Classification and Prediction

Databases are rich with hidden information that can be used for making intelligent business decisions. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Whereas *classification* predicts categorical labels, *prediction* models continuous-valued functions. For example, a classification model may be built to categorize bank loan applications as either safe or risky, while a prediction model may be built to predict the expenditures of potential customers on computer equipment given their income and occupation. Many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics, and neurobiology. Most algorithms are memory resident, typically assuming a small data size. Recent database mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data. These techniques often consider parallel and distributed processing.

In this chapter, you will learn basic techniques for data classification such as decision tree induction, Bayesian classification and Bayesian belief networks, and neural networks. The integration of data warehousing technology with classification is also discussed, as well as association-based classification. Other approaches to classification, such as k-nearest neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques are introduced. Methods for prediction, including linear, nonlinear, and generalized linear regression models, are briefly discussed. Where applicable, you will learn of modifications, extensions, and optimizations to these techniques for their application to data classification and prediction for large databases.

7.1 What Is Classification? What Is Prediction?

Data classification is a two-step process (Figure 7.1). In the first step, a model is built describing a predetermined set of data classes or concepts. The model is

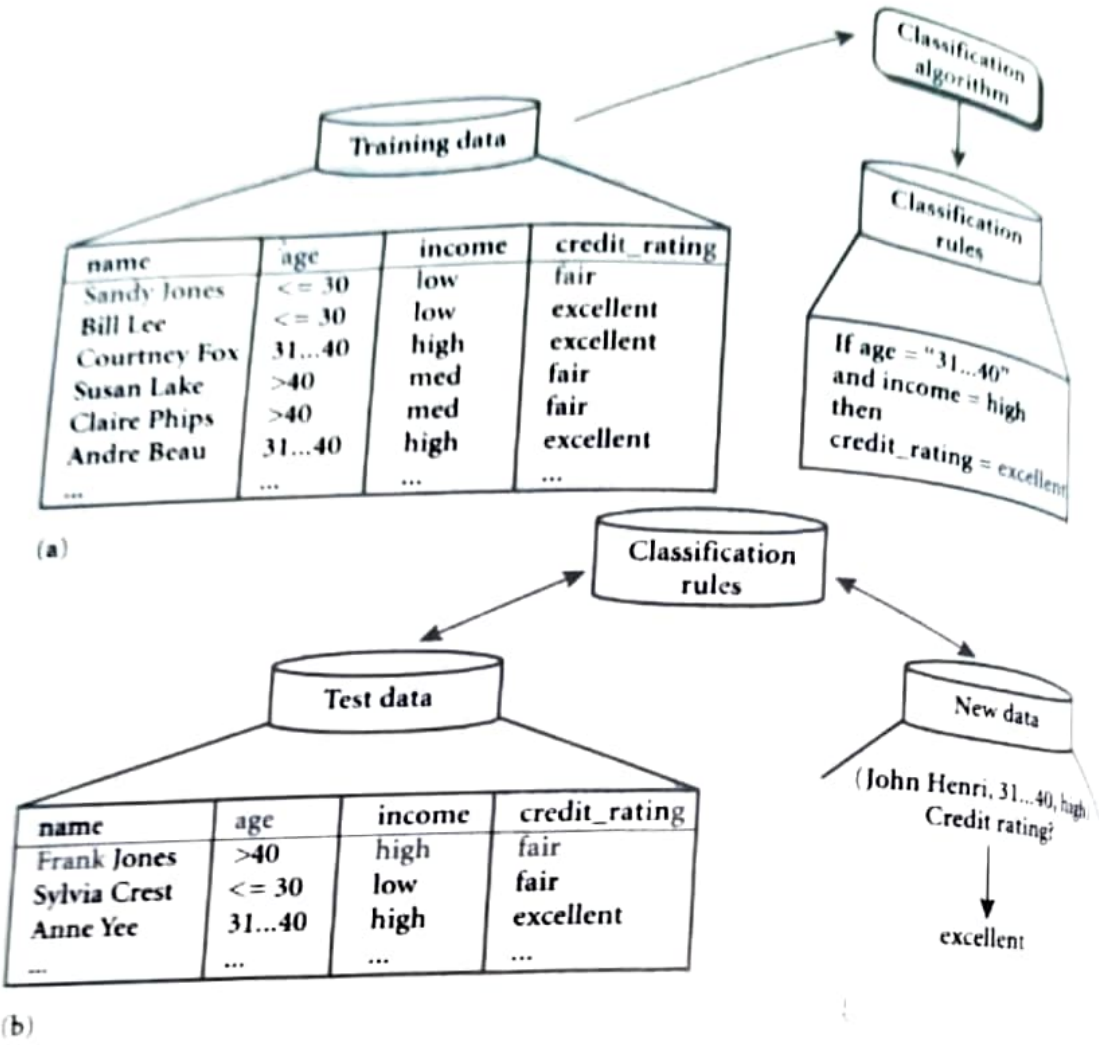

Figure 7.1 The data classification process: (a) *Learning:* Training data are analyzed by a classification algorithm. Here, the class label attribute is *credit_rating*, and the learned model or classifier is represented in the form of classification rules. (b) *Classification:* Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the **class label attribute**. In the context of classification, data tuples are also referred to as *samples, examples*, or *objects*. The data tuples analyzed to build the model collectively form the **training data set**. The individual tuples making up the training set are referred to as **training samples** and are randomly selected from the sample population. Since the class label of each training sample is provided, this step is also known as **supervised learning** (i.e., the learning of the model is "supervised" in that it is told to which class each training sample belongs).

contrasts with **unsupervised learning** (or **clustering**), in which the class label of each training sample is not known, and the number or set of classes to be learned may not be known in advance. Clustering is the topic of Chapter 8.

Typically, the learned model is represented in the form of classification rules, decision trees, or mathematical formulae. For example, given a database of customer credit information, classification rules can be learned to identify customers as having either excellent or fair credit ratings (Figure 7.1(a)). The rules can be used to categorize future data samples, as well as provide a better understanding of the database contents.

In the second step (Figure 7.1(b)), the model is used for classification. First, the predictive accuracy of the model (or classifier) is estimated. Section 7.9 describes several methods for estimating classifier accuracy. The **holdout method** is a simple technique that uses a **test set** of class-labeled samples. These samples are randomly selected and are independent of the training samples. The **accuracy** of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned model's class prediction for that sample. Note that if the accuracy of the model were estimated based on the training data set, this estimate could be optimistic since the learned model tends to **overfit** the data (that is, it may have incorporated some particular anomalies of the training data that are not present in the overall sample population). Therefore, a test set is used.

If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known. (Such data are also referred to in the machine learning literature as "unknown" or "previously unseen" data.) For example, the classification rules learned in Figure 7.1(a) from the analysis of data from existing customers can be used to predict the credit rating of new or future (i.e., previously unseen) customers.

"How is prediction different from classification?" **Prediction** can be viewed as the construction and use of a model to assess the class of an unlabeled sample, or to assess the value or value ranges of an attribute that a given sample is likely to have. In this view, classification and regression are the two major types of prediction problems, where classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered values. In our view, however, we refer to the use of prediction to predict class labels as *classification*, and the use of prediction to predict continuous values (e.g., using regression techniques) as *prediction*. This view is commonly accepted in data mining.

Classification and prediction have numerous applications including credit approval, medical diagnosis, performance prediction, and selective marketing.

Example 7.1 Suppose that we have a database of customers on the *AllElectronics* mailing list. The mailing list is used to send out promotional literature describing new products and upcoming price discounts. The database describes attributes of the customers, such as their name, age, income, occupation, and credit rating. The customers can be classified as to whether or not they have purchased a computer at *AllElectronics*.

Suppose that new customers are added to the database and that you would like to notify these customers of an upcoming computer sale. To send out promotional literature to every new customer in the database can be quite costly. A more cost-efficient method would be to target only those new customers who are likely to purchase a new computer. A classification model can be constructed and used for this purpose.

Suppose instead that you would like to predict the number of major purchases that a customer will make at *AllElectronics* during a fiscal year. Since the predicted value here is ordered, a prediction model can be constructed for this purpose.

7.2 Issues Regarding Classification and Prediction

This section describes issues regarding preprocessing the data for classification and prediction. Criteria for the comparison and evaluation of classification methods are also described.

7.2.1 Preparing the Data for Classification and Prediction

The following preprocessing steps may be applied to the data in order to help improve the accuracy, efficiency, and scalability of the classification or prediction process.

- **Data cleaning:** This refers to the preprocessing of data in order to remove or reduce *noise* (by applying smoothing techniques, for example) and the treatment of *missing values* (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics). Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.
- **Relevance analysis:** Many of the attributes in the data may be *irrelevant* to the classification or prediction task. For example, data recording the day of the week on which a bank loan application was filed is unlikely to be relevant to the success of the application. Furthermore, other attributes may be *redundant*. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. In machine learning, this step is known as *feature selection*. Including such attributes may otherwise slow down, and possibly mislead, the learning step.

Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting "reduced" feature subset, should be less than the time that would have been spent on learning from the original set of features. Hence, such analysis can help improve classification efficiency and scalability.

Data transformation: The data can be *generalized* to higher-level concepts. Concept hierarchies may be used for this purpose. This is particularly useful for continuous-valued attributes. For example, numeric values for the attribute *income* may be generalized to discrete ranges such as *low*, *medium*, and *high*. Similarly, nominal-valued attributes, like *street*, can be generalized to higher-level concepts, like *city*. Since generalization compresses the original training data, fewer input/output operations may be involved during learning.

The data may also be normalized, particularly when neural networks or methods involving distance measurements are used in the learning step. **Normalization** involves scaling all values for a given attribute so that they fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, *income*) from outweighing attributes with initially smaller ranges (such as binary attributes).

Data cleaning, relevance analysis, and data transformation are described in greater detail in Chapter 3 of this book. Relevance analysis is also described in Chapter 5.

7.2.2 Comparing Classification Methods

Classification and prediction methods can be compared and evaluated according to the following criteria:

- **Predictive accuracy:** This refers to the ability of the model to correctly predict the class label of new or previously unseen data.
- **Speed:** This refers to the computation costs involved in generating and using the model.
- **Robustness:** This is the ability of the model to make correct predictions given noisy data or data with missing values.
- **Scalability:** This refers to the ability to construct the model efficiently given large amounts of data.
- **Interpretability:** This refers to the level of understanding and insight that is provided by the model.

These issues are discussed throughout the chapter. The database research community's contributions to classification and prediction for data mining have emphasized the scalability aspect, particularly with respect to decision tree induction.