Chapter 1

PREVIEW: DATA WAREHOUSING/MINING

For many who are interested in the topics related to data mining, a common question is "What exactly is data mining?" In the community of statistics, data mining can be perceived as investigating and applying statistical techniques for EDA — Exploratory Data Analysis [Tukey 1977]. In the community of machine learning or artificial intelligence, data mining can be about algorithms that are computationally practical for "learning patterns" from large data sets [Weiss 1991]. In the community of scientific visualization, data mining could be about clustering and compressing high dimensional data, and transforming the data into a visual form for conveying "useful" information in data [Santos 2002]. In the community of database technology, it could be about OLAP (On-Line Analytical Processing) [Harinarayan 1996], warehouse model, data mart, and decision support.

Irrespective to the communities, there is one consensus about data mining. Data mining is a field of study about discovering "useful summary information" from data. As innocent as it may look, there are important questions behind discovering "useful summary information." We will take an approach on answering these questions as a starting point of our adventure into data mining.

1. WHAT IS SUMMARY INFORMATION?

What exactly is information? And More specifically, what is summary information?

Certainly we should not consider data as equivalent to information — at least not summary information. Let's suppose we have a data set of daily weather temperature for a year. We can abstract the data by *12* pieces of summary information where each is a monthly average, or by one piece of summary information where each is an annual average.

In essence, summary information should be an abstraction that preserves certain properties of data. Specifically, we will focus on the statistical and probability properties of the data. By focusing on the statistical and probability properties of the data, we will be able to better understand how to interpret the meaning behind the information, to verify its correctness in the analytical process, and to validate its truthfulness with respect to its attempt on the characterization of a physical real world phenomenon. To illustrate this point, let's consider Duke's parapsychologist Dr. Rhine's "extrasensory perception" experiment in the 1950s [Rhine 1983]:

David Rhine tested students for "extrasensory perception" (ESP) by asking them to guess 10 cards — red or black. He found about *1/1000* of them guessed all 10. Instead of realizing that this is what one would expect from random guessing, he declared them to have ESP. When he retested them, he found that they did no better than the average. His conclusion: telling people they have ESP causes them to lose it!

This is a case in point about the importance of the statistical and probability properties of the data. Let’s put the ESP experiment in perspective. The experiment requires a student to guess one of the two colors. To interpret this within the framework of information theory, there is $log_2 2=1$ bit of information to discern on each guess; i.e., black or red. Under the assumption of independence and uniform distribution, the probability for an individual to correctly guess all *10* cards is *1/1024*. For a group of n students, the probability that there are k students correctly guessing all *10* cards is characterized by a binomial distribution. The expected number of students guessing all 10 cards correct out of a group of n is $\sum_{k=0}^{n} k$ x $C(n,k)$ x $(1/1024)^k$ x $(1 - (1/1024))^{n-k} = n/1024$. In other words, for a group of (n=) *1024* student, we expect to have one student who will correctly guess all *10* cards! Although it may be interesting to discover someone who correctly guessed all *10* cards, it certainly does not convey useful information from the analytical point of view! It is because from the statistical point of view, the observed number of students who correctly guess all *10* cards is expected to converge to the mean value; i.e., one in *1024*.

Furthermore, the observed result does not convey useful information about the characterization of the physical real world problem. Note that the observed count is coherent with the expected count of events occurred randomly. In other words, the observation of the ESP experiment will fail the chi-square test of independence. Therefore, one cannot conclude that the discovered ESP does not happen by chance.

2. DATA, INFORMATION THEORY, STATISTICS

What is the relationship between information and data from the perspective of information theory and statistics?

Summary information is an abstraction of data. Many interesting properties of data may be revealed in form of patterns. Therefore, information may be considered as a manifestation of (data) patterns, and its abstraction is scalable. Different kinds of patterns may be found in data; e.g., an upward/downward trend pattern of time series data, periodic recurrence pattern of time series data, statistical association pattern between the occurrence of two events, and a mathematical pattern encapsulating the characteristics of data. There are three important essences about the concept of patterns. First, it exhibits regularities explicitly. Second, it often offers a good compression in terms of Kolmogorov complexity or MDL [Li 1997]. Third, it offers explanation, inference and/or prediction capabilities. It would be left as an exercise for readers to identify additional patterns that may be interesting for data mining!

Consider the number sequence 1 4 1 5 9 2 6 …, is there any regularity that can be captured in form of a pattern? The answer is "yes" and the mathematical pattern $S(n) = (trunc((22/7 - 3)\cdot 10^n)) \bmod 10$ (for $n = 1\ 2,\ 3,\ ...$) will reproduce the number sequence. After a closer look, one may realize that the numbers are indeed the decimal points of π. Let's consider yet another example 2.3 4.6 9.2 18.4 36.8 …, is there any regularity in this data set? The answer is again “yes”. One may realize that there is a mathematical pattern $S(n) = 2.3*2^n$ for $n > 0$, and there is an upward trend pattern; i.e., $S(m) > S(n)$ when $m > n$. There are two important observations about this example. First, the mathematical pattern is a loss less abstraction of data; i.e., it can reconstruct the original data set with low Kolmogorov complexity. But it does not offer a good compression in comparison to the trend pattern in terms of MDL (minimum descriptive length) if one is interested in capturing only the monotonically increasing behavior of the data.

Information theory is an important conceptual tool for evaluating the quality and quantity of information embedded in data. Suppose we are interested in only the trend patterns, there are four possibilities: upward trend, downward trend, flat trend, and no trend. We will need two bits to represent the four trend patterns. If we are given a data set, how do we know whether there is "valuable" information about possible trends? We can answer this question using expected Shannon entropy [Shannon 1972] in information theory. Let's assume the data is truly random. We will expect all four trends to occur more or less equal number of times. In other words, expected Shannon entropy will result in $4*(1/4)log_2 4 = 2$ bits. On the other hand, if only upward and downward trends occur, expected Shannon entropy will yield only one bit. This means that the trend information embedded in this latter case is more valuable than the former case. It is because there are fewer cases to discern when one has to interpret the trend information carried by the data. This indeed is the basis of the principle of minimum information commonly used in model selection in the statistics community.

3. DATA WAREHOUSING/MINING MANAGEMENT

What is the management cycle of data warehousing and data mining?

Database technology plays a central role in the management cycle of data warehousing and data mining. Database system, data warehouse, and data mining for decision support system, each plays a distinctive role in the management cycle of data warehousing and data mining. Database system is focused on the operational level, while data mining is focused on the analytical level. The emphasis of database system is on day-to-day business operation support, while the emphasis of data mining is on decision support for strategic planning.

It is preferable to maintain the data of day-to-day business operation as a single source (or at least homogenous sources) in a database system. This is particularly essential for situations that involve frequent and large volume of data transactions. For example, a customer order may require linking information in an invoice table with the information about the items (in an inventory table) being purchased, as well as linking customer information with shipping address information. If all these different kinds of information about a customer order have to be redundantly entered into different sources or systems (e.g., one for fulfillment department, one for accounting department), the data entry process will likely be error-prone. Foreign key constraints enforcing referential integrity will help reduce data entry errors.

While a database system is developed for supporting daily business operation, a data warehouse is developed to "pull" data/information together. Data/information pulled into a data warehouse are typically historical data; i.e., the need for update is rare, if any. However, it is not unusual to incorporate multiple data sources, or even external sources, into a data warehouse model. For example, in-house customer data about monthly spending, and data about the credit rating of the customers bought from an outside agency, may be combined in the process of developing a data warehouse.

In database design, we focus on issues such as referential integrity ("one fact at one place"), functional dependency, normalization, and tuning for performance improvement. In data warehouse design, we focus on issues such as data purification/cleansing, star transformation, ELT (extraction, load, and transformation), and indexing for efficient query performance. Query performance is sometimes seen as an issue for data mart but not data warehouse. In our case, we decide not to make such a distinction.

Data purification is a particularly important issue of data warehouse project. Since it is fairly common that data for a warehouse project are obtained from multiple sources, there are many data quality and applicability issues. For example, NYC Park scientists made biweekly bird observations for studying migration patterns. Bird observation data include basic weather information such as temperature but not precipitation. Weather data product made available by N.O.A.A. (National Oceanic and Atmospheric Administration) includes worldwide monthly precipitation data. If one were interested in studying possible correlation between bird-count data and precipitation data of the same location, one would have to estimate and summarize biweekly bird-count data in form of monthly data, or to apply statistical techniques/models to extrapolate biweekly precipitation data from the monthly record. Furthermore, it is not uncommon that there are missing data. Missing data will need to be filled in. Sometimes data will also require statistical adjustment for purposes such as homogeneity. Yet another subtle problem related to data purification is consolidation. Consider a customer of a car rental company who uses his home and business addresses under two different accounts, the ability to know this indeed is the very same person will be advantageous if the car rental company attempts to mail the customer about, let’s say, change of company address.

If we see database design and data warehouse design as the efforts towards improving schema design and creating a structure for efficient and effective data storage, Online Analytical Processing (OLAP) is about efficient information retrieval. OLAP is meant to provide users with the ability to gain insight into their data from multidimensional views. In formulating OLAP topology and design, we often ask questions about when, what, where, and who. In the simplest term, classifying sales by the time a product was sold by a sales agent in a particular store may entail a direct mapping between the sales information and the following four dimensions:

(When?) Date of sale	→	Time(and date)
(What?) Product	→	Category
(Where?) Store	→	Branch Location
(Who?) Salesperson	→	Employee

It should be kept in mind that the dimensions of OLAP are not arbitrary. Rather, the formulation of the dimensions are tightly integrated into the data warehouse design. In the data warehouse and OLAP design, we often think about "facts" and "measure" that we want to capture. For example, we may want to capture facts about sales, but to measure the sales by products, employees, and/or stores. When we try to measure sales by products, we may define products as a dimension and within this dimension we may want to apply OLAP to look into product sales in different categories/levels (e.g., a wine in a category of import from Europe, or a wine in a category of import from Italy). To measure sales by products, we may define a "fact" table with aggregated information derived/calculated from support tables. These support tables are dimension tables, and are referenced by the fact table via foreign key reference. The term star transformation mentioned earlier refers to the process of constructing the fact table and its surrounding dimension tables to support OLAP.

Data mining is a discovery process for uncovering patterns and trends in the data. An important goal of data mining is to help users to understand the structures and interrelationships among different dimensions of the data, and subsequently develop (predictive) models that may help for strategic planning. For example, in the database level one may be interested in a report to answer questions such as "How many students do we have in our CS graduate program?" In the data warehouse and decision support level one may be interested in answering questions such as "How many students in our CS graduate program are from CS undergraduate program?" In the data mining level, one may want to be able to answer questions such as "Is the population of applicants applying to our graduate program remains the same in the last three years? Is there any trend pattern of undergraduate CS students continuing in CS graduate study in the last 3 years? In addition, is there any association pattern between those joining our graduate program and the unemployment rate of CS profession?" If we could answer these questions during the course of data mining, we would gain valuable information for strategic planning, and for targeting and reaching out to specific groups of audience that could be most interested in joining our graduate program.

Finally, it is important to realize that data mining cannot be operated successfully without a good solid support of administration in the executive level and the technical support in the database and data warehouse level. In reality, data mining projects, like any IT projects, require corporate investment in human resources and information technologies. Approval on such investment is only possible if the data mining projects help to improve business intelligence and to gain additional insight into making timely decision to realize business goal(s), to meet business requirement, and to improve operational activities. In addition, it is equally important to formulate measures for evaluation when designing a data mining project. These measures, commonly referred to as Return On Investment (ROI), are a critical component when proposing a data mining project, or requesting continuing support for an existing project. From the technical point of view, mining valuable information from data can only be as good as the quality of the data — no matter how sophisticated and advanced a data mining methodology/technique may be. For example, if the integrity and accuracy of the data cannot be verified, we cannot establish the trustworthiness of the data. Then data mining can be no better than "garbage in, garbage out". Even if we have ascertained the quality and integrity of the data, we need to keep in mind that the techniques, and the enabling technologies for data mining, are typically more diverse compared to database and data warehouse technology. Techniques for data mining could range from association rule discovery, model identification, to prediction/classification based on probabilistic inference using a model encoded with association rules. In the spectrum of available techniques, there could be different choices even just for model discovery and model-based prediction. For example, neural network models, fuzzy classifiers, and Bayesian classifiers are just a few choices for model discovery (selection) for facilitating model-based prediction in the process of data mining.

4. ARCHITECTURE, TOOLS AND APPLICATIONS

Given the scope of the data warehousing and data mining discussed here, can we have an architectural overview that summarizes the relationship between database, data warehouse, and decision support system in the data mining level? Also, are there any real world applications of data warehousing and data mining? And what are the tools available?

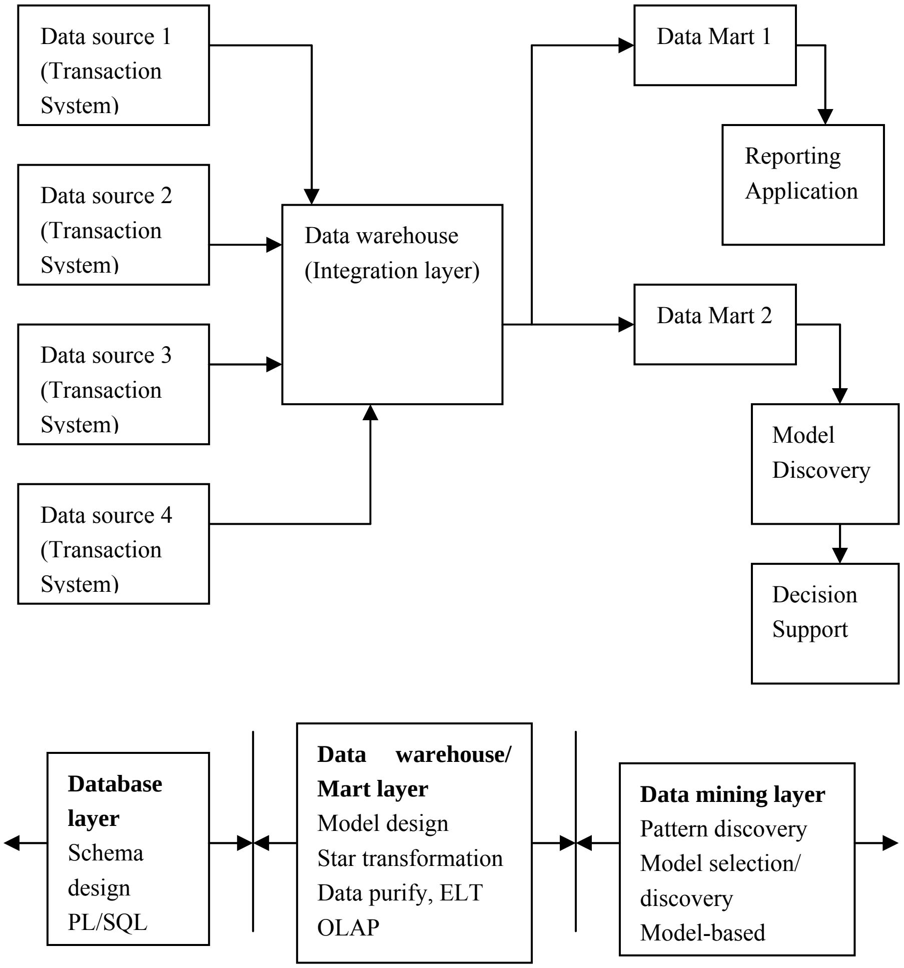
Data source 1 (Transaction System)
Data source 2 (Transaction System)
Data source 3 (Transaction System)
Data source 4 (Transaction System)
Data warehouse (Integration layer)
Data Mart 1
Reporting Application
Data Mart 2
Model Discovery
Decision Support
Database layer
Schema design
PL/SQL
Data warehouse/ Mart layer
Model design
Star transformation
Data purify, ELT
OLAP
Data mining layer
Pattern discovery
Model selection/ discovery
Model-based

Figure 1-1. Architecture overview

Are there any real world applications of data mining? The answer is definitely affirmative. American Express and AT&T analyzed and profiled their client based on KDD. BBC in UK applied data mining techniques to analyze viewing figures. Some of the well-known commercial data mining tools include Clementine (from Integral Solutions), Intelligent Miner (from IBM), 4Thought (from Livingstones), SAS data mining tool, Oracle Darwin, and Oracle data mining suite. Unfortunately, none of these tools costs less than $10,000 and some may cost up to the order of half of a million dollars.

In addition to commercial and business applications, data warehouse and data mining projects for scientific data addressing environmental issues are foreseen as a fast growing area. Affordable advanced statistical tools such as S-PLUS or I-Miner may become viable tools for relatively sophisticated users to conduct data mining in a scientific area such as one just mentioned. In our discussion, we will go into the details of building a scientific data warehouse hosting temperature, precipitation, water quality, and forest cover type data from four different independent sources. We will also illustrate how to apply data mining techniques to these data for discovering interesting patterns.

5. CONCEPTUAL/PRACTICAL MINING TOOLS

What conceptual/practical tools and techniques are we going to use in our quest for data mining?

It is not possible to cover all the techniques and tools for data mining in this book. Rather, we will focus on the conceptual tools for data mining that are based on information theory and statistics. In particular, we will introduce a concept of patterns and related techniques for data mining in the following scenario:

a) One important aspect about data mining is about discovering changes. Suppose we have made observations about a physical phenomenon, let's say, temperature, how do we know that there is a change in temperature phenomenon? And if so, what kind of change has occurred? In this example, if we do discover a change in temperature phenomenon due to, let's say, man-made environmental disturbance, is it a change towards global warming/cooling? Or do we have a change towards a more rapid rate of temperature fluctuation between seasons? We will discuss a technique referred to as change-point detection for the elicitation of information about changes from data.

b) Given a set of data representing observations made on a certain phenomenon, we would like to know the existence of association patterns that reveal the co-occurrence of events, and what are these joint events. We will discuss a concept of statistical significant pattern that reveals information about events not just co-occurring by chance. In other words, the objective of the specific data mining tasks is to discover events that their likelihood of occurrence is not insignificant while their occurrence is beyond a reasonable doubt of co-incidence.

c) Given a set of statistically significant event patterns, we would like to discover probability models that capture the statistical properties of the event patterns. Discovering such models is important in the sense that it offers one realization of probabilistic prediction rules and an inference mechanism within the framework of probability for prediction. We will discuss algorithms and techniques for model discovery and probabilistic inference, as well as available software that implements the algorithms and techniques.

Practical tools that implement the techniques mentioned above have been implemented as an ActiveX component written in C++, Java, S-PLUS programming, and MathCAD programming. Such practical tools can be found in the web site associated with this book [www http://www.techsuite.net/kluwer/].

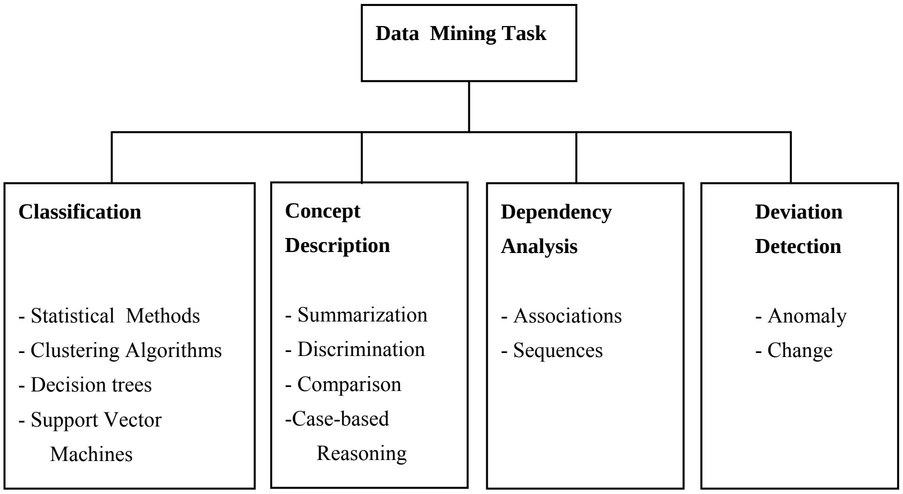

Figure 1-2. Data mining taxonomy

The diagram shown above is a possible taxonomy for data mining tasks by Mike Shaw of UIUC. Below is a brief summary of where the contribution of our work can be categorized:

1. For classification: A linkage between information theory and statistics for understanding the meaning behind the *support* and *confidence* of a rule learned via rule induction.

2. For concept description: A model discovering process based on an information-theoretic optimization approach for identifying an optimal probability model (with respect to minimum bias criterion) that summarizes the probability structure revealed by statistically significant data patterns.
3. For dependency analysis: A concept of event patterns and an information-theoretic approach for discovering event patterns with significant statistical association.
4. For deviation detection: A change point method based on Schwartz information criterion and a binary segmentation technique for identifying changes in population of a statistical parametric model (e.g., Gaussian).

6. CONCLUSION

In this chapter an overall view on data mining and data warehouse was presented. Several fundamental issues were addressed; for example, relationship between data and information, data mining and data warehouse management, architecture, and tools. A brief data mining taxonomy was also presented. Despite the contribution of the work described in this book was discussed, it is by no means covering all aspects and areas of the entire field of data mining and data warehouse. For example, case-based reasoning and support vector machines are two active topics in data mining, and are not covered in this book. Readers interested in these and related topics can refer to the publications elsewhere [Cristianini 2000][Minor 2000][Smyth 1995][Vapnik 1995].