Credit Risk Assessment and Data Mining

André Carlos Ponce de Leon Ferreira de Carvalho
Universidade de São Paolo, Brazil

João Manuel Portela Gama
Universidade do Porto, Portugal

Teresa Bernarda Ludermir
Universidade Federal de Pernambuco, Brazil

INTRODUCTION

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

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.

The behavior of the financial market is affected by a large number of political, economic, and psychological factors, which are correlated and interact among themselves in a complex way. The majority of these relations seems to be probabilistic and non-linear. Thus, these relations are hard to express through deterministic rules.

Simon (1960) classifies the financial management decisions in a continuous interval, whose limits are non-structure and highly structured. The highly structured decisions are those where the processes necessary for the achievement of a good solution are known beforehand and several computational tools to support the decisions are available. For non-structured decisions, only the managers' intuition and experience are used. Specialists may support these managers, but the final decisions involve a substantial amount of subjective elements. Highly non-structured problems are not easily adapted to the computer-based conventional analysis methods or decision support systems (Hawley, Johnson, & Raina, 1996).

BACKGROUND

The extraction of useful knowledge from large databases is named *knowledge discovery in databases* (KDD). KDD is a very 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 a task. According to Fayyad et al. (1996), KDD comprises a sequence of stages, including:

- Understanding the application domain,
- Selection,
- Pre-processing,
- Transformation,
- Data mining, and
- Interpretation/evaluation.

It is also 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 the previous listed steps. 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/or usefulness, so that it can be used to support a decision-making process.

Frequently, DM tools are applied to unstructured databases, where the data can, for example, be extracted from texts. In these situations, specific pre-processing techniques must be used in order to extract information in the attribute-value format from the original texts.

CREDIT RISK ASSESSMENT

Credit risk assessment is concerned with the evaluation of the profit and guaranty of a credit application. According to Dong (2006), the main approaches proposed in the literature for credit assessment can be divided into two groups: default models and credit scoring models. While default models assess the likelihood of default, credit scoring models assess the credit quality of the credit taker. This text covers credit scoring models.

A typical credit risk assessment database is composed of several thousands of credit applications. These credit applications can be related with either companies or people. Examples of personal credit applications are 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 assessment. Large financial institutions usually have whole departments dedicated to this problem.

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 dataset available for credit risk assessment, together with the high volume of data currently available, makes 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.

In 2004, the Basel Committee on Banking Supervision published a new capital measurement system, known as the New Basel Capital Accord, or Basel II, which implements a new credit risk assessment framework that supports the estimation of the minimum regulatory capital that should be allocated for the compensation of possible default loans or obligations (Basel, 2004; Van Gestel et al., 2006). The Basel Committee was created in 1974 by the central-bank of 10 countries. In 1988, the committee introduced a capital measurement system commonly referred to as the Basel Capital Accord. The first accord established general guidelines for credit risk assessment. The new accord, Basel II, stimulates financial institutions to adopt customized rating risk systems based on their credit transaction databases. As a consequence, DM techniques assume a very important role in credit risk assessment. They will allow the replacement of general risk assessment by careful analysis of each loan commitment.

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, for example, trends and patterns, which 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.

USING DATA MINING FOR CREDIT RISK ASSESSMENT

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 pieces of 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:

- Preprocessing of the data to the format specified by the algorithms to be used;
- Reduction of the number of samples/instances;
- Reduction of the number of features/attributes;
- Features construction, which is the combination of one or more attributes in order to transform irrelevant attributes to more significant attributes; and
- Noise elimination and treatment of missing values.

Once the data have been pre-processed, 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.

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; Horst, Padilha, Rocha, Rezende, & Carvalho, 1998; Lacerda, de Carvalho, Braga, & Ludermir, 2005; Dong, 2006; Huang, Hung, & Jiau, 2006).