

4.4 DATA MINING PROCESS

In order to systematically carry out data mining projects, a general process is usually followed. Based on best practices, data mining researchers and practitioners have proposed several processes (workflows or simple step-by-step approaches) to maximize the chances of success in conducting data mining projects. These efforts have led to several standardized processes, some of which (a few of the most popular ones) are described in this section.

One such standardized process, arguably the most popular one, Cross-Industry Standard Process for Data Mining—**CRISP-DM**—was proposed in the mid-1990s by a European consortium of companies to serve as a nonproprietary standard methodology for data mining (CRISP-DM, 2013). Figure 4.4 illustrates this proposed process, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfied the specific business need. Even though these steps are sequential in nature, there is usually a great deal of backtracking. Because the data mining is driven by experience and experimentation, depending on the problem situation and the knowledge/experience of the analyst, the whole process can be very iterative (i.e., one should expect to go back and forth through the steps quite a few times) and time-consuming. Because later steps are built on the outcomes of the former ones, one should pay extra attention to the earlier steps in order not to put the whole study on an incorrect path from the onset.

Step 1: Business Understanding

The key element of any data mining study is to know what the study is for. Answering such a question begins with a thorough understanding of the managerial need for new knowledge and an explicit specification of the business objective regarding the study to

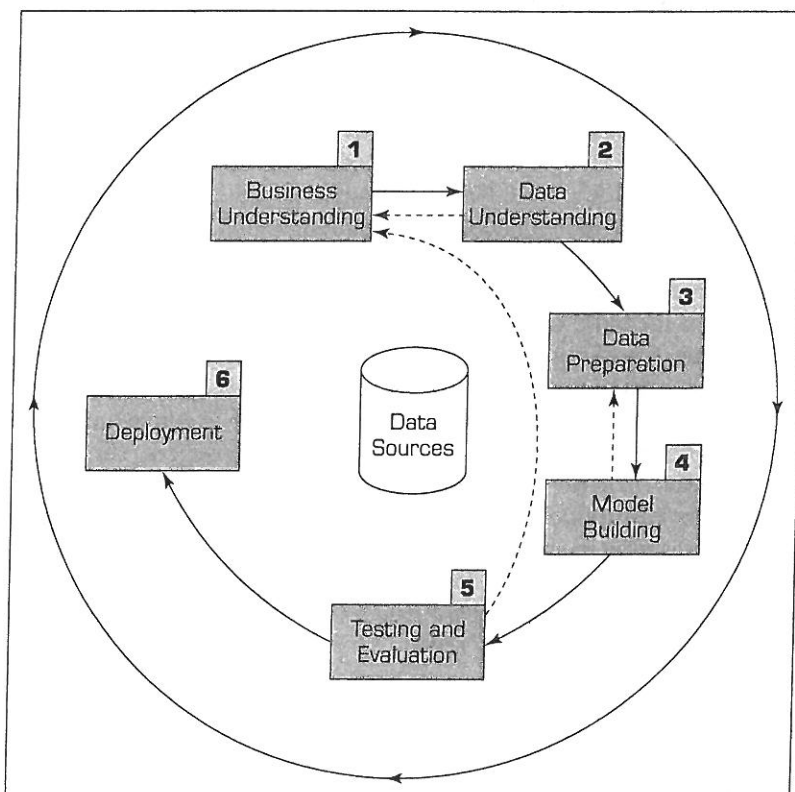


FIGURE 4.4 The Six-Step CRISP-DM Data Mining Process.

be conducted. Specific goals such as “What are the common characteristics of the customers we have lost to our competitors recently?” or “What are typical profiles of our customers, and how much value does each of them provide to us?” are needed. Then a project plan for finding such knowledge is developed that specifies the people responsible for collecting the data, analyzing the data, and reporting the findings. At this early stage, a budget to support the study should also be established, at least at a high level with rough numbers.

Step 2: Data Understanding

A data mining study is specific to addressing a well-defined business task, and different business tasks require different sets of data. Following the business understanding, the main activity of the data mining process is to identify the relevant data from many available databases. Some key points must be considered in the data identification and selection phase. First and foremost, the analyst should be clear and concise about the description of the data mining task so that the most relevant data can be identified. For example, a retail data mining project may seek to identify spending behaviors of female shoppers who purchase seasonal clothes based on their demographics, credit card transactions, and socioeconomic attributes. Furthermore, the analyst should build an intimate understanding of the data sources (e.g., where the relevant data are stored and in what form; what the process of collecting the data is—automated versus manual; who the collectors of the data are and how often the data are updated) and the variables (e.g., What are the most relevant variables? Are there any synonymous and/or homonymous variables? Are the variables independent of each other—do they stand as a complete information source without overlapping or conflicting information?).

In order to better understand the data, the analyst often uses a variety of statistical and graphical techniques, such as simple statistical summaries of each variable (e.g., for numeric variables the average, minimum/maximum, median, and standard deviation are among the calculated measures, whereas for categorical variables the mode and frequency tables are calculated), correlation analysis, scatterplots, histograms, and box plots. A careful identification and selection of data sources and the most relevant variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns.

Data sources for data selection can vary. Normally, data sources for business applications include demographic data (such as income, education, number of households, and age), sociographic data (such as hobby, club membership, and entertainment), transactional data (sales record, credit card spending, issued checks), and so on.

Data can be categorized as quantitative and qualitative. Quantitative data is measured using numeric values. It can be discrete (such as integers) or continuous (such as real numbers). Qualitative data, also known as categorical data, contains both nominal and ordinal data. Nominal data has finite nonordered values (e.g., gender data, which has two values: male and female). Ordinal data has finite ordered values. For example, customer credit ratings are considered ordinal data because the ratings can be excellent, fair, and bad.

Quantitative data can be readily represented by some sort of probability distribution. A probability distribution describes how the data is dispersed and shaped. For instance, normally distributed data is symmetric and is commonly referred to as being a bell-shaped curve. Qualitative data may be coded to numbers and then described by frequency distributions. Once the relevant data are selected according to the data mining business objective, data preprocessing should be pursued.

Step 3: Data Preparation

The purpose of data preparation (more commonly called *data preprocessing*) is to take the data identified in the previous step and prepare it for analysis by data mining methods. Compared to the other steps in CRISP-DM, data preprocessing consumes the most time

and effort; most believe that this step accounts for roughly 80 percent of the total time spent on a data mining project. The reason for such an enormous effort spent on this step is the fact that real-world data is generally incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), and inconsistent (containing discrepancies in codes or names). Figure 4.5 shows the four main steps needed to convert the raw real-world data into minable data sets.

In the first phase of data preprocessing, the relevant data is collected from the identified sources (accomplished in the previous step—Data Understanding—of the CRISP-DM process), the necessary records and variables are selected (based on an intimate understanding of the data the unnecessary sections are filtered out), and the records coming from multiple data sources are integrated (again, using the intimate understanding of the data, the synonyms and homonyms are to be handled properly).

In the second phase of data preprocessing, the data is cleaned (this step is also known as data scrubbing). In this step, the values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set (e.g., the *household income* field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out. Additionally, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

In the third phase of data preprocessing, the data is transformed for better processing. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables in order to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables (such as *number*

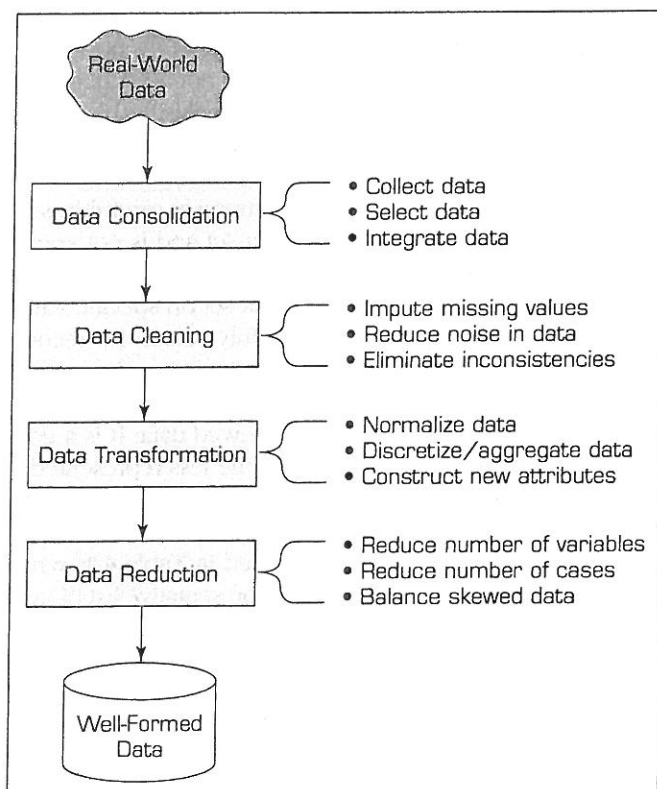


FIGURE 4.5 Data Preprocessing Steps.

of *dependents* or *years in service*, which may potentially be more important) having smaller values. Another transformation that takes place is discretization and/or aggregation. In some cases, the numeric variables are converted to categorical values (e.g., low, medium, high); in other cases a nominal variable's unique value range is reduced to a smaller set using concept hierarchies (e.g., as opposed to using the individual states with 50 different values, one may choose to use several regions for a variable that shows location) in order to have a data set that is more amenable to computer processing. Still, in other cases one might choose to create new variables based on the existing ones in order to magnify the information found in a collection of variables in the data set. For instance, in an organ transplantation data set one might choose to use a single variable showing the blood-type match (1: match, 0: no-match) as opposed to separate multinominal values for the blood type of both the donor and the recipient. Such simplification may increase the information content while reducing the complexity of the relationships in the data.

The final phase of data preprocessing is data reduction. Even though data miners like to have large data sets, too much data is also a problem. In the simplest sense, one can visualize the data commonly used in data mining projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (e.g., image processing and genome projects with complex microarray data), the number of variables can be rather large, and the analyst must reduce the number down to a manageable size. Because the variables are treated as different dimensions that describe the phenomenon from different perspectives, in data mining this process is commonly called *dimensional reduction*. Even though there is not a single best way to accomplish this task, one can use the findings from previously published literature; consult domain experts; run appropriate statistical tests (e.g., principal component analysis or independent component analysis); and, more preferably, use a combination of these techniques to successfully reduce the dimensions in the data into a more manageable and most relevant subset.

With respect to the other dimension (i.e., the number of cases), some data sets may include millions or billions of records. Even though computing power is increasing exponentially, processing such a large number of records may not be practical or feasible. In such cases, one may need to sample a subset of the data for analysis. The underlying assumption of sampling is that the subset of the data will contain all relevant patterns of the complete data set. In a homogenous data set, such an assumption may hold well, but real-world data is hardly ever homogenous. The analyst should be extremely careful in selecting a subset of the data that reflects the essence of the complete data set and is not specific to a subgroup or subcategory. The data is usually sorted on some variable, and taking a section of the data from the top or bottom may lead to a biased data set on specific values of the indexed variable; therefore, one should always try to randomly select the records on the sample set. For skewed data, straightforward random sampling may not be sufficient, and stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) may be required. Speaking of skewed data: It is a good practice to balance the highly skewed data by either oversampling the less represented or undersampling the more represented classes. Research has shown that balanced data sets tend to produce better prediction models than unbalanced ones (Wilson and Sharda, 1994).

The essence of data preprocessing is summarized in Table 4.1, which maps the main phases (along with their problem descriptions) to a representative list of tasks and algorithms.

Step 4: Model Building

In this step, various modeling techniques are selected and applied to an already prepared data set in order to address the specific business need. The model-building step also encompasses the assessment and comparative analysis of the various models built. Because there is not a universally known *best* method or algorithm for a data mining task,

TABLE 4.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
Data cleaning	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or -1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, Chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or under-sample the more represented classes.

one should use a variety of viable model types along with a well-defined experimentation and assessment strategy to identify the “best” method for a given purpose. Even for a single method or algorithm, a number of parameters need to be calibrated to obtain optimal results. Some methods may have specific requirements on the way that the data is to be formatted; thus, stepping back to the data preparation step is often necessary. Application Case 4.4 presents a research study where a number of model types are developed and compared to each other.

Depending on the business need, the data mining task can be of a prediction (either classification or regression), an association, or a clustering type. Each of these data mining tasks can use a variety of data mining methods and algorithms. Some of these data mining methods were explained earlier in this chapter, and some of the most popular algorithms, including decision trees for classification, *k*-means for clustering, and the Apriori algorithm for association rule mining, are described later in this chapter.

Step 5: Testing and Evaluation

In step 5, the developed models are assessed and evaluated for their accuracy and generality. This step assesses the degree to which the selected model (or models) meets the business objectives and, if so, to what extent (i.e., do more models need to be developed and assessed). Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit. Even though the outcome of the developed models is expected to relate to the original business objectives, other findings that are not necessarily related to the original business objectives but that might also unveil additional information or hints for future directions often are discovered.

The testing and evaluation step is a critical and challenging task. No value is added by the data mining task until the business value obtained from discovered knowledge patterns is identified and recognized. Determining the business value from discovered knowledge patterns is somewhat similar to playing with puzzles. The extracted knowledge patterns are pieces of the puzzle that need to be put together in the context of the specific business purpose. The success of this identification operation depends on the interaction among data analysts, business analysts, and decision makers (such as business managers). Because data analysts may not have the full understanding of the data mining objectives and what they mean to the business and the business analysts and decision makers may not have the technical knowledge to interpret the results of sophisticated mathematical solutions, interaction among them is necessary. In order to properly interpret knowledge patterns, it is often necessary to use a variety of tabulation and visualization techniques (e.g., pivot tables, cross-tabulation of findings, pie charts, histograms, box plots, scatterplots).

Step 6: Deployment

Development and assessment of the models is not the end of the data mining project. Even if the purpose of the model is to have a simple exploration of the data, the knowledge gained from such exploration will need to be organized and presented in a way that the end user can understand and benefit from. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

The deployment step may also include maintenance activities for the deployed models. Because everything about the business is constantly changing, the data that reflect the business activities also are changing. Over time, the models (and the patterns embedded within them) built on the old data may become obsolete, irrelevant, or misleading.

Therefore, monitoring and maintenance of the models are important if the data mining results are to become a part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process, which may not be a trivial task for complex data mining models.

Application Case 4.4

Data Mining in Cancer Research

According to the American Cancer Society, half of all men and one-third of all women in the United States will develop cancer during their lifetimes; approximately 1.5 million new cancer cases will be diagnosed in 2013. Cancer is the second most common cause of death in the United States and in the world, exceeded only by cardiovascular disease. This year, over 500,000 Americans are expected to die of cancer—more than 1,300 people a day—accounting for nearly 1 of every 4 deaths.

Cancer is a group of diseases generally characterized by uncontrolled growth and spread of abnormal cells. If the growth and/or spread are not controlled, it can result in death. Even though the exact reasons are not known, cancer is believed to be caused by both external factors (e.g., tobacco, infectious organisms, chemicals, and radiation) and internal factors (e.g., inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors may act together or in sequence to initiate or promote carcinogenesis. Cancer is treated with surgery, radiation, chemotherapy, hormone therapy, biological therapy, and targeted therapy. Survival statistics vary greatly by cancer type and stage at diagnosis.

The 5-year relative survival rate for all cancers is improving, and decline in cancer mortality has reached 20 percent in 2013, translating into the avoidance of about 1.2 million deaths from cancer since 1991. That's more than 400 lives saved per day! The improvement in survival reflects progress in diagnosing certain cancers at an earlier stage and improvements in treatment. Further improvements are needed to prevent and treat cancer.

Even though cancer research has traditionally been clinical and biological in nature, in recent years data-driven analytic studies have become a common

complement. In medical domains where data- and analytics-driven research have been applied successfully, novel research directions have been identified to further advance the clinical and biological studies. Using various types of data, including molecular, clinical, literature-based, and clinical-trial data, along with suitable data mining tools and techniques, researchers have been able to identify novel patterns, paving the road toward a cancer-free society.

In one study, Delen (2009) used three popular data mining techniques (decision trees, artificial neural networks, and support vector machines) in conjunction with logistic regression to develop prediction models for prostate cancer survivability. The data set contained around 120,000 records and 77 variables. A k -fold cross-validation methodology was used in model building, evaluation, and comparison. The results showed that support vector models are the most accurate predictor (with a test set accuracy of 92.85%) for this domain, followed by artificial neural networks and decision trees. Furthermore, using a sensitivity-analysis-based evaluation method, the study also revealed novel patterns related to prognostic factors of prostate cancer.

In a related study, Delen et al. (2004) used two data mining algorithms (artificial neural networks and decision trees) and logistic regression to develop prediction models for breast cancer survival using a large data set (more than 200,000 cases). Using a 10-fold cross-validation method to measure the unbiased estimate of the prediction models for performance comparison purposes, the results indicated that the decision tree (C5 algorithm) was the best predictor, with 93.6 percent accuracy on the holdout sample (which was the best prediction accuracy reported in the literature); followed by

(Continued)

Application Case 4.4 (Continued)

artificial neural networks, with 91.2 percent accuracy; and logistic regression, with 89.2 percent accuracy. Further analysis of prediction models revealed prioritized importance of the prognostic factors, which can then be used as basis for further clinical and biological research studies.

These examples (among many others in the medical literature) show that advanced data mining techniques can be used to develop models that possess a high degree of predictive as well as explanatory power. Although data mining methods are capable of extracting patterns and relationships hidden deep in large and complex medical databases, without the cooperation and feedback from the medical experts their results are not of much use. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable, and novel to warrant new research directions. In short,

data mining is not meant to replace medical professionals and researchers, but to complement their invaluable efforts to provide data-driven new research directions and to ultimately save more human lives.

QUESTIONS FOR DISCUSSION

1. How can data mining be used for ultimately curing illnesses like cancer?
2. What do you think are the promises and major challenges for data miners in contributing to medical and biological research endeavors?

Sources: D. Delen, "Analysis of Cancer Data: A Data Mining Approach," *Expert Systems*, Vol. 26, No. 1, 2009, pp. 100–112; J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space," *Expert Systems with Applications*, 2009, *in press*; D. Delen, G. Walker, and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," *Artificial Intelligence in Medicine*, Vol. 34, No. 2, 2005, pp. 113–127.

Other Data Mining Standardized Processes and Methodologies

In order to be applied successfully, a data mining study must be viewed as a process that follows a standardized methodology rather than as a set of automated software tools and techniques. In addition to CRISP-DM, there is another well-known methodology developed by the SAS Institute, called SEMMA (2009). The acronym **SEMMA** stands for "sample, explore, modify, model, and assess."

Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. A pictorial representation of SEMMA is given in Figure 4.6.

By assessing the outcome of each stage in the SEMMA process, the model developer can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data; that is, as with CRISP-DM, SEMMA is driven by a highly iterative experimentation cycle. The main difference between CRISP-DM and SEMMA is that CRISP-DM takes a more comprehensive approach—including understanding of the business and the relevant data—to data mining projects, whereas SEMMA implicitly assumes that the data mining project's goals and objectives along with the appropriate data sources have been identified and understood.

Some practitioners commonly use the term **knowledge discovery in databases (KDD)** as a synonym for data mining. Fayyad et al. (1996) defined *knowledge discovery in databases* as a process of using data mining methods to find useful information and patterns in the data, as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDD process. KDD is a comprehensive process