1

Introduction to Data Mining Principles

Objectives:

- This section deals with detailed study of the principles of data warehousing, data mining, and knowledge discovery.
- The availability of very large volumes of such data has created a problem of how to extract useful, task-oriented knowledge.
- The aim of data mining is to extract implicit, previously unknown and potentially useful patterns from data.
- Data warehousing represents an ideal vision of maintaining a central repository of all organizational data.
- Centralization of data is needed to maximize user access and analysis.
- Data warehouse is an enabled relational database system designed to support very large databases (VLDB) at a significantly higher level of performance and manageability.
- Due to the huge size of data and the amount of computation involved in knowledge discovery, parallel processing is an essential component for any successful large-scale data mining application.
- Data warehousing provides the enterprise with a memory. Data mining provides the enterprise with intelligence.
- Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, visualization, and neural networks.
- We analyze the knowledge discovery process, discuss the different stages of this process in depth, and illustrate potential problem areas with examples.

Abstract. This section deals with a detailed study of the principles of data warehousing, data mining, and knowledge discovery. There exist limitations in the traditional data analysis techniques like regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, and stochastic models. Even though these techniques have been widely used for solving many practical problems, they are however primarily oriented toward the extraction

A. Lew and H. Mauch: *Introduction to Data Mining Principles*, Studies in Computational Intelligence (SCI) **38**, 1–20 (2006)
www.springerlink.com

of quantitative and statistical data characteristics. To satisfy the growing need for new data analysis tools that will overcome the above limitations, researchers have turned to ideas and methods developed in machine learning. The efforts have led to the emergence of a new research area, frequently called *data mining and knowledge discovery*. Data mining is a multidisciplinary field drawing works from statistics, database technology, artificial intelligence, pattern recognition, machine learning, information theory, knowledge acquisition, information retrieval, high-performance computing, and data visualization. Data warehousing is defined as a process of centralized data management and retrieval.

1.1 Data Mining and Knowledge Discovery

An enormous proliferation of databases in almost every area of human endeavor has created a great demand for new, powerful tools for turning data into useful, task-oriented knowledge. In the efforts to satisfy this need, researchers have been exploring ideas and methods developed in machine learning, pattern recognition, statistical data analysis, data visualization, neural nets, etc. These efforts have led to the emergence of a new research area, frequently called *data mining and knowledge discovery*.

The current Information Age is characterized by an extraordinary growth of data that are being generated and stored about all kinds of human endeavors. An increasing proportion of these data is recorded in the form of computer databases, so that the computer technology may easily access it. The availability of very large volumes of such data has created a problem of how to extract form useful, task-oriented knowledge.

Data analysis techniques that have been traditionally used for such tasks include regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, stochastic models, time series analysis, nonlinear estimation techniques, and others. These techniques have been widely used for solving many practical problems. They are, however, primarily oriented toward the extraction of quantitative and statistical data characteristics, and as such have inherent limitations.

For example, a statistical analysis can determine covariances and correlations between variables in data. It cannot, however, characterize the dependencies at an abstract, conceptual level and procedure, a casual explanation of reasons why these dependencies exist. Nor can it develop a justification of these relationships in the form of higher-level logic-style descriptions and laws. A statistical data analysis can determine the central tendency and variance of given factors, and a regression analysis can fit a curve to a set of datapoints. These techniques cannot, however, produce a qualitative description of the regularities and determine their dependence of factors not explicitly provided in the data, nor can they draw an analogy between the discovered regularity and regularity in another domain.

A numerical taxonomy technique can create a classification of entities and specify a numerical similarity among the entities assembled into the same or

different categories. It cannot, however, build qualitative description of the classes created and hypothesis reasons for the entities being in the same category. Attributes that define the similarity, as well as the similarity measures, must be defined by a data analyst in advance. Also, these techniques cannot by themselves draw upon background domain knowledge in order to automatically generate relevant attributes and determine their changing relevance to different data analysis problems.

To address such tasks as those listed above, a data analysis system has to be equipped with a substantial amount of background and be able to perform symbolic reasoning tasks involving that knowledge and the data. In summary, traditional data analysis techniques facilitate useful data interpretations and can help to generate important insights into the processes behind the data. These interpretations and insights are the ultimate knowledge sought by those who build databases. Yet, such knowledge is not created by these tools, but instead has to be derived by human data analysis.

In efforts to satisfy the growing need for new data analysis tools that will overcome the above limitations, researchers have turned to ideas and methods developed in machine learning. The field of machine learning is a natural source of ideas for this purpose, because the essence of research in this field is to develop computational models for acquiring knowledge from facts and background knowledge. These and related efforts have led to the emergence of a new research area, frequently called *data mining and knowledge discovery.*

There is confusion about the exact meaning of the terms "data mining" and "KDD." KDD was proposed in 1995 to describe the whole process of extraction of knowledge from data. In this context, knowledge means relationships and patterns between data elements. "Data mining" should be used exclusively for the discovery stage of the KDD process.

The last decade has experienced a revolution in information availability and exchange via the Internet. The World Wide Web is growing at an exponential rate and we are far from any level of saturation. E-commerce and other innovative usages of the worldwide electronic information exchange have just started. In the same spirit, more and more businesses and organizations have begun to collect data on their own operations and market opportunities on a large scale. This trend is rapidly increasing, with recent emphasis being put more on collecting the right data rather than storing all information in an encyclopedic fashion without further using it. New challenges arise for business and scientific users in structuring the information in a consistent way. Beyond the immediate purpose of tracking, accounting for, and archiving the activities of an organization, this data can sometimes be a *gold mine* for strategic planning, which recent research and new businesses have only started to tap. Research and development in this area, often referred to as *data mining and knowledge discovery*, has experienced a tremendous growth in the last couple of years. The goal of these methods and algorithms is to extract useful regularities from large data archives, either directly in the form of "knowledge" characterizing the relations between the variables of interest,

or indirectly as functions that allow to predict, classify, or represent regularities in the distribution of the data.

What are the grand challenges for information and computer science, statistics, and algorithmics in the new field of data mining and knowledge discovery? The huge amount of data renders it possible for the data analysis to infer data models with an unprecedented level of complexity. Robust and efficient algorithms have to be developed to handle large sets of high-dimensional data. Innovations are also required in the area of database technology to support interactive data mining and knowledge discovery. The user with his knowledge and intuition about the application domain should be able to participate in the search for new structures in data, e.g., to introduce a *priori* knowledge and to guide search strategies. The final step in the inference chain is the validation of the data where new techniques are called for to cope with the large complexity of the models.

Statistics as the traditional field of inference has provided models with more or less detailed assumptions on the data distribution. The classical theory of Bayesian inference has demonstrated its usefulness in a large variety of application domains ranging from medical applications to consumer data and market basket analysis. In addition to classical methods, neural networks and machine learning have contributed ideas, concepts, and algorithms to the analysis of these data sets with a distinctive new flavor. The new approaches put forward by these researchers in the last decade depart from traditional statistical data analysis in several ways: they rely less on statistical assumptions on the actual distribution of the data, they rely less on models allowing simple mathematical analysis, but they use sophisticated models that can learn complicated nonlinear dependencies from large data sets. Whereas statistics has long been a tool for testing theories proposed by scientists, machine learning and neural network research are rather evaluated on the basis of how well they *generalize on new data*, which come from the same unknown process that generated the training data. Measuring the generalization performance to select models has to be distinguished from the widespread but questionable current practice of data inquisition where "the data are tortured until they confess."

During the last 15 years, various techniques have been proposed to improve the generalization properties of neural estimators. The basic mechanism is to control the richness of the class of possible functions that can be obtained through training, which has been quantified with the seminal work of Vapnik and Chervonenkis on the "*capacity of a hypothesis class*." The combinational concept of the *VC* dimensions and its generalizations parameterize a rigorous but loose upper bound on large deviations of the empirical risk from the expected risk of classification or regression. Such theoretical bounds can help us understand the phenomenon of generalization. To answer a numerical question about a particular algorithm and data set, purely quantitative empirical bounds on the expected generalization error can be obtained by repeating many training/test simulations, and they are tighter than the analytic

theoretical bounds. Heuristics that essentially implement complexity control in one way or another are the widely used weight decay in training multilayer perceptrons or the early stopping rule during training. It is also possible to view capacity control in terms of penalty terms for too complex estimators.

Complexity control is particularly relevant for data mining. In this area, researchers look for complex but still valid characterizations of their large data sets. Despite the large size of the data sets inference often takes place in the small sample size limit. It should be noted that the ratio of samples to degrees of freedom might be small even for large data sets when complex models like deep decision trees or support vector machines in high-dimensional spaces are used. Complexity control, either by numerical techniques like cross validation or by theoretical bounds from computational learning theory with empirical rescaling, is indispensable for data mining practitioners.

The enterprise of knowledge discovery aims at the automation of the millennium-old effort of humans to gain information and build models and theories about phenomena in the world around us. Data miners and knowledge discoverers can learn a lot and, i.e., sharpen their awareness, by looking at the scientific method of experimentation, modeling, and validation/falsification in the natural sciences, engineering sciences, social sciences, economics, as well as philosophy.

The next decade of research in network-based information services promises to deliver widely available access to unprecedented amounts of constantly expanding data. Users of many commercial, government, and private information services will benefit from new machine learning technologies that mine new knowledge by integrating and analyzing very large amounts of widely distributed data to uncover and report upon subtle relationships and patterns of events that are not immediately discernible by direct human inspection.

1.2 Data Warehousing and Data Mining - Overview

The past decade has seen an explosive growth in database technology and the amount of data collected. Advances in data collection, use of bar codes in commercial outlets, and the computerization of business transactions have flooded us with lots of data. We have an unprecedented opportunity to analyze this data to extract more intelligent and useful information, and to discover interesting, useful, and previously unknown patterns from data. Due to the huge size of data and the amount of computation involved in knowledge discovery, parallel processing is an essential component for any successful large-scale data mining application.

Data mining is concerned with finding hidden relationships present in business data to allow businesses to make predictions for future use. It is the process of data-driven extraction of not so obvious but useful information from large databases. Data mining has emerged as a key business intelligence technology.

The explosive growth of stored data has generated an information glut, as the storage of data alone does not bring about knowledge that can be used: (a) to improve business and services and (b) to help develop new techniques and products. Data is the basic form of information that needs to be managed, sifted, mined, and interpreted to create knowledge. Discovering the patterns, trends, and anomalies in massive data is one of the grand challenges of the Information Age. Data mining emerged in the late 1980s, made great progress during the Information Age and in the 1990s, and will continue its fast development in the years to come in this increasingly data-centric world. Data mining is a multidisciplinary field drawing works from statistics, database technology, artificial intelligence, pattern recognition, machine learning, information theory, knowledge acquisition, information retrieval, high-performance computing, and data visualization.

The aim of data mining is to extract implicit, previously unknown and potentially useful (or actionable) patterns from data. Data mining consists of many up-to-date techniques such as classification (decision trees, naïve Bayes classifier, k-nearest neighbor, neural networks), clustering (k-means, hierarchical clustering, density-based clustering), association (one-dimensional, multidimensional, multilevel association, constraint-based association). Many years of practice show that data mining is a process, and its successful application requires data preprocessing (dimensionality reduction, cleaning, noise/outlier removal), postprocessing (understandability, summary, presentation), good understanding of problem domains and domain expertise.

Today's competitive marketplace challenges even the most successful companies to protect and retain their customer base, manage supplier partnerships, and control costs while at the same time increasing their revenue. In a world of accelerating change, competitive advantage will be defined by the ability to leverage information to initiate effective business decisions before competition does. Hence in this age of global competition accurate information plays a vital role in the insurance business. Data is not merely a record of business operation – it helps in achieving competitive advantages in the insurance sector. Thus, there is growing pressure on MIS managers to provide information technology (IT) infrastructure to enable decision support mechanism. This would be possible provided the decision makers have online access to previous data. Therefore, there is a need for developing a data warehouse. Data mining as a tool for customer relationship management also has proved to be a means of controlling costs and increase revenues.

In the last decade, machine learning had come of age through a number of ways such as neural networks, statistical pattern recognition, fuzzy logic, and genetic algorithms. Among the most important applications for machine learning are classification, recognition, prediction, and data mining. Classification and recognition are very significant in a lot of domains such as multimedia, radar, sonar, optical character recognition, speech recognition, vision, agriculture, and medicine. In this section, the concept of data warehousing and data mining is briefly presented.

1.2.1 Data Warehousing Overview

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. *Data warehousing* is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access these data freely. The data analysis software is what supports data mining. Hence, data warehousing provides the enterprise with a memory. Data mining provides the enterprise with intelligence.

Data warehouse is an enabled relational database system designed to support very large databases (VLDB) at a significantly higher level of performance and manageability. Data warehouse is an environment, not a product. It is an architectural construct of information that is hard to access or present in traditional operational data stores.

Any organization or a system in general is faced with a wealth of data that is maintained and stored, but the inability to discover valuable, often previously unknown information hidden in the data, prevents it from transferring these data into knowledge or wisdom.

To satisfy these requirements, these steps are to be followed.

1. Capture and integrate both the internal and external data into a comprehensive view "Mine" for the integrated data information
2. Organize and present the information and knowledge in ways that expedite complex decision making.

Access Tools for Data Warehousing

The principal purpose of data warehousing is to provide information to users for strategic decision making. These users interact with the data warehouse using front-end tools. Many of these tools require an information specialist, although many end users develop expertise in the tools. The access tools are divided into five main groups.

1. Data query and reporting tools
2. Application development tools
3. Executive information system (EIS) tools
4. Online analytical preprocessing tools and
5. Data mining tools

Data mining tools are considered for information extraction from data. In recent research, data mining through pattern classification is an important area of concentration.

1.2.2 Concept of Data Mining

Database technology has been used with great success in traditional business data processing. There is an increasing desire to use this technology in new application domains. One such application domain that is likely to acquire considerable significance in the near future is database mining. An increasing number of organizations are creating ultralarge databases (measured in gigabytes and even terabytes) of business data, such as consumer data, transaction histories, sales records, etc.; such data forms a potential gold mine of valuable business information.

Data mining is a relatively new and promising technology. It can be defined as the process of discovering meaningful new correlation, patterns, and trends by digging into (mining) large amounts of data stored in warehouse, using statistical, machine learning, artificial intelligence (AI), and data visualization techniques. Industries that are already taking advantage of data mining include medical, manufacturing, aerospace, chemical, etc. Knowledgeable observers generally agree that in-depth decision support requires new technology. This new technology should enable the discovery of trends and predictive patterns in data, the creation and testing of hypothesis, and generation of insight-provoking visualizations.

Data mining helps the end users to extract useful information from large databases. These large databases are present in data warehouses, i.e., "data mountain," which are presented to data mining tools. In short data warehousing allows one to build the data mountain. Data mining is the nontrivial extraction of implicit, previously unknown and potentially useful information from the data mountain. This data mining is not specific to any industry – it requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. Data mining is also referred to as knowledge discovery in databases (KDD).

Data, Information, and Knowledge

Data: Data are any facts, numbers, or text that can be processed by a computer. Today organizations are accumulating vast and growing amounts of data in different formats and databases.

This includes: Operational or transactional data such as sales, cost, inventory, payroll, and accounting.

Nonoperational data like industry sales, forecast data, and macroeconomic data.

Metadata: data about the data itself such as logical database design or data dictionary definitions.

Information: The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point-of-sale transaction data can yield information on which products are selling and when.

Knowledge: Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge or consumer buying behavior. Thus a manufacturer or a retailer could determine those items that are most susceptible to promotional efforts.

Data Mining Definitions

- Data mining is the efficient discovery of valuable, nonobvious information from a large collection of data.
- Knowledge discovery in databases is the nontrivial process of identifying valid novel potentially useful and ultimately understandable patterns in the data.
- It is the automatic discovery of new facts and relationships in data that are like valuable nuggets of business data.
- It is not a complex query where the user already has a suspicion about a relationship in the data and wants to pull all such information.
- The information discovered should give competitive advantage in business.
- Data mining is the induction of understandable models and patterns from a database.
- It is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions.

It is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, visualization, and neural networks.

Data mining is streamlining the transformation of masses of information into meaningful knowledge. It is a process that helps identify new opportunities by finding fundamental truths in apparently random data. The patterns revealed can shed light on application problems and assist in more useful, proactive decision making. Typical techniques for data mining involve decision trees, neural networks, nearest neighbor clustering, fuzzy logic, and genetic algorithms.

Now we focus on the relationship between data mining and data warehousing.

What is a data warehouse and why do we need it?

In most organizations we find really large databases in operation for normal daily transactions. These databases are known as operational databases; in

most cases they have not been designed to store historical data or to respond to queries but simply support all the applications for day-to-day transactions. The second type of database found in organizations is the data warehouse, which is designed for strategic decision support and largely built up from operational databases. Small, local data warehouses are called data marts.

Rules for Data Warehouses:

- Time dependent
- Nonvolatile data in data warehousing is never updated but used only for queries. This means that a data warehouse will always be filled with historical data.
- Subject oriented
- Integrated

A data warehouse is designed especially for decision support queries; therefore only data that is needed for decision support will be extracted from the operational data and stored in data warehouse. Setting up a data warehouse is the most appropriate procedure for carrying out decision support. A decision support system can constantly change if the requirement of the organization alters, then the data model must also change. The data warehouse requires a high-speed machine and a wide variety of optimization processes.

- Metadata: describes the structure of the contents of a database.

Designing Decision Support Systems

The design of a decision support system differs considerably from that of an online transaction processing system. The main difference is that decision support systems are used only for queries, so their structure should be optimized for this use. When designing a decision support system, particular importance should be placed on the requirements of the end user and the hardware and software products that will be required.

The Requirements of EndUser

Some end users need specific query tools so that they can build their queries themselves, others are interested only in a particular part of the information. They may also need trend analysis tools and GUI user interface.

Software Products of Decision Support Systems

The types of software we choose depend very much on the requirements of end users. For data mining we can split the software into two parts: the first works with the algorithms on the database server and the second on the local workstation. The latter is mostly used to generate screens and reports for end users for visualizing the output of algorithm.

Hardware Products of Decision Support Systems

The hardware requirements depend on the type of data warehouse and techniques with which we want to work.

Integration with Data Mining

The application of data mining techniques can be carried out in two ways: from the existing data warehouse, or by extracting from the existing data warehouse the part of the information that is of interest to the end user and copying it to a specific computer, possibly a multiprocessing machine.

Integration of data mining in a decision support system is very helpful. There are several types of data mining technique and each uses the computer in a specific way. For this reason it is important to understand the demands of the end user so that we are able to build a proper data warehouse for data mining. In many cases we will find that we need a separate computer for data mining.

Client/Server and Data Warehousing

The end user would ideally like to have available all kinds of techniques such as graphical user interfaces, statistical techniques, windowing mechanisms, and visualization techniques so that they can easily access the data being sought. This means that a great deal of local computer power is needed at each workstation, and the client/server technique is the solution to this problem.

With client/server we only have to change the piece of software that is related to the end use—the other applications do not require alteration. Of all the techniques currently available on the market, client/server represents the best choice for building a data warehouse.

Replication techniques are used to load the information from the operational database to the data warehouse. If we need immediate access to the latest information, then we need to work with the more advanced replication tools; if the update of the data warehouse is less urgent, then we can work with batch update of the database server.

Two basic techniques, known as the “top-down” and the “bottom-up” approaches, are used to build a data warehouse.

In the “top-down” approach, we first build a data warehouse for the complete organization and from this select the information needed for our department or for local end users. In the “bottom-up” approach, smaller local data warehouses, known as data marts, are used by end users at a local level for their specific local requirements.

Multiprocessing Machines

A data mining environment has specific hardware requirements. There are several types of multiprocessing machines and we describe the most important ones here:

- Symmetric multiprocessing

All processors work on one computer, are equal, and they communicate via shared storage. Symmetric multiprocessing machines share the same hard disk and the internal memory. At present, approximately twelve processors are the maximum.

- Massively parallel

This is a computer where each processor has its own operating system, memory, and hard disk. Although each processor is independent, communication between the systems is possible. In this type of environment one can work with thousands of processors.

Not all databases will support parallel machines but most modern databases are able to work with symmetric parallel machines. At present, only a few database vendors such as IBM with DB/2, Oracle, and Tandem are able to operate with massively parallel computers.

- Cost justification

It is difficult to give a cost justification for the implementation of a KDD environment. Basically the cost of using machine-learning techniques to recognize patterns in data must be compared with the cost of a human performing the same task.

The Knowledge Discovery Process

We analyze the knowledge discovery process, discuss the different stages of this process in depth, and illustrate potential problem areas with examples.

The knowledge discovery process consists of six stages:

- Data selection
- Cleaning
- Enrichment
- Coding
- Data mining
- Reporting

It is impossible to describe in advance all the problems that can be expected in a database, as most will be discovered in mining stage.

Data Selection and Cleaning: A very important element in a cleaning operation is the de-duplication of records. Although data mining and data cleaning are two different disciplines, they have a lot in a common and pattern recognition algorithms can be applied in cleaning data. One kind of errors is spelling errors. The second type of pollution that frequently occurs is lack of domain consistency. For instance, a transaction listed in table was completed in 1901 but the company was set up after 1901.

Enrichment: New information can easily be joined to the existing client records.

Coding: In most tables that are collected from operational data, a lot of desirable data is missing, and most is impossible to retrieve. We therefore have to make a deliberate decision either to overlook or to delete it. A general rule states that any deletion of data must be a conscious decision, after a thorough analysis of the possible consequences. We can remove some unrelated attributes form current tables. By this time, the information in database is much too detailed to be used as input for pattern recognition algorithms. For instance, address to region, birth data to age, divide income by 1000, etc.

Data mining: The discovery stage of the KDD process is fascinating. We now see that some learning algorithms do well on one part of the data set where others fail, and this clearly indicates the need for hybrid learning.

Although various different techniques are used for different purposes, those that are of interest in the present context are:

Query tool
Statistical techniques
Visualization
Online analytical processing (OLAP)
Case-based learning (K-Nearest Neighbor)
Decision trees
Association rules
Neural networks
Genetic algorithm

Preliminary Analysis of the Data Set Using Traditional Query Tools: The first step in a data mining project should always be a rough analysis of the data set using traditional query tools. Just by applying simple structured query language (SQL) to a data set, we can obtain a wealth of information. We need to know the basic aspects and structures of the data set. For the most part 80% of the interesting information can be abstracted from a database using SQL. The remaining 20% of hidden information needs more advanced techniques. A trivial result that is obtained by an extremely simple method is called a naïve prediction. We can never judge the performance of an advanced learning algorithm properly if we have no information concerning the naïve probabilities of what it is supposed to predict.

Visualization Techniques: Visualization techniques are a very useful method of discovering patterns in data sets and may be used at the beginning of a data mining process to get a rough feeling of the quality of the data set and where patterns are to be found. An elementary technique that can be of great value is the so-called *scatter diagram*. Scatter diagrams can be used to identify interesting subsets of the data sets so that we can focus on the rest of the data mining process. There is a whole field of research dedicated to the search for interesting projections of data sets – this is called projection pursuit. A much better way to explore a data set is through an interactive three-dimensional environment.

Likelihood and Distance: The space metaphor is very useful in data mining context. Records that are closer to each other are very alike, and those that are very far from each other represent individuals that have little in common. Sometimes it is possible to identify interesting clusters merely by visual inspection.

OLAP Tools: This idea of dimensionality can be expanded: a table with n independent attributes can be seen as an n-dimensional space. We need to explore the relationship between these dimensions as standard relational database is not very good at this. OLAP tools were developed to solve this problem. These tools store their data in a special multidimensional format.

OLAP can be an important stage in a data mining processes. However there is an important difference between OLAP and data mining: OLAP tools do not learn; data mining is more powerful than OLAP and also needs no special multi-dimensional storage.

K-Nearest Neighbor: When we interpret records as points in a data space, we can define the concept of neighborhood records that are close to each other live in each other's neighborhood. In terms of the metaphor of our multi-dimensional data space, a type is nothing more than a region in this data space. Based on this insight, we can develop a very simple but powerful learning algorithm – the k-nearest neighbor. The basic philosophy of k-nearest neighbor is "do as our neighbors do." If we want to predict the behavior of a certain individual, we start to look at the behaviors of its neighbors. The letter k stands for the number of neighbors we have investigated. Simple k-nearest neighbor is not really a learning algorithm, but more of a search method. In general data mining algorithms should not have a complexity higher than n (log n) (where n is the number of records). The other techniques such as decision trees, association rules, neural networks, and genetic algorithms are discussed in the following sections.

Principles of Data Mining

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. It discovers

information within the data that queries and reports cannot effectively reveal. The section explores many aspects of data mining in the following areas:

- Data rich, information poor
- Data warehouses
- What is data mining?
- What can data mining do?
- The evolution of data mining
- How data mining works
- Data mining technologies
- Real-world examples
- The future of data mining
- Privacy concerns

Data Rich, Information Poor

The amount of raw data stored in corporate databases is exploding. From trillions of point-of-sale transactions and credit card purchases to pixel-by-pixel images of galaxies, databases are now measured in gigabytes and terabytes. (One terabyte = one trillion bytes. A terabyte is equivalent to about 2 million books!) For instance, every day, Wal Mart uploads 20 million point-of-sale transactions to an A&T massively parallel system with 483 processors running a centralized database. Raw data by itself, however, does not provide much information. In today's fiercely competitive business environment, companies need to rapidly turn these terabytes of raw data into significant insights for their customers and markets to guide their marketing, investment, and management strategies.

Data Warehouses

The drop in price of data storage has given companies willing to make the investment a tremendous resource: Data about their customers and potential customers stored in "data warehouses." Data warehouses are becoming part of the technology. Data warehouses are used to consolidate data located in disparate databases. A data warehouse stores large quantities of data by specific categories; so it can be more easily retrieved, interpreted, and sorted by users. Warehouses enable executives and managers to work with vast stores of transactional or other data to respond faster to markets and make more informed business decisions. It has been predicted that every business will have a data warehouse within ten years. Companies will want to learn more about that data to improve knowledge of customers and markets. The companies benefit when meaningful trends and patterns are extracted from the data.

What is Data Mining?

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that were traditionally too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable one. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

What Can Data Mining Do?

Although data mining is still in its infancy, companies in a wide range of industries – including finance, health care, manufacturing, transportation,– are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques of sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions, and anomalies that might otherwise go unnoticed.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

Market segmentation – Identify the common characteristics of customers who buy the same products from your company.
Customer churn – Predict those customers who are likely to leave the company and go to a competitor.
Fraud detection – Identify transactions that are most likely to be fraudulent.
Direct marketing – Identify the prospects who should be included in a mailing list to obtain the highest response rate.
Interactive marketing – Predict what each individual accessing a web site is most likely interested in seeing.
Market basket analysis – Understand what products or services are commonly purchased together, e.g., beer and diapers.
Trend analysis – Reveal the difference in a typical customer between the current month and the previous one.

Data mining technology can generate new business opportunities by:

- Automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in large database. Questions

that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default and identifying segments of a population likely to respond similarly to given events.

- Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Using massively parallel computers, companies dig through volumes of data to discover patterns about their customers and products. For example, grocery chains have found that when men go to a supermarket to buy diapers, they sometimes walk out with a six-pack of beer as well. Using that information, it is possible to lay out a store so that these items are closer.

AT&T, A.C. Nielsen, and American Express are among the growing ranks of companies implementing data mining techniques for sales and marketing. These systems are crunching through terabytes of point-of-sale data to aid analysts in understanding consumer behavior and promotional strategies. Why? To gain a competitive advantage and increase profitability!

Similarly, financial analysts are plowing through vast sets of financial records, data feeds, and other information sources in order to make investment decisions. Health-care organizations are examining medical records to understand trends of the past so that they can reduce costs in the future.

The Evolution of Data Mining

Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts. Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship and verifies it or discounts it with a series of queries against the data. For example, an analyst might hypothesize that people with low income and high debt are bad credit risks and query the database to verify or disprove this assumption. Data mining can be used to generate a hypothesis. For example, an analyst might use a neural net to discover a pattern that analysts did not think to try – for example, that people over 30 years with low incomes and high debt but who own their own homes and have children are good credit risks.

How Data Mining Works

How is data mining able to tell us important things that we did not know or what is going to happen next? The technique that is used to perform these feats is called *modeling*. Modeling is simply the act of building a model (a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers are not known. Modeling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect and store huge amounts of data, and the computational power to automate modeling techniques to work directly on the data, have been available.

As a simple example of building a model, consider the director of marketing for a telecommunications company. He would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long-distance services. He knows a lot about his customers, but it is impossible to discern the common characteristics of his best customers because there are so many variables. From this existing database of customers, which contains information such as age, sex, credit history, income, zip code, occupation, etc., he can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long-distance calls. For instance, he might learn that his best customers are unmarried females between the ages of 34 and 42 who earn in excess of $60,000 per year. This, then, is his model for high-value customers, and he would budget his marketing efforts accordingly.

Data Mining Technologies

The analytical techniques used in data mining are often well-known mathematical algorithms and techniques. What is new is the application of those techniques to general business problems made possible by the increased availability of data, and inexpensive storage and processing power. Also, the use of graphical interface has led to tools becoming available that business experts can easily use.

Some of the tools used for data mining are:

Artificial neural networks – Nonlinear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees – Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Rule induction – The extraction of useful if-then rules from databases on statistical significance.

Genetic algorithms – Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.

Nearest neighbor – A classification technique that classifies each record based on the records most similar to it in a historical database.

Real-World Examples

Details about who calls whom, how long they are on the phone, and whether a line is used for fax as well as voice can be invaluable in targeting sales of services and equipment to specific customers. But these tidbits are buried in masses of numbers in the database. By delving into its extensive customer-call database to manage its communications network, a regional telephone company identified new types of unmet customer needs. Using its data mining system, it discovered how to pinpoint prospects for additional services by measuring daily household usage for selected periods. For example, households that make many lengthy calls between 3 p.m. and 6 p.m. are likely to include teenagers who are prime candidates for their own phone and lines. When the company used target marketing that emphasized convenience and value for adults – "Is the phone always tied up?"- hidden demand surfaced. Extensive telephone use between 9 a.m and 5 p.m. characterized by patterns related to voice, fax, and modem usage suggests a customer has business activity. Target marketing offering those customers "business communications capabilities for small budgets" resulted in sales of additional lines, functions, and equipment.

The ability to accurately gauge customer response to changes in business rules is a powerful competitive advantage. A bank searching for new ways to increase revenues from its credit card operations tested a nonintuitive possibility: Would credit card usage and interest earned increase significantly if the bank halved its minimum required payment? With hundreds of gigabytes of data representing two years of average credit card balances, payment amounts, payment timeliness, credit limit usage, and other key parameters the bank used a powerful data mining system to model the impact of the proposed policy change on specific customer categories, such as customers consistently near or at their credit limits who make timely minimum or small payments. The bank discovered that cutting minimum payment requirements for small, targeted customer categories could increase average balances and extend indebtedness periods, generating more than $25 million in additional interest earned.

The Future of Data Mining

In the short term, the results of data mining will be in profitable, if mundane, business-related areas. Micromarketing campaigns will explore new niches. Advertising will target potential customers with new precision.

In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to New York, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers.

The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on subatomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe. There are potential dangers, though, as discussed below.

Privacy Concerns

What if every telephone call we make, every credit purchase we make, every flight we take, every visit to the doctor we make, every warranty card we send in, every employment application we fill out, every school record we have, our credit record, every web page we visit... were all collected together? A lot would be known about us! This is an all-too-real possibility. Much of this kind of information is already stored in a database. Remember that phone interview we gave to a marketing company last week? Our replies went into a database. Remember that loan application we filled out? It is now in a database Too much information about too many people for anybody to make sense of? Not with data mining tools running on massively parallel processing computers! Would we feel comfortable about someone (or lots of someones) having access to all this data about us? And remember, all this data does not have to reside in one physical location, as the Net growing information of this type becomes more available to more people.

1.3 Summary

Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship. Data mining can be used to generate a hypothesis. The results of data mining will be in profitable. Data warehouses are used to consolidate data located in disparate databases. A data warehouse stores large quantities of data by specific categories, so it can be more easily retrieved, interpreted, and sorted by users. Thus in this section we have seen the overview of the data mining and data warehousing.

1.4 Review Questions

1. Give an overview of data mining and data warehousing.
2. Explain the concept of data mining.
3. Define the terms *data warehousing* and *data mining*.
4. What is data warehousing and why do we need it?
5. What are the rules for data warehouses?
6. What is the necessity for multiprocessing machines?
7. Explain the stages involved in knowledge discovery process with a neat schematic diagram.
8. What are the various techniques adapted of data mining?
9. State the principles of data mining.
10. Explain the performance of various data mining techniques.
11. How does the data mining methodology works?
12. Give some of the data mining technologies used at present.