1 Data-Mining Concepts

CHAPTER OBJECTIVES

- Understand the need for analyses of large, complex, information-rich data sets.
- Identify the goals and primary tasks of the data-mining process.
- Describe the roots of data-mining technology.
- Recognize the iterative character of a data-mining process and specify its basic steps.
- Explain the influence of data quality on a data-mining process.
- Establish the relation between data warehousing and data mining.

1.1 INTRODUCTION

Modern science and engineering are based on using *first-principle models* to describe physical, biological, and social systems. Such an approach starts with a basic scientific model, such as Newton's laws of motion or Maxwell's equations in electromagnetism, and then builds upon them various applications in mechanical engineering or electrical engineering. In this approach, experimental data are used to verify the underlying first-principle models and to estimate some of the parameters that are difficult or sometimes impossible to measure directly. However, in many domains the underlying first principles are unknown, or the systems under study are too complex to be mathematically formalized. With the growing use of computers, there is a great amount of data being generated by such systems. In the absence of first-principle models, such readily available data can be used to derive models by estimating useful relationships between a system's variables (i.e., unknown input–output dependencies). *Thus there is currently a paradigm shift from classical modeling and analyses based on first principles to developing models and the corresponding analyses directly from data.*

We have grown accustomed gradually to the fact that there are tremendous volumes of data filling our computers, networks, and lives. Government agencies, scientific institutions, and businesses have all dedicated enormous resources to collecting and storing data. In reality, only a small amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analyzed effectively. How could this happen? The primary reason is that the original effort to create a data set is often focused on issues such as storage efficiency; it does not include a plan for how the data will eventually be used and analyzed.

The need to understand large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering. In the business world, corporate and customer data are becoming recognized as a strategic asset. The

ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining.

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

In practice, the two primary goals of data mining tend to be *prediction* and *description*. *Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description*, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories:

1) Predictive data mining, which *produces the model* of the system described by the given data set, or
2) Descriptive data mining, which *produces new, nontrivial information* based on the available data set.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably. The goals of prediction and description are achieved by using data-mining techniques, explained later in this book, for the following *primary data-mining tasks:*

1. *Classification* – discovery of a predictive learning function that classifies a data item into one of several predefined classes.
2. *Regression* – discovery of a predictive learning function, which maps a data item to a real-value prediction variable.
3. *Clustering* – a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
4. *Summarization* – an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.
5. *Dependency Modeling* – finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
6. *Change and Deviation Detection* – discovering the most significant changes in the data set.

The more formal approach, with graphical interpretation of data-mining tasks for complex and large data sets and illustrative examples, is given in Chapter 4. Current introductory classifications and definitions are given here only to give the reader a feeling of the wide spectrum of problems and tasks that may be solved using data-mining technology.

The success of a data-mining engagement depends largely on the amount of energy, knowledge, and creativity that the designer puts into it. In essence, data mining is like solving a puzzle. The individual pieces of the puzzle are not complex structures in and of themselves. Taken as a collective whole, however, they can constitute very elaborate systems. As you try to unravel these systems, you will probably get frustrated, start forcing parts together, and generally become annoyed at the entire process; but once you know how to work with the pieces, you realize that it was not really that hard in the first place. The same analogy can be applied to data mining. In the beginning, the designers of the data-mining process probably do not know much about the data sources; if they did, they would most likely not be interested in performing data mining. Individually, the data seem simple, complete, and explainable. But collectively, they take on a whole new appearance that is intimidating and difficult to comprehend, like the puzzle. Therefore, being an analyst and designer in a data-mining process requires, besides thorough professional knowledge, creative thinking and a willingness to see problems in a different light.

Data mining is one of the fastest growing fields in the computer industry. Once a small interest area within computer science and statistics, it has quickly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets. Since data mining is a natural activity to be performed on large data sets, one of the largest target markets is the entire data warehousing, data-mart, and decision-support community, encompassing professionals from such industries as retail, manufacturing, telecommunications, healthcare, insurance, and transportation. In the business community, data mining can be used to discover new purchasing trends, plan investment strategies, and detect unauthorized expenditures in the accounting system. It can improve marketing campaigns and the outcomes can be used to provide customers with more focused support and attention. Data-mining techniques can be applied to problems of business process reengineering, in which the goal is to understand interactions and relationships among business practices and organizations.

Many law enforcement and special investigative units, whose mission is to identify fraudulent activities and discover crime trends, have also used data mining successfully. For example, these methodologies can aid analysts in the identification of critical behavior patterns in the communication interactions of narcotics organizations, the monetary transactions of money laundering and insider trading operations, the movements of serial killers, and the targeting of smugglers at border crossings. Data-mining techniques have also been employed by people in the intelligence community who maintain many large data sources as a part of the activities relating to matters of national security. Appendix B of the book gives a brief overview of typical commercial applications of data-mining technology today.

1.2 DATA-MINING ROOTS

Looking at how different authors describe data mining, it is clear that we are far from a universal agreement on the definition of data mining or even what constitutes data mining. Is data mining a form of statistics enriched with learning theory or is it a revolutionary new concept? In our view, most data-mining problems and corresponding solutions have roots in classical data analysis. Data mining has its origins in various disciplines, of which the two most important are *statistics* and *machine learning*. Statistics has its roots in mathematics, and therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice. In contrast, the machine-learning community has its origins very much in computer practice. This has led to a practical orientation, a willingness to test something out to see how well it performs, without waiting for a formal proof of effectiveness.

If the place given to mathematics and formalizations is one of the major differences between statistical and machine-learning approaches to data mining, another is in the relative emphasis they give to models and algorithms. Modern statistics is almost entirely driven by the notion of a model. This is a postulated structure, or an approximation to a structure, which could have led to the data. In place of the statistical emphasis on models, machine learning tends to emphasize algorithms. This is hardly surprising; the very word ''learning'' contains the notion of a process, an implicit algorithm.

Basic modeling principles in data mining also have roots in *control theory*, which is primarily applied to engineering systems and industrial processes. The problem of determining a mathematical model for an unknown system (also referred to as the target system) by observing its input–output data pairs is generally referred to as system identification. The purposes of system identification are multiple and, from a standpoint of data mining, the most important are to predict a system's behavior and to explain the interaction and relationships between the variables of a system.

System identification generally involves two top-down steps:

1. *Structure identification* – In this step, we need to apply a priori knowledge about the target system to determine a class of models within which the search for the most suitable model is to be conducted. Usually this class of models is denoted by a parametrized function $y = f(u,t)$, where y is the model's output, u is an input vector, and t is a parameter vector. The determination of the function f is problem-dependent, and the function is based on the designer's experience, intuition, and the laws of nature governing the target system.
2. *Parameter identification* – In the second step, when the structure of the model is known, all we need to do is apply optimization techniques to determine parameter vector t such that the resulting model $y^* = f(u,t^*)$ can describe the system appropriately.

In general, system identification is not a one-pass process: both structure and parameter identification need to be done repeatedly until a satisfactory model is found. This iterative process is represented graphically in Figure 1.1. Typical steps in every iteration are as follows:

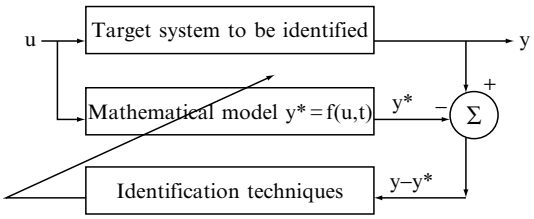

FIGURE 1.1 Block diagram for parameter identification

1. Specify and parametrize a class of formalized (mathematical) models, $y^* = f(u,t)$, representing the system to be identified.
2. Perform parameter identification to choose the parameters that best fit the available data set (the difference $y - y^*$ is minimal).
3. Conduct validation tests to see if the model identified responds correctly to an unseen data set (often referred as test, validating, or checking data set).
4. Terminate the process once the results of the validation test are satisfactory.

If we do not have any *a priori* knowledge about the target system, then structure identification becomes difficult, and we have to select the structure by trial and error. While we know a great deal about the structures of most engineering systems and industrial processes, in a vast majority of target systems where we apply data-mining techniques, these structures are totally unknown, or they are so complex that it is impossible to obtain an adequate mathematical model. Therefore, new techniques were developed for parameter identification and they are today a part of the spectra of data-mining techniques.

Finally, we can distinguish between how the terms ''model'' and ''pattern'' are interpreted in data mining. A model is a ''large scale'' structure, perhaps summarizing relationships over many (sometimes all) cases, whereas a pattern is a local structure, satisfied by few cases or in a small region of a data space. It is also worth noting here that the word ''pattern'', as it is used in pattern recognition, has a rather different meaning for data mining. In pattern recognition it refers to the vector of measurements characterizing a particular object, which is a point in a multidimensional data space. In data mining, a pattern is simply a local model. In this book we refer to n-dimensional vectors of data as *samples*.

1.3 DATA-MINING PROCESS

Without trying to cover all possible approaches and all different views about data mining as a discipline, let us start with one possible, sufficiently broad definition of data mining:

DEF: Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data.

The word "process" is very important here. Even in some professional environments there is a belief that data mining simply consists of picking and applying a computer-based tool to match the presented problem and automatically obtaining a solution. This is a misconception based on an artificial idealization of the world. There are several reasons why this is incorrect. One reason is that data mining is not simply a collection of isolated tools, each completely different from the other, and waiting to be matched to the problem. A second reason lies in the notion of matching a problem to a technique. Only very rarely is a research question stated sufficiently precisely that a single and simple application of the method will suffice. In fact, what happens in practice is that data mining becomes an iterative process. One studies the data, examines it using some analytic technique, decides to look at it another way, perhaps modifying it, and then goes back to the beginning and applies another data-analysis tool, reaching either better or different results. This can go round and round many times; each technique is used to probe slightly different aspects of data—to ask a slightly different question of the data. What is essentially being described here is a voyage of discovery that makes modern data mining exciting. Still, data mining is not a random application of statistical, machine learning, and other methods and tools. It is not a random walk through the space of analytic techniques but a carefully planned and considered process of deciding what will be most useful, promising, and revealing.

It is important to realize that the problem of discovering or estimating dependencies from data or discovering totally new data is only one part of the general experimental procedure used by scientists, engineers, and others who apply standard steps to draw conclusions from the data. The general experimental procedure adapted to data-mining problems involves the following steps:

1. State the problem and formulate the hypothesis

Most data-based modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initial phase; it continues during the entire data-mining process.

2. Collect the data

This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a *designed experiment*. The second possibility is when the expert cannot influence the data-

generation process: this is known as the *observational approach*. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results.

3. Preprocessing the data

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

1. *Outlier detection (and removal)* – Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such nonrepresentative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

   a) Detect and eventually remove outliers as a part of the preprocessing phase, or
   b) Develop robust modeling methods that are insensitive to outliers.

2. *Scaling, encoding, and selecting features* – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process.

Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding. More about these techniques and the preprocessing phase in general will be given in Chapters 2 and 3, where we have functionally divided preprocessing and its corresponding techniques into two subphases: data preparation and data-dimensionality reduction.

4. Estimate the model

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task. The basic principles of learning and discovery from data are given in Chapter 4 of this book. Later, Chapters 5 through 13 explain and analyze specific techniques that are applied to perform a successful learning process from data and to develop an appropriate model.

5. Interpret the model and draw conclusions

In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high-dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision-making.

Even though the focus of this book is on steps 3 and 4 in the data-mining process, we have to understand that they are just two steps in a more complex process. All phases, separately, and the entire data-mining process, as a whole, are highly iterative, as has been shown in Figure 1.2. A good understanding of the whole process is important for any successful application. No matter how powerful the data-mining method used in step 4 is, the resulting model will not be valid if the data are not collected and preprocessed correctly, or if the problem formulation is not meaningful.

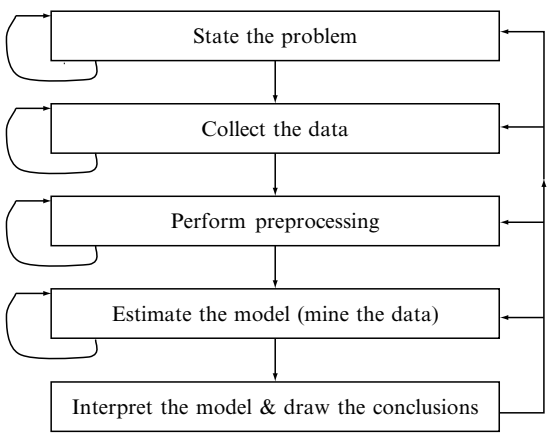

FIGURE 1.2 The data-mining process

1.4 LARGE DATA SETS

As we enter into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive *data sets*, as we call large data, is far behind our ability to gather and store the data. Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge business databases. Scientists are at the higher end of today's data-collection machinery, using data from different sources—from remote-sensing platforms to microscope probing of cell details. Scientific instruments can easily generate terabytes of data in a short period of time and store them in the computer. The information age, with the expansion of the Internet, has caused an exponential growth in information sources and also in information-storage units. An illustrative example is given in Figure 1.3, where we can see a dramatic increase of Internet hosts in just the last three years, where these numbers are directly proportional to the amount of data stored on the Internet.

There is a rapidly widening gap between data-collection and data-organization capabilities and the ability to analyze the data. Current hardware and database technology allows efficient, inexpensive, and reliable data storage and access. However, whether the context is business, medicine, science, or government, the data sets themselves, in their raw form, are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer goods company may yield knowledge of the correlation between sales of certain items and certain demographic groupings. This knowledge can be used to introduce new, targeted marketing campaigns with a predictable financial return, as opposed to unfocused campaigns.

The root of the problem is that the data size and dimensionality are too large for manual analysis and interpretation, or even for some semiautomatic computer-based analyses. A scientist or a business manager can work effectively with a few hundred or thousand records. Effectively mining millions of data points, each described with tens or hundreds of characteristics, is another matter. Imagine the analysis of terabytes of sky-image data with thousands of photographic high-resolution images (23,040 $\times$ 23,040 pixels per image), or human genome databases with billions of components. In theory, "big data" can lead to much stronger conclusions,

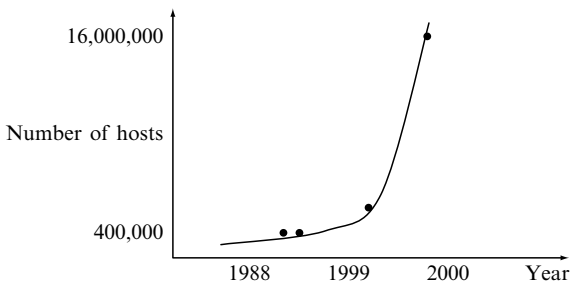

FIGURE 1.3 Growth of Internet hosts

but in practice many difficulties arise. The business community is well aware of today's information overload, and one analysis shows that

1. 61% of managers believe that information overload is present in their own workplace,
2. 80% believe the situation will get worse,
3. over 50% of the managers ignore data in current decision-making processes because of the information overload,
4. 84% of managers store this information for the future; it is not used for current analysis,
5. 60% believe that the cost of gathering information outweighs its value.

What are the solutions? Work harder. Yes, but how long can you keep up, because the limits are very close. Employ an assistant. Maybe, if you can afford it. Ignore the data. But then you are not competitive in the market. The only real solution will be to replace classical data analysis and interpretation methodologies (both manual and computer-based) with a new data-mining technology.

In theory, most data-mining methods should be happy with large data sets. Large data sets have the potential to yield more valuable information. If data mining is a search through a space of possibilities, then large data sets suggest many more possibilities to enumerate and evaluate. The potential for increased enumeration and search is counterbalanced by practical limitations. Besides the computational complexity of the data-mining algorithms that work with large data sets, a more exhaustive search may also increase the risk of finding some low-probability solutions that evaluate well for the given data set, but may not meet future expectations.

In today's multimedia-based environment that has a huge Internet infrastructure, different types of data are generated and digitally stored. To prepare adequate data-mining methods, we have to analyze the basic types and characteristics of datasets. The first step in this analysis is systematization of data with respect to their computer representation and use. Data that is usually the source for a data-mining process can be classified into structured data, semi-structured data, and unstructured data.

Most business databases contain structured data consisting of well-defined fields with numeric or alphanumeric values, while scientific databases may contain all three classes. Examples of semi-structured data are electronic images of business documents, medical reports, executive summaries, and repair manuals. The majority of web documents also fall in this category. An example of unstructured data is a video recorded by a surveillance camera in a department store. Such visual and, in general, multimedia recordings of events or processes of interest are currently gaining widespread popularity because of reduced hardware costs. This form of data generally requires extensive processing to extract and structure the information contained in it.

Structured data is often referred to as traditional data, while the semi-structured and unstructured data are lumped together as nontraditional data (also called multimedia data). Most of the current data-mining methods and commercial tools are applied to traditional data. However, the development of data-mining tools for

nontraditional data, as well as interfaces for its transformation into structured formats, is progressing at a rapid rate.

The standard model of structured data for data mining is a collection of cases. Potential measurements called features are specified, and these features are uniformly measured over many cases. Usually the representation of structured data for data-mining problems is in a tabular form, or in the form of a single relation (term used in relational databases), where columns are features of objects stored in a table and rows are values of these features for specific entities. A simplified graphical representation of a data set and its characteristics is given in Figure 1.4. In the data-mining literature, we usually use the terms samples or cases for rows. Many different types of features (attributes or variables)—i.e., fields—in structured data records are common in data mining. Not all of the data-mining methods are equally good at dealing with different types of features.

There are several ways of characterizing features. One way of looking at a feature—or in a formalization process, the more often-used term; variable—is to see whether it is an *independent variable* or a *dependent variable*; i.e, whether or not it is a variable whose values depend upon values of other variables represented in a data set. This is a model-based approach to classifying variables. All dependent variables are accepted as outputs from the system for which we are establishing a model, and independent variables are inputs to the system, as represented in Figure 1.5.

There are some additional variables that influence system behavior, but the corresponding values are not available in a data set during a modeling process. The reasons are different: from high complexity and the cost of measurements for these features to a modeler's not understanding the importance of some factors and their influences on the model. These are usually called unobserved variables, and they are the main cause of ambiguities and estimations in a model.

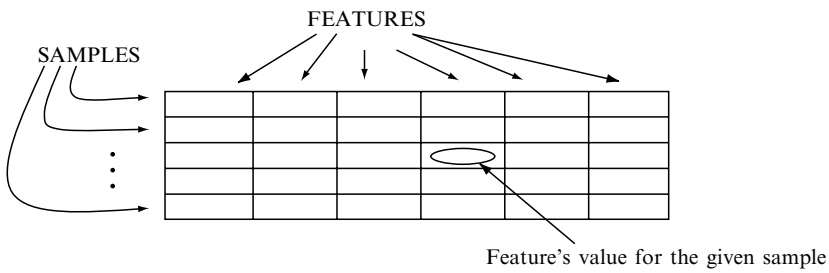

FIGURE 1.4 Tabular representation of a data set

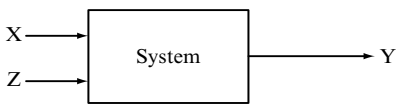

FIGURE 1.5 A real system, besides input (independent) variables X and (dependent) outputs Y, often has unobserved inputs Z

Today's computers and corresponding software tools support the processing of data sets with millions of samples and hundreds of features. Large data sets, including those with mixed data types, are a typical initial environment for application of data-mining techniques. When a large amount of data is stored in a computer, one cannot rush into data-mining techniques, because the important problem of data quality has first to be resolved. Also, it is obvious that a manual quality analysis is not possible at that stage. Therefore, it is necessary to prepare a data-quality analysis in the earliest phases of the data-mining process; usually it is a task to be undertaken in the data-preprocessing phase. The quality of data has a profound effect on the image of the system and determines the corresponding model that is implicitly described; it could also limit the ability of end users to make informed decisions. Using the available data-mining techniques, it will be difficult to undertake major qualitative changes in an organization if the data is of a poor quality; similarly, to make new sound discoveries from poor quality scientific data will be almost impossible. There are a number of indicators of data quality:

1. The data should be accurate. The analyst has to check that the name is spelled correctly, the code is in a given range, the value is complete, etc.
2. The data should be stored according to data type. The analyst must ensure that the numeric value is not presented in character form, that integers are not in the form of real numbers, etc.
3. The data should have integrity. Updates should not be lost because of conflicts among different users; robust backup and recovery procedures should be implemented if they are not already part of the Data Base Management System (DBMS).
4. The data should be consistent. The form and the content should be the same after integration of large data sets from different sources.
5. The data should not be redundant. In practice, redundant data should be minimized and reasoned duplication should be controlled. Duplicated records should be eliminated.
6. The data should be timely. The time component of data should be recognized explicitly from the data or implicitly from the manner of its organization.
7. The data should be well understood. Naming standards are a necessary but not the only condition for data to be well understood. The user should know that the data corresponds to an established domain.
8. The data set should be complete. Missing data, which occurs in reality, should be minimized. Missing data could reduce the quality of a global model. On the other hand, some data-mining techniques are robust enough to support analyses of data sets with missing values.

How to work with and solve some of these problems of data quality is explained in greater detail in Chapters 2 and 3 where basic data-mining preprocessing methodologies are introduced. These processes are performed very often using data-warehousing technology, briefly explained in Section 1.5.

1.5 DATA WAREHOUSES

Although the existence of a data warehouse is not a prerequisite for data mining, in practice, the task of data mining, especially for some large companies, is made a lot easier by having access to a data warehouse. A primary goal of a data warehouse is to increase the ''intelligence'' of a decision process and the knowledge of the people involved in this process. For example, the ability of product marketing executives to look at multiple dimensions of a product's sales performance—by region, by type of sales, by customer demographics—may enable better promotional efforts, increased production, or new decisions in product inventory and distribution. It should be noted that average companies work with averages. The superstars differentiate themselves by paying attention to the details. They may need to slice and dice the data in different ways to obtain a deeper understanding of their organization and to make possible improvements. To undertake these processes, users have to know what data exists, where it is located, and how to access it.

A data warehouse means different things to different people. Some definitions are limited to data; others refer to people, processes, software, tools, and data. One of the global definitions is that

> the data warehouse is a collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time.

Based on this definition, a data warehouse can be viewed as an organization's repository of data, set up to support strategic decision-making. The function of the data warehouse is to store the historical data of an organization in an integrated manner that reflects the various facets of the organization and business. The data in a warehouse are never updated but used only to respond to queries from end users who are generally decision-makers. Typically, data warehouses are huge, storing billions of records. In many instances, an organization may have several local or departmental data warehouses often called data marts. A data mart is a data warehouse that has been designed to meet the needs of a specific group of users. It may be large or small, depending on the subject area.

At this early time in the evolution of data warehouses, it is not surprising to find many projects floundering because of the basic misunderstanding of what a data warehouse is. What *does* surprise is the size and scale of these projects. Many companies err by not defining exactly what a data warehouse is, the business problems it will solve, and the uses to which it will be put. Two aspects of a data warehouse are most important for a better understanding of its design process: the first is the specific types (classification) of data stored in a data warehouse, and the second is the set of transformations used to prepare the data in the final form such that it is useful for decision making. A data warehouse includes the following categories of data, where the classification is accommodated to the time-dependent data sources:

1. Old detail data
2. Current (new) detail data

3. Lightly summarized data
4. Highly summarized data
5. Metadata (the data directory or guide).

To prepare these five types of elementary or derived data in a data warehouse, the fundamental types of data transformation are standardized. There are four main types of transformations, and each has its own characteristics:

1. *Simple transformations* – These transformations are the building blocks of all other more complex transformations. This category includes manipulation of data that is focused on one field at a time, without taking into account its values in related fields. Examples include changing the data type of a field or replacing an encoded field value with a decoded value.
2. *Cleansing and scrubbing* – These transformations ensure consistent formatting and usage of a field, or of related groups of fields. This can include a proper formatting of address information, for example. This class of transformations also includes checks for valid values in a particular field, usually checking the range or choosing from an enumerated list.
3. *Integration* – This is a process of taking operational data from one or more sources and mapping it, field by field, onto a new data structure in the data warehouse. The common identifier problem is one of the most difficult integration issues in building a data warehouse. Essentially, this situation occurs when there are multiple system sources for the same entities and there is no clear way to identify those entities as the same. This is a challenging problem, and in many cases it cannot be solved in an automated fashion. It frequently requires sophisticated algorithms to pair up probable matches. Another complex data-integration scenario occurs when there are multiple sources for the same data element. In reality, it is common that some of these values are contradictory, and resolving a conflict is not a straightforward process. Just as difficult as having conflicting values is having no value for a data element in a warehouse. All these problems and corresponding automatic or semiautomatic solutions are always domain-dependent.
4. *Aggregation and summarization* – These are methods of condensing instances of data found in the operational environment into fewer instances in the warehouse environment. Although the terms aggregation and summarization are often used interchangeably in the literature, we believe that they do have slightly different meanings in the data-warehouse context. Summarization is a simple addition of values along one or more data dimensions; e.g., adding up daily sales to produce monthly sales. Aggregation refers to the addition of different business elements into a common total; it is highly domain-dependent. For example, aggregation is adding daily product sales and monthly consulting sales to get the combined, monthly total.

These transformations are the main reason why we prefer a warehouse as a source of data for a data-mining process. If the data warehouse is available, the preprocessing phase in data mining is significantly reduced, sometimes even eliminated. Do not forget that this preparation of data is the most time-consuming phase.

Although the implementation of a data warehouse is a complex task, described in many texts in great detail, in this text we are giving only the basic characteristics. A three-stage data-warehousing development process is summarized through the following basic steps:

1. *Modeling* – In simple terms, to take the time to understand business processes, the information requirements of these processes, and the decisions that are currently made within processes.
2. *Building* – To establish requirements for tools that suit the types of decision support necessary for the targeted business process; to create a data model that helps further define information requirements; to decompose problems into data specifications and the actual data store, which will, in its final form, represent either a data mart or a more comprehensive data warehouse.
3. *Deploying* – To implement, relatively early in the overall process, the nature of the data to be warehoused and the various business intelligence tools to be employed; to begin by training users. The deploy stage explicitly contains a time during which users explore both the repository (to understand data that are and should be available) and early versions of the actual data warehouse. This can lead to an evolution of the data warehouse, which involves adding more data, extending historical periods, or returning to the build stage to expand the scope of the data warehouse through a data model.

Data mining represents one of the major applications for data warehousing, since the sole function of a data warehouse is to provide information to end users for decision support. Unlike other query tools and application systems, the data-mining process provides an end-user with the capacity to extract hidden, nontrivial information. Such information, although more difficult to extract, can provide bigger business and scientific advantages and yield higher returns on "data warehousing and data mining" investments.

How is data mining different from other typical applications of a data warehouse, such as structured query languages (SQL) and on-line analytical processing tools (OLAP), which are also applied to data warehouses? SQL is a standard relational database language that is good for queries that impose some kind of constraints on data in the database in order to extract an answer. In contrast, data-mining methods are good for queries that are exploratory in nature, trying to extract hidden, not so obvious information. SQL is useful when we know exactly what we are looking for and we can describe it formally. We will use data-mining methods when we know only vaguely what we are looking for. Therefore, these two classes of data-warehousing applications are complementary.

OLAP tools and methods have become very popular in recent years as they let users analyze data in a warehouse by providing multiple views of the data, supported by advanced graphical representations. In these views, different dimensions of data correspond to different business characteristics. OLAP tools make it very easy to look at dimensional data from any angle or to slice-and-dice it. Although OLAP tools, like data-mining tools, provide answers that are derived from data, the similarity between them ends here. The derivation of answers from data in OLAP is analogous to calculations in a spreadsheet; because they use simple and

given-in-advance calculations, OLAP tools do not learn from data, nor do they create new knowledge. They are usually special-purpose visualization tools that can help end-users draw their own conclusions and decisions, based on graphically condensed data. OLAP tools are very useful for the data-mining process; they can be a part of it but they are not a substitute.

1.6 ORGANIZATION OF THIS BOOK

After introducing the basic concepts of data mining in Chapter 1, the rest of the book follows the basic phases of a data-mining process. In Chapters 2 and 3 are explained common characteristics of raw, large, data sets, and the typical techniques of data preprocessing. The text emphasizes the importance and influence of these initial phases on the final success and quality of data-mining results. Chapter 2 provides basic techniques for transforming raw data, including data sets with missing values and with time-dependent attributes. Outlier analysis is a set of important techniques for preprocessing of messy data, and is also explained in this chapter. Chapter 3 deals with reduction of large data sets and introduces efficient methods for reduction of features, values, and cases. When the data set is preprocessed and prepared for mining, a wide spectrum of data-mining techniques is available, and the selection of a technique or techniques depends on the type of application and the data characteristics. In Chapter 4, before introducing particular data-mining methods, we present the general theoretical background and formalizations applicable for all mining techniques. The essentials of the theory can be summarized with the question: How can one learn from data? The emphasis in Chapter 4 is on statistical learning theory and the different types of learning methods and learning tasks that may be derived from the theory.

Chapters 5 to 12 give an overview of common classes of data-mining techniques. Selected statistical inference methods are presented in Chapter 5, including Bayesian classifier, predictive and logistic regression, ANOVA analysis, and log-linear models. Chapter 6 explains the complexity of clustering problems and introduces agglomerative, partitional, and incremental clustering techniques. Chapter 7 summarizes the basic characteristics of the C4.5 algorithm as a representative of logic-based techniques for classification problems. Different aspects of local modeling in large data sets are addressed in Chapter 8, and common techniques of association-rules mining, Web mining, and text mining are presented. Chapter 9 discusses the basic components of artificial neural networks and introduces two classes: multilayer perceptrons and competitive networks as illustrative representatives of a neural-network technology. Most of the techniques explained in Chapters 10 and 11, about genetic algorithms and fuzzy systems, are not directly applicable in mining large data sets. The author believes that these technologies, derived from soft computing, become more important, perhaps not as separate techniques for data mining but as methodologies combined with other techniques, in better representing and computing with data. Finally, Chapter 12 recognizes the importance of data-mining visualization techniques, especially those for representation of large-dimensional samples.

It is our hope that we have succeeded in producing an informative and readable text supplemented with relevant examples and illustrations. All chapters in the

book have a set of review problems and reading lists. The author is preparing a solutions manual for instructors, who might use the book for undergraduate or graduate classes. For an in-depth understanding of the various topics covered in this book, we recommend to the reader a fairly comprehensive list of references, given at the end of each chapter. Although many of these references are from various journals, magazines, conference and workshop proceedings, it is obvious that during the last few years there are many more books available, covering different aspects of data mining and knowledge discovery. Finally, the book has two appendices with useful background information for practical applications of data-mining technology. In Appendix A we provide an overview of commercial and publicly available data-mining tools, and in Appendix B an extensive list of important Web sites and data-mining vendors has been included.

The reader should have some knowledge of the basic concepts and terminology associated with data structures and databases. In addition, some background in elementary statistics and machine learning may also be useful, but it is not necessarily required, as the concepts and techniques discussed within the book can be utilized without knowledge of the underlying theory.

1.7 REVIEW QUESTIONS AND PROBLEMS

1. Explain why it is not possible to analyze some large data sets using classical modeling techniques.
2. Do you recognize in your business or academic environment some problems in which the solution can be obtained through classification, regression, or deviation? Explain with examples.
3. Explain the differences between statistical and machine-learning approaches to the analysis of large data sets.
4. Why are preprocessing and dimensionality reduction important phases in successful data-mining applications?
5. Give examples of data where the time component may be recognized explicitly, and other data where the time component is given implicitly in a data organization.
6. Why is it important that the data miner understand data well?
7. Give examples of structured, semi-structured, and unstructured data from everyday situations.
8. Can a set with 50,000 samples be called a large data set? Explain your answer.
9. Enumerate the tasks that a data warehouse may solve as a part of the data-mining process.
10. Many authors include OLAP tools as a standard data-mining tool. Give the arguments for and against this classification.

1.8 REFERENCES FOR FURTHER STUDY

1. Berson, A., S. Smith, K. Thearling, *Building Data Mining Applications for CRM*, McGraw-Hill, New York, 2000.

   The book is written primarily for the business community, explaining the competitive advantage of data-mining technology. It bridges the gap between understanding this vital technology and implementing it to meet a corporation's specific needs. Basic phases in a data-mining process are explained through real-world examples.

2. Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2000.

   This book gives a sound understanding of data-mining principles. The primary orientation of the book is for database practitioners and professionals, with emphasis on OLAP and data warehousing. In-depth analysis of association rules and clustering algorithms is an additional strength of the book. All algorithms are presented in easily understood pseudocode and they are suitable for use in real-world, large-scale data-mining projects, including advanced applications such as Web mining and text mining.

3. Hand, D., H. Mannila, P. Smith, *Principles of Data Mining*, MIT Press, Cambridge: MA, 2001.

   The book consists of three sections. The first, foundations, provides a tutorial overview of the principles underlying data-mining algorithms and their applications. The second section, data-mining algorithms, shows how algorithms are constructed to solve specific problems in a principled manner. The third section shows how all of the preceding analyses fit together when applied to real-world data-mining problems.

4. Westphal, C. and T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, John Wiley, New York, 1998.

   This introductory book gives a refreshing "out-of-the-box" approach to data mining that will help the reader to maximize time and problem-solving resources, and prepare for the next wave of data-mining visualization techniques. An extensive coverage of data-mining software tools is valuable to readers who are planning to set up their own data-mining environment.