C

Credit Card Users' Data Mining

André de Carvalho
USP São Carlos, Brazil

Antonio P. Braga
UFMG, Brazil

Teresa Ludermir
UFPE, Brazil

MINING CREDIT RISK ANALYSIS DATA

The widespread use of databases and the fast increase in the volume of data they store are creating problems and new opportunities for credit companies. These companies are realizing the necessity of making efficient use of the information stored in their databases, extracting useful knowledge to support their decision-making processes.

Nowadays, knowledge is the most valuable asset a company or nation may have. Several companies are investing large sums of money in the development of new computational tools able to extract meaningful knowledge from large volumes of data collected over many years. Among such companies, companies working with credit risk analysis have invested heavily in sophisticated computational tools to perform efficient data mining in their databases.

Credit risk analysis is concerned with the evaluation of the profit and guaranty of a credit application. A typical credit risk analysis database is composed of several thousands of credit applications. These credit applications can be related to either companies or people. Examples of personal credit applications include for student loans, personal loans, credit card concessions, and home mortgages. Examples of company credits are loans, stocks, and bonds (Ross, Westerfield, & Jaffe, 1993).

Usually, the higher the value of the credit asked, the more rigorous is the credit risk analysis. Some large companies have whole departments dedicated to credit risk analysis.

The traditional approach employed by bank managers largely depends on their previous experience and does not follow the procedures defined by their institutions. Besides, several deficiencies in the data set available for credit risk assessment, together with the high volume of data currently available, make the manual analysis almost impossible. The treatment of these large databases overcomes the human capability of understanding and efficiently dealing with them, creating the need for a new generation of computational tools and techniques to perform automatic and intelligent analysis of large databases.

Credit analysis is essentially a classification task that involves the evaluation of the reliability and profitability of a credit application. The application of data mining techniques for credit risk analysis may provide important information that can improve the understanding of the current credit market and support the work of credit analysts (Carvalho, Braga, Rezende, Ludermir, & Martineli, 2002; Eberlein, Breckling, & Kokic, 2000; He, Hawkins, Graco, & Yao, 2000; Horst, Padilha, Rocha, Rezende, & Carvalho, 1998).

The extraction of useful knowledge from large databases is called knowledge discovery in databases (KDD). KDD is a demanding task and requires the use of sophisticated computing techniques (Brachman & Anand, 1996; Fayyad, Piatetsky-Shapiro, Amith, & Smyth, 1996). The recent advances in hardware and software make possible the development of new computing tools to support such tasks. According to Fayyad, Piatetsky-Shapiro, Amith, and Smyth (1996), KDD comprises a sequence of stages:

1. Selection
2. Preprocessing
3. Transformation
4. Data mining
5. Interpretation/evaluation

It is then important to stress the difference between KDD and data mining (DM). While KDD denotes the whole process of knowledge discovery, DM is a component of this process. The DM stage is used as the extraction of patterns or models from observed data. KDD can be understood as a process that contains, at least, the steps of application domain understanding, data selection and preprocessing, DM, knowledge evaluation and consolidation, and use of the knowledge. At the core of the knowledge discovery process, the DM step usually takes only a small part (estimated at 15–25%) of the overall effort (Brachman & Anand, 1996).

The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample, selected according to statistical techniques, is removed from the database, preprocessed, and submitted to the methods and tools of the DM stage, with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated regarding its quality and usefulness, so that it can be used to support a decision-making process.

Credit analysis databases usually cover a huge number of transactions performed over several years. The analysis of these data may lead to a better understanding of the customer's profile, thus supporting the offer of new products or services. These data usually hold valuable information, e.g., trends and patterns, that can be employed to improve credit assessment. The large amount makes its manual analysis an impossible task. In many cases, several related features need to be simultaneously considered in order to accurately model credit user behavior. This need for automatic extraction of useful knowledge from a large amount of data is widely recognized.

DM techniques are employed to discover strategic information hidden in large databases. Before they are explored, these databases are cleaned. Next, a representative set of samples is selected. Machine-learning techniques are then applied to these selected samples. The use of data mining techniques on a credit risk analysis database allows the extraction of several relevant information regarding credit card transactions.

The data present in a database must be adequately prepared before data mining techniques can be applied to it. The main steps employed for data preparation are as follows:

1. Preprocess the data to the format specified by the algorithm to be used
2. Reduce the number of samples/instances
3. Reduce the number of features/attributes
4. Features construction, which is the combination of one or more attributes in order to transform irrelevant attributes to more significant attributes
5. Elimination of noise and treatment of missing values

Once the data have been prepared, machine-learning (ML) techniques can be employed to discover useful knowledge. The quality of a knowledge extraction technique can be evaluated by different measures, such as accuracy, comprehensibility, and new, useful knowledge.

CRITICAL ISSUES OF CREDIT RISK ANALYSIS DATA MINING

There are, of course, a few problems that can arise from the use of data mining techniques for credit risk analysis. The main advantages and problems associated with this technology are briefly presented in Table 1. Current Internet technology may allow the invasion of company databases, which may lead to the access of confidential data. Besides, the stored data may be made available to other companies or used without the knowledge or authoriza-

Table 1. A summary of critical issues of credit risk analysis data mining

Training of employees Credit analysts should be trained with the new tools available and shown the benefits of the new technology
User ignorance and perceptions Lack of adequate understanding of the data mining and its usefulness
Inadequate information Applicants may have supplied incorrect information,intentionally or not
Maintain integrity of data Maintain up-to-date and accurate information on the databases
Security Maintain secure and safe systems and denying unauthorized users access
Privacy and confidentiality agreements Address individual's right to privacy and the sharing of confidential information

Related Content