Chapter 2
Tasks and Data Flow

2.1 Data Mining Tasks

Data mining is often defined as a set of mathematical models and data manipulation techniques that perform functions aiming at the discovery of new knowledge in databases. The functions, or *tasks*, performed by these techniques can be classified in terms of either the analytical function they entail or their implementation focus. The first classification scheme takes the point of view of the data mining analyst. In this case, the analyst would classify the tasks on the basis of the problem type as one of the following.

1. *Classification*

   In these problems, the operative is to assign each record in the database a particular class or a category label from a finite set of predefined class labels. For example, a bank would be interested in classifying each of its customers as potentially interested in a new credit card or not. All decisions involving Yes/No selection, such as classifying insurance claims according to the possibility of fraud, also belong to classification problems. Classification problems may involve three or more levels, such as "high," "medium," and "low." The main point is that the number of classes is finite. Note that there could be an implicit order relationship in the definition of the classes, such as "high," "medium," and "low."

2. *Estimation*

   These problems are focused on estimating the unknown value of a continuous variable. For example, taxation authorities might be interested in estimating the *real income* of households. The number of possible outcomes of an estimation problem is infinite by definition.

3. *Prediction*

   Prediction is the task of estimating a value in the future. Typical examples include attempting to predict stock prices, prices of commodities, and future values of air pollution indices.

4. *Clustering*

   Clustering, which is also known as *segmentation*, is the task of dividing a heterogeneous population into a number of more or less homogeneous subgroups or clusters. It is also sometimes defined as *finding islands of simplicity in the data*. Typical examples include customer and market segmentation. In dealing with very large datasets, clustering is also used as an initial analysis tool to simplify the data into smaller groups or to generate hypotheses about the data.

5. *Affinity Analysis*

   Other names for affinity analysis include *market basket analysis* and *association analysis*. It is concerned with finding *things* that usually go together. These *things* could be products, transactions, sequences of operations, or any objects stored in a database. A typical example is the analysis of the supermarket basket, where we attempt to find the likelihood of specific products being purchased together in the same basket. For example, we might be interested to know whether chicken and barbecue sauce are more likely to be purchased together than, say, chicken and canned soup.

The preceding classification scheme of data mining tasks focuses on their analytical nature. Businesses, on the other hand, define data mining in terms of the application. For example, in banking and finance one speaks about *credit scoring* and *risk analysis*, and in marketing applications, data mining is described as the tool for modeling *customer behavior*, *churn analysis*, *customer acquisition*, and *cross-selling*. We can set a simple framework for the classification of data mining tasks in terms of the *business view* by dividing the applications into the following three areas of interest.

- *Sales and marketing:* This domain includes CRM (customer relationship management) applications such as customer acquisition, cross-selling, customer service, churn and retention, product affinity, and lifetime value estimation.
- *Operations management:* This classification applies to areas such as process control, inventory management, supply chain management, financial risk analysis, and maintenance.
- *Finance:* This category includes areas such as prediction and management of cash flow, loans and mortgages, credit card issuing, and assignment of credit limits.

A second business-based view of data mining tasks could be structured by grouping applications that relate to the management of "Products," "Operations," or "Customers." These three domains cover almost all aspects of any business. For

example, customer retention is related to management of customers, and risk management belongs to operations.

The range of problems for which data mining modeling has been used outside business applications is too wide to classify into specific categories. For example, clustering methods have been used to identify star formations in astronomy, and classification models are used to select jury members. Other applications include weather prediction, clinical trials, drug discovery, genetic engineering, and social studies, to mention a few.

2.2 Data Mining Competencies

A successful implementation of data mining modeling requires competencies in three areas.

1. *Understanding the Problem Domain*

   This first requirement necessitates the full understanding of the objectives of the project, the value added by the engagement and the expected return on investment (ROI), and how the business processes is being impacted by the implementation of the data mining technology. For example, in credit card risk scoring applications, the analyst must understand the basics of credit card risk management strategies and the basics of the legal as well as the business procedures involved.

2. *Understanding the Data*

   Understanding the data is not limited to the names and descriptions of fields in the database or data warehouse, but also concerns the content of the fields, the meaning of each category, the meaning of outliers, missing values, any preprocessing that has been done on the data, and the sources of the data.

3. *Data Mining Modeling Methods and Software*

   This area of competency covers the methodology of data mining, the strengths and limitations of each data mining technique, and the modeling software. The analyst should know which technique to use, with which dataset, and when.

Although nobody is an expert in everything, to achieve good results using data mining modeling, these three areas of competency are necessary. Therefore, in large organizations, where no single individual could possess high competency in all these areas, data mining is performed by a team consisting of business domain experts, data analysts and programmers, and modelers.

Many good textbooks provide the details of the business aspect of data mining and how modeling fits in the general scheme of things. Similarly, numerous good texts are dedicated to the explanation of the different data mining algorithms and software. We will not dwell much on these two areas.

2.3 The Data Flow

Figure 2.1 depicts the typical stages of data flow. In this process, many of the steps may be repeated several times in order to fit the flow of operations within a certain data mining methodology. The process can be described as follows.

1. The data is extracted from the database or the data warehouse to the mining view. The mining view is the dataset that will be used to create the predictive models.
2. The data in the mining view is divided into three partitions for training, validation, and testing of the different models. Often, only two partitions are used: training and validation.
3. A scoring view is extracted from the database or data warehouse. The scoring view is similar to the mining view, except that it contains only the fields necessary to calculate the score by the trained and validated predictive model(s). The scoring view does not contain the value of the dependent variable.
4. One or more of the trained and tested predictive models is used to produce scores using the data of the scoring view. The scores may take the form of discrete

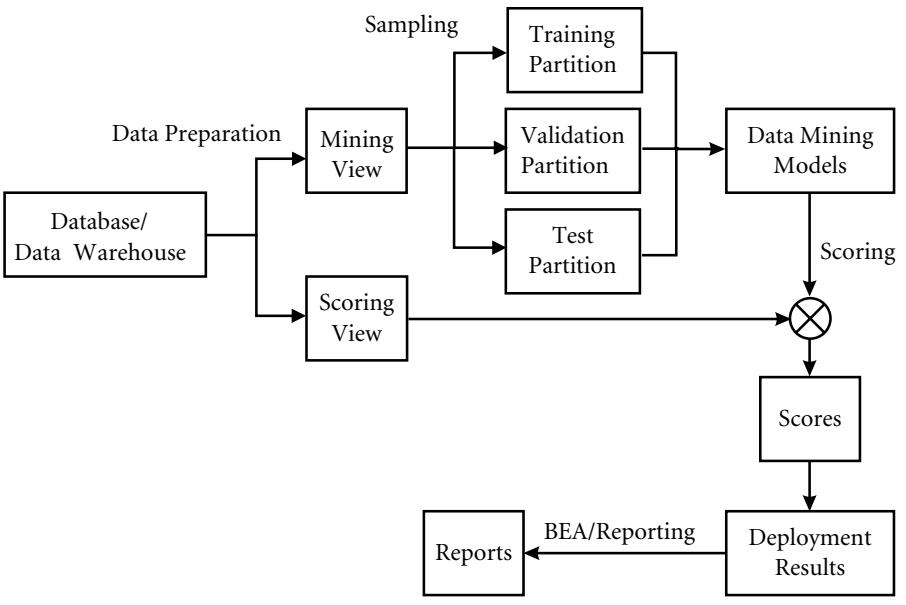

Figure 2.1 Steps of data flow.

values (1/0, Yes/No), probabilities (in the range 0–1), real values (predicted sales), or segment indexes (e.g., 1 to 9).

5. The scores, along with additional information that would be collected from the results of implementing the model (e.g., in a campaign), are then used to compile business reports.
6. After the deployment of the model and collection of the actual response or the predicted value, a Back End Analysis (BEA) is conducted to assess the model and the business process performance.

A closer examination of the preceding steps reveals that all the procedures that reshape the data can be categorized into three groups.

1. Procedures for extracting, integrating, and reshaping the mining and scoring views
2. Procedures for sampling
3. Procedures for reporting

We next discuss in more detail the types of variables and the contents of the mining and scoring views.

2.4 Types of Variables

In general, we can categorize the variables that would be included in the mining and scoring views into the following groups.

1. *Dependent Variables*

   The dependent variables (DVs) contain the quantities being estimated or predicted. In the cases of clustering and association rule models, there is no dependent variable. In all other models (classification, estimation, and prediction models), there is at least one dependent variable. Sometimes, the mining view can be used to model different objectives simultaneously (e.g., calculating the risk of loan default as well as estimating the amount to be lost in such loans).

2. *Independent Variables*

   The independent variables (IVs) are used to build the model. In the case of classification, estimation, or prediction models, these variables are the "predictors" of the dependent variable.

3. *Record Identification Variables*

   The identification (ID) variables allow the analyst to revert to the data to identify the records that show certain interesting features. For example, the ID variable in

a loan default model would allow the analyst to identify the details of data on the customers who have been flagged by the model as high-risk customers.

4. *Explanation Variables*

   The explanation variables (labels) provide extra information on each record. They are not to be used, however, in the modeling process. For example, one may find a product variable having the categories (AK1, ASU3). A descriptive variable indicating that AK1 is "long-distance package A with calling feature package K" would not add value to analysis and is too long to be used in the model, but it is useful to keep in the mining view to make understanding the data easier.

5. *Sample Weights*

   Sample weights should be included in all samples in order to allow the proper calculation of the population statistical properties, and for use in the modeling algorithms.

2.5 The Mining View and the Scoring View

The mining view is a table or, in the case of SAS, a dataset that contains the following fields.

- Identification fields: Typical examples of ID fields include customer number and transaction number.
- Independent variables: These are only *possible* predictors because one of the tasks of building a good predictive model is to identify the best predictors and use them in the model. Therefore, the mining view will contain many candidate IVs that may not be used in the development of the model.
- Dependent variables: As mentioned, with the exception of the clustering and association models, we always have one or more dependent variable in the model. In cases of classification models, it is always better to have the DV in a binary form (1/0). In spite of the fact that most classification techniques (decision trees, neural networks, logistic regression) allow the multicategory DVs, when the DV has more than two categories, the quality of the models using each output separately is better than that achieved by using all the categories simultaneously.
- Explanation variables.
- Sample weights.
- *Other reporting fields:* The mining view can include other fields that would be used for producing reports.

The scoring view is a table that contains the following fields.

- ID fields

- All the predictors that were used in the predictive model(s) to calculate the score or the predicted value, or the actual model IVs
- Description variables
- Reporting variables

Therefore, the scoring view is very similar to the mining view with the following exceptions.

- It contains only the IVs that were used in the development of the final predictive model used for scoring. Not all the other possible IVs, which were eliminated from the model, need be included.
- It does not contain the DV field. This will be the result of the scoring process.
- It usually contains many fewer fields but far more records.

2.6 Steps of Data Preparation

The typical steps involved in the preparation of the mining view include the following.

1. Extracting and sampling data from the operational database or the data warehouse; at this stage, the data may be spread among several datasets.
2. Checking the integrity of the extracted and sampled data: These checks may be performed on the individual datasets or after the integration of all data elements into one dataset, which is the next step.
3. Integrating data: In this step, the different elements of the mining view are integrated into one table or view.
4. Transforming the data and creating new variables (analytical variables): This operation can also be performed on each table extracted from the data warehouse before the integration of the data. This can sometimes result in improved model performance.
5. Removing independent variables that have low or no predictive power.

This order of steps may be altered to improve performance and to accommodate the specific needs of the data warehouse or the data sources.

The preparation of the scoring view is always much easier than that for the mining view. However, because the number of records in the scoring view is usually much larger than that in the mining view, the execution of the data preparation programs could be slower.

In today's practical business applications of data mining, the most common models are classification models, followed by estimation models. Clustering models are also popular as a first step of analysis to reduce the data or describe large populations in terms of smaller segments.

This Page Intentionally Left Blank