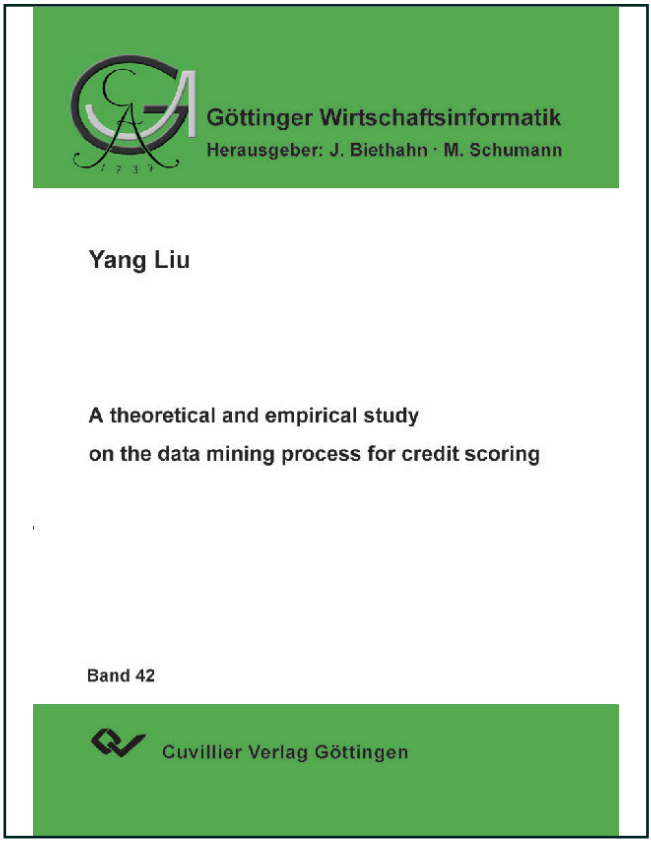
Yang Liu (Autor)

A theoretical and empirical study on the data mining process for credit scoring



https://cuvillier.de/de/shop/publications/3168

1 Introduction

1.1 Motivations and aims

With the continuous changing and development in the credit industry, credit products and services are becoming more and more important in the economy. The increasing demand and increasing competition resulting from new economic environment offer new opportunities, but also put forward new requirements. Credit granting institutions pursue urgently cost savings and efficiencies. This has led them to expanding the role of technology in their credit management process.

The new economic environment can be recognized in the following aspects:

1. Increasing credit risk

   As the volume of credits increases, the volume of insolvent credits presents an increasing trend too. In Germany, for example, the total number of insolvency (private and companies) in 1991 is about 13,300. After ten years in 2001 the number increases to almost 50,000, among them above 32,000 are company insolvency. In 2002 the total number of insolvency is expected to increase to around 60,000 and the company insolvency to 43,000 (cf. DeSt02; Böhm02, P. 52). Both financial institutions and regulation institutions pay much more attention to the increasing credit and the risks associated with it. Financial institutions need to invest considerable resources to develop efficient and sophisticated tools to evaluate and control credit risks.

2. Online credit channel

   Economic globalization and newly emerging service channels such like call center, internet and mobile communication provide possibilities for the customers to seek and choose their creditors without regional and time limitations. Because of this trend, creditor must now be ready, willing and able to extend credit to business in other countries or regions around the world. Banks and other financial service organizations are therefore facing a more drastic worldwide competition. They are realizing that their online-channel means adding more demand on their services. The challenge to the decision-makers behind all of these channels is how to best serve and protect the customer. In a complete, end-to-end, robust Internet lending solution, they must make consistent and intelligent real time decisions on their large quantity of online credit applications.

3. Supervision requirement

   According to the recently published ‘‘The New Basle Capital Accord’’, banking supervisors have moved towards accepting the internal-ratings based approach as a basis for the determination of adequate capital reserves for credit risks. This will generate significant advantages for those credit institutions that have a sound internal rating system. The establishment of a rating system is a challenge to many credit institutions, especially to small and middle banks (cf. LeDö02, P. 50). Experiences and techniques in credit risk modeling are demanded more pressingly than before.

Credit scoring model technology can supply the basic part of credit decision support systems that serve the new requirements for credit risk management. Increasing competition, pressuring margins and decreasing customer satisfaction have placed modern scoring model techniques at the forefront of priorities for credit services. Today, advanced scoring model techniques play a necessary and crucial role in helping credit decision making and management.

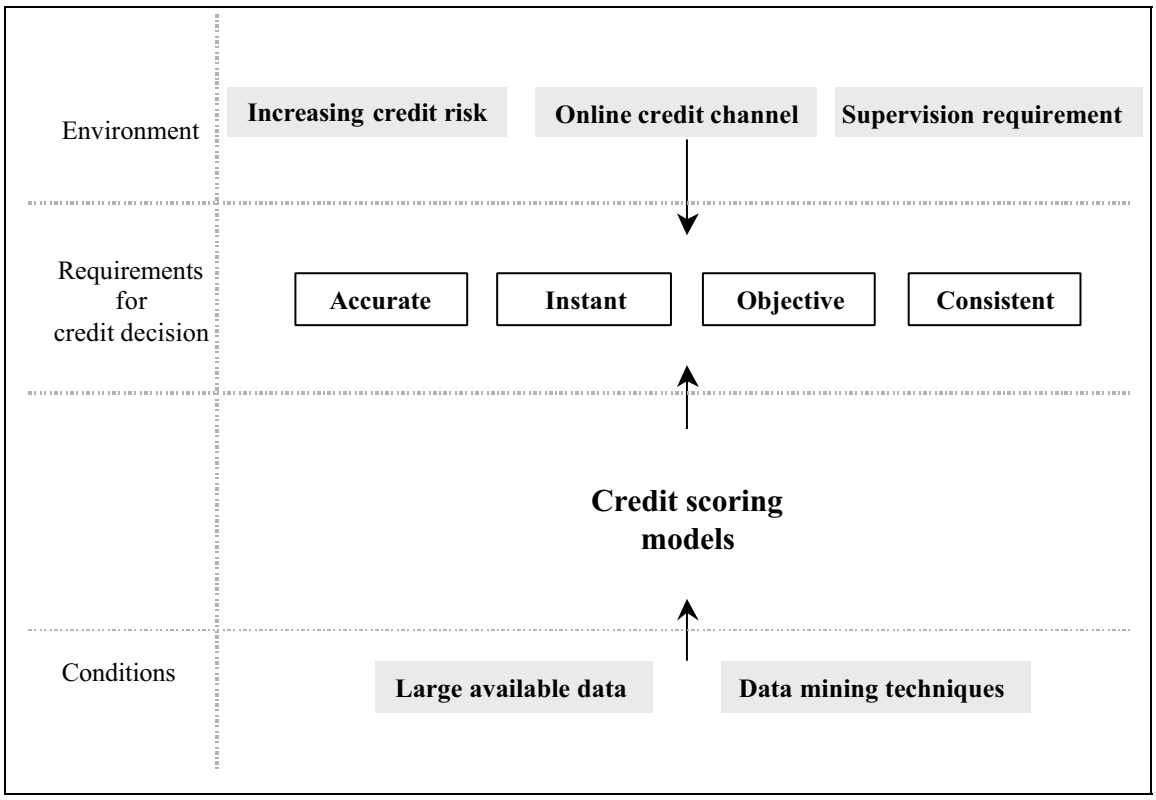


Figure 1.1/1: Promotions of the research and the application of credit scoring

Furthermore, new conditions are also provided for the development of scoring models. The large available data coming from both the public credit information companies and the data warehouse of financial service institutions have enabled researchers to

investigate the use of these data (cf. Eise96, P. 273). The advent of data mining techniques means that the technical problems of analyzing such vast amounts of data are being addressed (cf. Thom00, P. 165) (see Figure 1.1/1).

All these requirements and conditions have jointly served to prompt serious attempts to develop scoring models both for consumer credit and for business credit. Credit scoring, a technique that can support quick, objective, accurate, consistent credit decisions, can be expected to be widely used. The standardization and the automatization of credit decision that is already realized to some extent in the section of private customers will be also extended to business customers in the future.

The credit scoring models involve the techniques that are called today the techniques of data mining. Classification methods are the most commonly used data mining techniques that have been applied in the domain of credit scoring to predict the default probabilities of credit takers. Many methods, such as linear and logistic regression, decision trees, neural networks, etc., have been used for developing credit scoring models. The search for commercial advantage in the credit industry has led to interest in a new and emerging technology --- data mining.

Data mining as an approach to support computer-based decision making is actually not a purely new technology but borrows many algorithms from statistics, artificial intelligence and other fields. It is not the algorithms of data mining but the idea of automatically getting knowledge from large databases is revolutionary. Nowadays, large amount of clean and well-documented data in organizations and more cost-effective IT solutions in terms of storage and processing ability make this idea more realistic. New algorithms from research centers and universities are able to enter into commercial software (cf. Cabe98, P. 11; Schi99, P. 99). Although the implementation of totally automatic knowledge discovery from database is still far away from the expected ideality, this new concept and the continuous research endeavors on it give the opportunity for the future's revolution in computer-based decision making.

The research of data mining necessarily involves many different areas, including the background areas of different data mining techniques like statistic, machine learning, and the areas of computer science like database and parallel computing aiming at assisting and speeding data mining. In this dissertation data mining is considered as a decision support process that enable users to solve business problems (specifically, credit scoring problem). From this view the process of data analysis employing data mining techniques is mainly concerned. Compared with other aspects of the data mining, research on the process of data mining is very rarely mentioned in previous

studies. However, a successful application of data mining in practice sometimes depends decisively on the strategies of controlling the data mining process.

In the view of data mining, the simple use of one or several algorithms on the available data set is replaced with a complex time-consuming process, which is full of trials and iterations. With the maturity of the mining algorithms and the best understanding of the problem domains, the most challenging is how to control the mining process. The control of mining process is related to but still beyond the mining algorithms and the domain knowledge. In this sense, the aim of data mining is not only the good performance of particular algorithms, but also to get the most applicable results with the least time and cost.

There are some essentially similar descriptions on the process of data mining in literatures. Some standard methodologies are also proposed, such as “Cross-Industry Standard Process for Data Mining” (cf. CRISP-DM00). Many descriptions of data mining process gave a list of steps of the data mining procedure which lacks for detailed discussions of how these steps to be accomplished for a specific application problem: which mining techniques are involved, which evaluation criterion are used to determine whether going to the next step or iterating to earlier steps.

This dissertation aims to study the data mining process for the credit scoring problem. Credit scoring can be looked as the type of classification problem of data mining. Meanwhile, its practical applications accompany many problems relevant to the credit industry. Due to the complex decision process credit scoring has always been based on a pragmatic approach: A solution can not be the optimal one for everywhere, just for specific circumstances. The process of credit scoring is not standardized. A serious problem with this nonstandard model building process is an aimless, repeated and expensive data analysis process that cannot yet guarantee an optimal model solution.

Data Mining approach not only utilizes a multi-strategy combining multiple statistics and machine learning algorithms of classification, but also provides a framework for data analysis process, which includes necessary pre-processing of the real-world large data sets and post-processing of the model outputs in order to support a standard model building process.

In this dissertation, the techniques of data mining and the strategies in controlling the data mining process are going to be concerned in the application area of credit risk management. Both the theoretical problems in the data mining approach and the practical problems in the credit scoring models' application are going to be addressed. The work intends to develop a systematic data mining process framework that can be

applied particularly on the credit scoring problem. Some data mining techniques are empirically studied under this framework to show their effectiveness and performance in the real-world credit scoring model building.

1.2 Organization of the dissertation

The second and third chapters provide general overviews of the research and practice in the credit scoring and the data mining areas. They establish the theoretical fundament of the dissertation and position the work within the scope of these two areas.

The second chapter introduces the origins, development, and current problems associated with credit scoring. It covers the definition of credit scoring, its relationship with credit rating and how the application range of scoring methods are expanded facing new economic environment. After that we go deep into some data and application issues faced by credit scoring model builders.

In Chapter 3 the background knowledge of data mining is introduced. A general data mining process is given, which will be specialized in Chapter 4 for the problem of credit scoring. Afterwards, the classification problem is specially introduced. The professional backgrounds and the approaches of model generation are explained in order to give a deep understanding of the classification model building. Five classification algorithms are described which are going to be used in the empirical studies in later chapters. A comparison of the introduced methods is given allowing us to identify the relative advantages and disadvantages of these methods and their applicability in credit scoring.

In Chapter 4 an overall framework of data mining process for the problem of credit scoring is constructed. To incorporate new data mining techniques in the credit scoring models, three subprocesses are formed for the data analysis stage of data mining. The empirical studies in the following chapters are constructed under this general framework.

In Chapter 5, 6 and 7, the evaluation subprocess, the input relevant subprocess and the model-relevant subprocess are further developed (see Figure 1.2/1). The scoring model techniques, model evaluation methods, feature selection techniques, and model combination techniques are studied empirically with real-world credit data.

Finally, summarization and conclusions are given in the Chapter 8. Proposals for further research are also pointed out.