Symbolic methods create models that are represented either as decision trees or as if-then rules. A supervised induction technique is particularly suitable for data mining if it has three characteristics.

It can produce high-quality models even when the data in the training set is noisy and incomplete.

The resulting models are comprehensible and explainable, so that the user can understand how the system makes the decision.

It can accept domain knowledge, which can expedite the induction task while simultaneously improving the quality of the induced model.

Association discovery. Given a collection of items and a set of records, each of which contains some number of items from the given collection, an association discovery function is an operation against this set of records that returns affinities existing among the collection of items. These affinities can be expressed by rules such as, "72% of all the records that contain items A, B and C also contains items D and E." The specific percentage of occurrences (in this case 72) is *the confidence factor* of the association. Also in this association, A, B, and C are said to be on opposite side of the association to D and E. Association discovery can involve any number of items on either side of the association.

Sequence discovery. In the transaction log discussed earlier, the identification of the customer who made the purchase generally remains unknown. If this information exists, analysis can be made on the collection of records of the same structure as above (that is consisting of the number of items drawn for a given collection of items). The records are related by the identity of the customer who made the repeated purchases.

Clustering. This is used to segment the database into subsets, or clusters with the members of each cluster sharing a number of interesting properties. The results of a clustering operation fulfill one of two functions.

Summarizing the contents of the target database by considering the characteristics of each created cluster, rather than those of each record in database.

Serving is an input to other methods, such as supervised induction. A cluster is smaller and more manageable data set to the supervised inductive learning component.

Clusters use neural and symbolic unsupervised induction methods. The various neural and symbolic methods are distinguished by the type of attributes values they allow the records in the target database to take (numeric, nominal, structured objects), the way they represent each cluster, and the way they organize the set of clusters. After clustering the database, the analyst can use visualization component to examine the created clusters to locate ones that are useful or interesting.

Neural clustering methods, such as feature maps, represent the cluster as a prototype with which they associate a subset of the instances in the data set being clustered. Symbolic clustering methods operate primarily on

records with nominal-valued attributes. They consider all the attributes that characterize each instances and use AI-based search methods to establish the subset of these attributes that will desire each created cluster.

Characteristics of data mining: Let us begin our discussion about data mining applications by examining the available, albeit preliminary, statistics. 80% to 90% of the operational data mining applications today employ only a single operation. There are two reasons for this. First, because tool developers usually understand only one or two determining operations, the tools on which these applications are based usually contain only one technique. Second when initially applying data mining end-user organizations select narrowly focused problems that require only one operation to solve.

In 70% of the applications, users perform data mining using verification-driven operations. Data analysts and business analysts alike thoroughly understand query and reporting, multidimensional analysis, and statistical analysis. Because many of them are graduates of business schools, they usually receive training such operations, so they feel comfortable trying to solve their decision support problems with them. Neural and symbolic induction methods have only recently appeared in business school curricula.

To date, two factors have inhibited the broad deployment of applications that incorporate discovery-driven data mining techniques: the significant effort necessary to develop each data mining application and the inappropriate state of the data that application must mine.

Application development. Most deployed data mining applications are not developed by business analysts but through the collaboration of data mining tool vendors, data analysts, and end users. Because the tool vendors and data analysts usually first must develop an understanding of the end user's problem, such collaborations are time consuming. Furthermore, the current generations of data mining tools are aimed at the data analyst, not the business analyst.

Data. Data is often incorrect and incomplete, distributed across many departmental databases, organized using incompatible data models and formats, and using often-incomprehensible naming conventions. Because of such problems, and the lack of appropriate data integration and cleaning tools, the Gartner group has estimated that only about 10% of the collected data is ever analyzed. Many organizations realize the poor quality of the data they have been collecting only after they start developing a data mining application. The development of data warehouses should alleviate this problem. A data warehouse ensures that disparate data is integrated consistently under a single data model and is cleaned in the process, and that metadata is encoded explicitly. Over 40% of the organizations surveyed by the metagroup across 10 industries have started developing data warehouses during the past six months, with the intention of using them in conjunction with multidimensional databases and discovery-driven data mining tools. These warehouses will vary from departmental data marts, size ranging from 20 to 60 Gbytes, to corporate warehouses, which often hold several terabytes of data.

7.2 Data Mining: Tasks, Techniques, and Applications

Data mining is the core part of the knowledge discovery in database (KDD) process. The KDD process may consist of the following steps: data selection, data cleaning, data transformation, pattern searching (data mining), and finding presentation, finding interpretation, and finding evaluation. Data mining and KDD are often used interchangeably because data mining is the key to the KDD process. These steps are illustrated in Fig. 7.2.

7.2.1 Data Mining Tasks

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

Summarization. Summarization is the abstraction or generalization of data. A set of task-relevant data is summarized and abstracted. This results in a smaller set, which gives a general overview of the data, usually with aggregate information.

For example, the long-distance calls of a customer can be summarized into total minutes, total spending, total calls, and so forth. Such high-level, summary information, instead of the individual elements of each call, is presented to the sales managers for customer analysis.

The summarization can go to different abstraction levels and can be viewed from different angles. For example, calling minutes and spending can

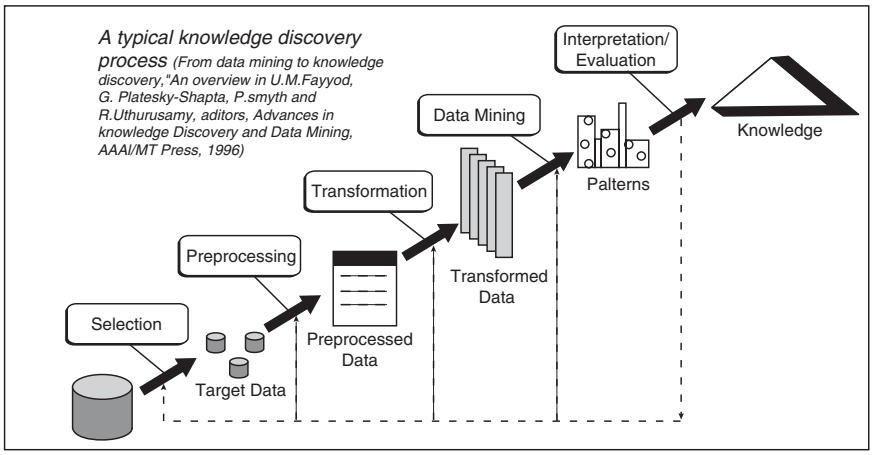

Fig. 7.2. A Typical Knowledge Discovery Process

be totaled along various calling periods: weeks, months, quarters, or years. Similarly, the calls can be summarized into in-state calls, state-to-state calls, calls to Asia, calls to Europe, etc. These groupings 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 derives 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 it, 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. This function or model can then classify future objects. This helps us 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. This model concludes a patient's disease from his/her diagnostic data. The classification model can diagnose a new patient's disease based on data such as age, sex, weight, temperature, blood pressure, etc.

Association. Association is the discovery of togetherness or the connection objects. Such togetherness or connection is termed an *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, we may find a rule that "call waiting" is associated with "call display," denoted as "call waiting → call display." It states that if a customer subscribes to the "call waiting" service, he or she very likely also has "call display."

Association rules can be useful for marketing, commodity management, advertising, and so forth. For example, a retail store may discover that people tend to buy soft drinks and potato chips together. Store personnel then place the potato chips near the soft drinks to promote the sale of both. They may even discount one to entice buying the other, since these shoppers now will be "saving money."

Clustering. Clustering identifies classes – also called *clusters* or *groups* – for a set of objects whose classes are unknown. The objects are so clustered that the interclass similarities are maximized and the interclass similarities are minimized. This is done based on some criteria defined on the attributes of the objects. Once the clusters are decided the objects are labeled with their corresponding clusters. The common features for object 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, and residence. The characteristics, the customers in a group share can be used to describe that group of customers.

These clusters help the bank understand its customers better and provide more suitable products and customized services.

Trend analysis. Time series data are records accumulated over time. For example, a company's sales, a customer's credit card transactions, and stock prices are all time series data. Such data can be viewed as objects with an attribute time. The objects are snapshots of entries with values that change over time. Finding the patterns and regularities in the data evolution along the dimension of time can be fascinating.

One topic in trend analysis is identifying patterns in an object's evolution, such as its ups and downs, or peaks and valleys. A model or a function is constructed to simulate the behavior of the object to predict future behavior. For example, we can estimate this year's profit for a company based on its profits last year and the estimated annual increasing rate.

Another topic in trend analysis is matching 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. These can help us to understand the behavior of the objects. For example, a company's sales and profit figures can be analyzed to find the disagreeing trends. These trends can be further researched to discover the reasons behind such disagreements.

7.2.2 Data Mining Techniques

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 and cluster analysis, and correlation analysis. Usually statistical models are built from a set of training data. An optimal model, based on a defined statistical measure, is searched among the hypothesis space. Rules, patterns, and regularities are then drawn from the models.

A Bayesian network is directed graph computed using the Bayesian probability theorem. It represents the causal relation ships among the variables. Regression is the function derivation, 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 x^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 Fig. 7.3. Nodes in a Bayesian network represent variables or states; edges represent the dependencies between nodes directed from the cause to the effect. The figure shows that a patient's age, occupation, and diet affect the disease, which causes the symptoms.

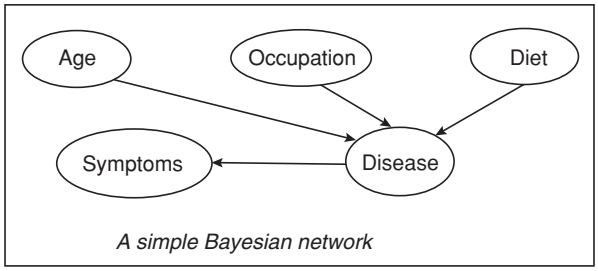

Fig. 7.3. A Simple Bayesian network

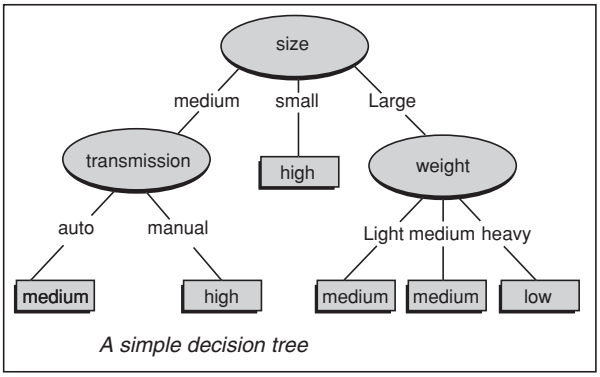

Fig. 7.4. A Simple Decision Tree

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

The most common machine learning methods used for data mining include decision tree, 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. It chooses the branches according to the attribute values of the object. Decision trees are induced from the training set. 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 Fig. 7.4. It determines a car's mileage from its size, transmission type, and weight. The leaf nodes are in square boxes that represent the three mileage classes. From the decision tree, we can conclude, for example, a medium size, automatic car will have medium mileage.

Database-oriented approaches. Database-oriented methods do not search for a best model, as do the previous two methods. Instead, data modeling or database specific heuristics are used to exploit the characteristics of 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 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 the patterns.

The left side of Fig. 7.5 shows a simple conceptual hierarchy for students. The right side shows an example of attribute-oriented induction. In the example, the students of a local IEEE chapter are summarized.

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 device that computes a function of its inputs. The inputs 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 class.

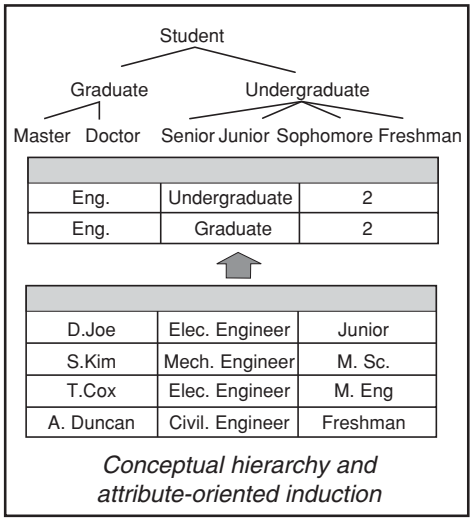

Fig. 7.5. Conceptual hierarchy and attribute-oriented induction

A rough set is a set whose membership is fuzzy. A set of objects can be arranged to form a group of rough sets for use, in say, classification and clustering.

Visual exploration is another interesting data mining technique. Data are transformed into visual objects such as dots, lines, and areas. The data is then displayed in a two- or three-dimensional space. Users can interactively explore the interesting spots by visual examination.

These methods can be integrated or combined to deal with complicated probabilities, or provide solutions. For example, data is visually summarized in charts, graphs, and such. This is done to help us understand the results and allow 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.

7.2.3 Applications

Data mining techniques have been applied successfully in many areas, from business, science, to sports.

Business applications. Many organizations now employ data mining as a secret weapon to keep or gain a competitive edge. Data mining has been used in data base marketing, retail data analysis, stock selection, credit approval, etc. For example, Mellon Bank, a Pittsburgh-based $40 billion financial services company, is applying IBM's intelligent Miner on its customer database in an effort to retain profitable customers.

- Database marketing is a very successful and popular business application of data mining. By mining historical customer databases, patterns and trends are extracted. Customer profiles are built from these results to produce more effective marketing.
- Retail databases contain customer-shopping transactions. Data mining can find customer-shopping patterns that can be used, for example, in a sales campaign.
- Using data mining techniques, investors can build models to predict the performance of stocks. By searching trends and patterns in stock data, data mining can also help investors to find stocks with good performances.
- 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, the 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 satellite images.

Other applications. Data mining techniques have also been used in health care management, tax fraud detection, money laundering monitoring, and 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.

7.2.4 Data Mining Applications – Survey

Numerous verification data driven and some discovery-driven data mining applications and application prototypes are already available for domains that range from credit risk analysis, fraud detection, to astronomy. Several of these applications, especially those using verification-driven data mining techniques, are currently in daily use and have been described in detail elsewhere. Of the already-deployed applications that use discover-driven data-mining techniques, three warrant particular mention:

- The Falcon credit-risk assessment system from HNC Inc., used by a large percentage of retail banks to detect suspicious credit-card transactions.
- JPL's Skicat system, which astronomers use to identify and classify new types of celestial objects, and
- The FAIS system from the Financial Crimes Enforcement Network (FinCEN), which identifies financial transactions that might indicate money laundering.

These three applications use the model creation as their primary data-mining operation, often supplemented with clustering and visualization techniques. They balance high predictive accuracy with the ability to use mined knowledge to explain the actions they recommended to their users. They also took significant time to implement and deploy. The Falcon system builds upon HNC's neural network shell. The fact that only a few companies capture credit card transaction data facilitated Falcon's deployment. Even though each of the companies uses its own data format, every bank issuing credit cards uses one of these few formats. Therefore, an application that works effectively with even one format can be easily adopted by a large number of banks.

The Skicat system, developed from scratch for one particular group of astronomers at JPL, uses data collected from astronomical instruments. It is also finding a larger user community, because astronomers also use data with common formats.

The FIAS application uses data from common government forms and consists of a combination of off-the-shelf and custom-built components. Its use should expand to a variety of government agencies concerned with detecting suspicious financial transactions that might indicate money laundering. Of the three systems, this last one faces the hardest data quality problem because it mines data that is often entered by hand, frequently hand written.

Financial analysis applications that employ predictive modeling techniques – whether they are statistical regression, neural networks, or symbolic model-

ing techniques – run into a special problem. Such applications perform portfolio creation and optimization, trading model creation and so forth, and have been in use for several years. To maintain a competitive advantage, the users and developers of such applications do not publicize their use and effectiveness. In these applications, predictive accuracy is much more important than the ability to use the extracted knowledge to explain a recommended action. Thus, the modeling methods used must first ensure that they do not overfit the data. The term *data mining* has always been associated with model overfit and has consequently been rejected by the financial analysis community.

Developers have recently incorporated discovery-driven data mining techniques in two classes of decision support applications: Market basket analysis and customer segmentation. Off the shelf versions of these two applications currently assist strategic marketing personnel in gaining significant insight about consumer behavior.

Market basket analysis: MBA refers to the process of examining point-of-scale data to identify affinities between products and services purchased by the customer. For example, after analyzing POS data from a particular set of supermarkets, one might identify that when the sales of lettuce increase, so do the sales of tomatoes and salad dressing. MBA is usually one step in larger-scale analyses aimed at such tasks as planning the replenishment of a store, assessing the effectiveness of an organization's promotions, and fine-tuning manufacturing and distribution operations. Because retail organizations have traditionally used MBA, typical users of these applications include marketing analysis, merchandise buyers, and store managers, catalog managers, and brand managers.

Early implementations of MBA applications used query-and-reporting data mining techniques and applied them to summaries aggregations and specialized views of the captured POS data, as well as to data about the purchases of closely monitored customer groups. For example, A.C Nielsen monitors the purchase of 40,000 consumers and sells the data to corporations interested in performing MBA. By using query-and-reporting techniques on this type of data, organizations can extract item movement information such as "when the sales of brand X coffee increase so do the sales of brand Y sugar."

Extracting information from summarized data was, and often still is, necessary because of the volumes of the captured data. For example, because a supermarket **or** a department store routinely stocks over 100,000 type of merchandise, the resulting POS databases of retail chains routinely consume several gigabytes per week. Until parallel database technology recently emerged, the query processing times needed to execute each query deterred any one from posing queries against such databases.

More recently, tools that can perform multidimensional analysis, such as DSPlus from Kelly Information Systems and DSS Agent from Micro strategy, have emerged for further improving the quality of MBA results produced by query and reporting tools. Using multidimensional analysis tools, an analyst can easily ask queries such as "show me the three highest selling products

of the produce department for the first quarter of 1994–1995, for the northeastern United States, organized store." The dimensions used in this query are department, time, geographic region, and organization (store). These tools continue to work with summarized data, and like any other verification-driven data mining tools rely on the analyst to postulate hypothesis and convert the returned data into information. The size of the multidimensional databases being deployed with MBA application varies from 10 to 30 Gbytes, to more than 100Gbytes. Corporations such as Mervyn's and Target Stores have recently reported successful deployment of multidimensional databases for MBA.

Two major retailing trends today are the emergence of megastore chains (home improvement chains and discount clothing chains being two examples), and the lack of differentiation in the customer's eyes between the offering of any two retail chains with in a particular segment, such as home improvement supermarkets. To improve their profits under such conditions, retailers attempt to identify micromarkets and exploit them before competition does. The recent mergers, Chase Bank with Chemical, for example, of large banks and insurance companies also create large financial institutions with little differentiation among themselves. Such organizations will start relying heavily on micromarket identification for improving their financial performance and growth. While the remainder of this section discusses how retailers use MBA, large financial institutions are also adapting MBA Applications. Summarized data that is often used today in conjunction with verification-driven data mining is not effective for micromarket identification; summarizing data removes individual consumer characteristics which once detected, could have helped identity micromarkets.

Micromarket identification requires:

- The use of the item's actual POS transactions, so that one can find out exactly what each customer is buying at the transaction level rather than at some aggregated level, whether that customer is buying an item on sale, at what time he or she is making a particular trip to the store, and so forth.
- Overlaying POS data with demographic and life style consumer data to identify other customer characteristics that may be correlated to the consumer's choices and
- The use of discovery-driven techniques (primarily tools that support database segmentation and link analysis operations) that can automatically identify important buying patterns and the types of consumers exhibiting such patterns.

For example, a marketing executive could use database segmentation tools on a set of supermarket POS transactions to identify product drivers ("lettuce and whole grain breads drive a large percentage of the sales in suburban supermarkets in Area X") in particular subset of the transactions. She could then use association rule discovery techniques on each of the identified clusters to automatically extract rules, such as "Customers who buy lettuce and tomatoes also buy bacon 80% of the time."

Recently, IBM consultants worked with a supermarket chain's business analysts to perform MBA using neural segmentation and association rule discovery techniques. Together, they discovered that shoppers use certain of the chain's supermarket as convenience stores during Friday and Saturday nights. Based on this discovery, a super market chain decided to start offering video rentals in these stores. This decision provided a chain with a good way to differentiate itself from competing supermarket chains and to identify a new source of revenue. The number of consumers going to the chain's supermarkets increased, there by increasing the general sales volume, and the sales of items such as popcorn, potato chips, and soft drinks also jumped significantly. Through this action the chain realized significant incremental sales and profit.

The application of discovery-driven data mining techniques for MBA analysis imposes two requirements on database management systems. First, they must deal with large volumes of transactional and overlaid data. In particular, while it has been a common practice to perform discovery-driven data mining on sampled data rather than the entire database, one might miss a particular micro market by doing so. Second, the systems must perform the mining operation within a time interval that will allow an organization to respond to a market opportunity before the competition does. For example, catalog companies send out a new issue of each catalog about every two months. During this eight-week period, they must collect the data about a set of consumers (collecting sales data from the previous issue of the catalog and combining it with demographic data), mine the collected data, determine the consumers to whom a particular catalog will be sent (identify the customers segments or micromarkets), and prepare and send the catalog.

Customer segmentation: Customer segmentation is the process of analyzing the data about the actual customers or general consumers to identify characteristics and behaviors that can be exploited in the market place. For example, an insurance company might want to identify the characteristics of customers with automotive policies who have let their policies lapse, so that it can establish a set of strategies for convincing the highest valued of these customers to renew their policies. Ultimately, customer segmentation allows an organization to view each of its customers as a segment of one, thereby forming a highly personalized relationship with each customer. Organizations apply customer segmentation to both actual customers and general consumers; consumer data analysis identifies characteristics that will allow the creation of strategies for tuning targeted consumers into actual customers.

Organizations use customer segmentation to address two basic marketing problems: understanding the causes of customer action so that it can be reduced and identifying new customers.

Customer attrition. Policy lapsing is a form of attrition. Others include switching credit cards, switching mortgage institutions, and buying merchandise from different direct-mail company. In order to identify the causes of an attrition problem, a customer segmentation application must also automati-

cally establish which customers the organization should try to keep and which should be left to defect – most organizations derive their profits from only a percentage of their customer base.

Target marketing. Organizations are invested in two types of target market. First, they want to know how to offset particular product or service to a specific set of consumers to turn them into customers. Every direct-mail company addresses this problem each time it creates a mailing list of the households that receives the next issue of its catalog. Second, companies want to determine which customers of one of their other products they should offer a product or service. Also known as cross selling, this action aims to convince customers to also the offered product. For example, a company might try to convince certain of its credit card customers to apply for a second mortgage. The two questions that will be answered in cross- selling scenario are "To which customers should a particular product be offered?" and "Which products should be offered to a particular customer?"

Typical users of this application include marketing analysts, brand managers, risk managers, and other types of business analysts. As in MBA, early implementations of customer segmentation applications rely on verification-driven data mining techniques. In particular, such applications used statistical analysis techniques – for example, binary and logistic regression – to create sector models. Analysts then applied these models to larger databases to assess an individual risk or propensity to lapse for example, they selected the data used to create model either by sampling a database with customer or customer data, or by selecting a set of customers through queries.

The use of sampling led to models could only cover broad customer characteristics. The use of queries to select a customer could lead to more targeted models, but the analyst must determine the features to include in the query, and thus cannot detect and capitalize market shifts and micromarket characteristics. Models developed using linear and logistic regression techniques can only account for certain characteristics in the data; essentially these techniques cannot account for "every ounce of information" in the data. The data that was used was also not very rich itself. Limitations in database technology and in the means of obtaining customer and customer data prevented users from performing fine-grained analysis even with the available techniques. Finally because statistical analysis tools are primarily targeted to data analysis, early implementations of customer segmentation applications were used by organizations staffed primarily by statisticians. Such organizations reside in large corporations and specialized service bureaus.

The customer segmentation applications being developed and deployed today arise in response to several trends. First, the variety and quality of the captured a data have increased. In fact, organizations currently capture more data than they use, because the improvements in database technology have simplified data maintenance. Second, having solved the data capture and maintenance problem, business units increasingly want to take ownership of their data, rather than relying on specialized departments and service

bureaus. Also early experimentation with discovery driven data mining techniques – primarily clustering, deviation detection, and modeling (from the symbolic and neural induction regimes) has demonstrated their advantages over traditional techniques.

The challenge that has to be addressed was making the applications embedding such techniques more appropriate to business users, not just statisticians. Symbolic learning techniques (modeling or clustering) have contributed significantly to this need, because they express the information they extract in forms that are comprehensible by business users. However, the new generation of customer-segmentation application goes a step further. By embedding this logic and knowledge into the application itself, they have begun eliminating the need for users to deeply understand data analysis and modeling processes.

Customer segmentation applications with these characteristics are now at work in the research and strategy departments of major banks, the customer-relations management departments of banks, insurance companies, retail organizations, as well as the marketing departments of major airlines. IBM recently reported that through the use of such an application, Legal and General, a major insurance company, conducted a successful target marketing campaign whose results improved significantly on similar campaigns run using more traditional customer segmentation applications.

The effectiveness of discovery-driven data mining techniques in extracting valuable and previously unknown information from large, complex databases is speeding acceptance of the two marketing applications now being deployed aggressively by various organizations. As the value these early adopters derive becomes apparent, applications using discovery-driven data mining techniques will proliferate, drawing data mining into the mainstream.

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. More work is needed, although a great deal of progress has been made in research and development.

7.3 Summary

Data mining is becoming an integral part of the operations in organizations of varying sizes. Many organizations that only recently have begun analyzing their data have started to successfully use applications employing verifications-driven data mining techniques. Applications using discovery-driven techniques are also finding increased use. While many of the deployed applications primarily employ predictive modeling techniques, application developers and end users alike are beginning to recognize the need to use additional techniques from the discovery-driven data mining repertory. Applications with broad market appeal such as MBA and customer segmentation have successfully demonstrated the advantage of using such techniques.

Today, the development of data mining applications remains a tedious process. Applications recognized as successful invariably require the cooperation of business analysts and developers of generic data mining tools. At these early phases, this is not necessarily a drawback; most successful applications in any discipline require the collaboration of end users and developers. However, a phase must quickly ensue in which business analyst can quickly develop and specialize data mining applications, much as is happening with database technology.

7.4 Review Questions

1. Explain the data mining process and its operational techniques.
2. What are the characteristics of data mining?
3. Give some of the data mining tasks and techniques involved in processing of the data.
4. Give details on performance survey on data mining applications.