Research on Risk Management of Green Credit in Commercial Banks Based on Distributed Database

Qingqing Duan

Department of Finance and Economics, Guangdong University of Science and Technology, Dongguan , China

Abstract. Data mining techniques are used in order to develop a credit risk management system for commercial banks. Firstly, the concepts related to data mining and data warehouse technology were introduced, and the existing credit management situation was analyzed. Subsequently, a credit risk management model was designed and implemented based on data mining, taking into account the actual characteristics of China's commercial banking industry. On this basis, attributes were selected before the classification, which not only improved the overall performance of the classifier but also reduced the cost of data acquisition, enabling commercial banks to improve the efficiency of their credit work.

Keywords: Distributed database; Commercial banks; credit risk management

1. Introduction

For a long time, due to the limitation of hardware facilities and data processing technology, the managers of commercial banks cannot grasp the relevant information of loans in a comprehensive manner, and often fail to make a correct assessment of the risk of credit assets, which leads to wrong decisions. One of the major challenges facing the commercial banking industry is how to discover the patterns of risk assessment from the vast amount of data to enable commercial banks to avert risks.

Approved by the Chinese government, Postal Savings Bank of China was incorporated by law on March 6, 2007. The micro loan business of postal savings does not require pledge or collateral, and has the advantages of flexible guarantee methods, quick payment of loans and high loan amount. It is of great significance to broaden the financing channels for urban and rural residents, effectively alleviate the problem of difficult lending for farmers, and boost the income of farmers and economic development in rural areas. However, risk prevention is the eternal theme of the financial industry. It should strengthen the analysis of macroeconomic situation and market research while launching business, pay close attention to various risks that may appear and maintain financial security and stability. In the credit risk management of commercial banks, commercial banks can use data mining technology to discover the objective laws hidden in the huge amount of credit data and to better reduce the business risks of financial institutions. This paper applies data mining technology to solve the credit risk management problem of a postal commercial bank.

2. Credit problems in commercial banks

Data mining is a new technique for discovering and extracting hidden information and knowledge from large amounts of data that are not known to people beforehand but may be useful [1]. It is a non-trivial process of discovering previously unknown, regular, and hidden information and knowledge from a large amount of data [2]. Data mining and knowledge discovery theories have been extensively researched since their introduction in August 1989. Their research involves fundamental theory, discovery algorithms, data warehouse, visualization techniques, qualitative and quantitative interchange models [3], knowledge representation methods, maintenance and reuse of discovered knowledge, knowledge discovery in semi-structured and unstructured data, and online data mining.

There are many general-purpose data mining systems that are applicable to horizontal solutions for various business applications, while the applications of dedicated data mining systems are mostly concentrated in sectors such as telecommunications, insurance, biomedicine, and retail. All the work of data mining should be linked to the actual business of the fields of work, and the commercial

banking industry should determine the specific data analysis themes according to the characteristics of different credit businesses of each commercial bank. Many commercial banks across the country are influenced by various factors such as the level of economic development of each place, and some of their businesses will definitely have some differences and operational characteristics. Therefore, only when a suitable data analysis theme is determined by closely combining with the actual business of a specific commercial banking industry, the data mining results will be of practical significance. In this paper, loan risk classification and analysis is defined as the goal of data mining.

In loan management, loan quality is generally classified following a risk-based approach. The People's Bank of China classifies loans into different classes according to the degree of risk and promulgated the Guidelines on Risk-Based Loan Classification in December 2001, where loans are classified into five categories, pass, special-mention, substandard, doubtful and loss; with the last three categories recognized as non-performing loans [5]. Commercial banks can directly adopt this criterion to classify their loan risks, or they can develop their unique loan classification system based on the Guidelines. However, it is required that the loan classification system developed by commercial banks should have highly defined correspondence and conversion relationship with the risk-based loan classification approach adopted by the People's Bank of China.

Commercial banks can analyze the characteristics that different categories of loans have through data mining and build a model. When there is a new loan application, commercial banks can estimate its category using the model, so that they can take appropriate measures for different categories of loan applications. For example, commercial banks can directly approve and pass loans that fall within the pass category, while they need to strengthen the examination of loans below the special-mention level, or strengthen the post-loan inspection of the enterprise, or reject the loan, thus improving the security of credit assets. The risk level of credit assets will also change with the business operation of the enterprise. Commercial banks need to re-analyze the current classification of each loan at certain intervals to improve the management of credit risks and reduce the loss of their credit assets.

3. Solutions to credit risk management problems of commercial banks based on data mining

In this paper, WEKA 3.5.7 was used as a data mining platform. Waikato Environment for Knowledge Analysis (WEKA) is a data mining application with open source code [6].

3.1 Raw data description

According to statistics, since two-thirds of postal savings outlets are located in counties and townships, 80% of loans have been granted to rural areas since the start of micro pledged loans and micro loan business of postal savings. The micro loan business of postal savings is further divided into two types: Farmer Loan Express and Merchant Loan Express. Among them, Farmer Loan Express refers to short-term loans granted to farmers to meet their needs in crop cultivation, farming or non-agricultural (articles of daily use, production and processing, services, construction, transportation, etc.) production and operation. Merchant Loan Express refers to the loans provided to micro and small business owners engaged in wholesale and retail, service (catering), production and processing sectors to meet their capital needs in business. In this paper, Merchant Loan Express, one of the micro loans of postal savings, is selected as the object of study. Merchant Loan Express is divided into two types: merchant joint-guarantee loan and merchant guarantee loan.

For the Merchant Loan Express business studied in this paper, there are many data tables involved, such as customer and household information table, business information table, purchase information table, seasonal analysis table, gross margin calculation table, balance sheet, income statement, guarantor information table, and group guarantee information table. Although all of these messages are relevant to the business, not all of them are beneficial to the research in this paper. In order not to violate and leak the secrets of merchants, we filtered the attributes such as business license number, merchant's name, residence address, store or factory name, and contact information in the process of

extracting data in this paper. After analysis, we extracted 17 fields of customer code, marital status, loan type, education level, age, loan amount, loan term, repayment method, main business, years of operation, total current assets, total fixed assets, liabilities, monthly net income, monthly investment, credit, and classification result as the data of the fact table.

3.2 Data pre-processing

The source data collected after preliminary acquisition is often incomplete, noisy, and inconsistent. A large amount of noisy data is present in the databases of commercial banks due to manual input errors, malfunctioning of data collection equipment, and errors in data transmission [7]. And some attributes, such as the income status of the customer, as well as the source of income, are not recorded exhaustively and accurately. Some data, such as housing status, working organization, job title, and family size, are entered as null values in the database. Therefore, it is necessary to pre-process these incorrect and null data first.

In this stage, the data were collected, selected, cleaned, and transformed. Seventeen attribute fields were selected for data extraction, and 100 records were randomly selected and collated from the database. Among them, the respondents' marital status are married (No loans will be granted to those who are not married), and the repayment method is phased repayment with average capital plus interest per instalment. These two attributes have no reference value for loan classification, so these 2 attributes are removed. The customer code is removed from the attributes because it takes many values and has no generalization operation. The generalized results for the other attribute fields are shown in Table 1.

Table 1. Generalized attribute fields

Attribute name	Value range	Description
Type of loan	1~2	1: Merchant guarantee 2: Merchant joint-guarantee
Education level	1~3	1: Junior high school and below 2: Senior high school or secondary school 3: College and above
Age	1~3	1: 18~30 2: 31~45 3: 46~60
Loan limit	1~4	1: 0≤…<50 000 2: 50 000≤…<100 000 3: 100 000≤…<150 000 4: 150 000≤…≤200 000

There are 52 classified cases in the sorted customer profiles. Among them, 30 were in the pass category, 9 were in the special-mention category, 6 were in the substandard category, 5 were in the doubtful category, and 2 were in the loss category.

Since most of the borrowers' financial information in the loss category was not available, only the first 4 categories were involved. The actual cases were 30 in the pass category, 9 in the special-mention category, 6 in the substandard category, and 5 in the doubtful category, for a total of 50.

Based on the above data preparation, the training data set of this model was obtained as shown in Table 2.

Table 2. Part of the training data set

Loan type	Education level	Age	Loan limit	Loan term (months)	Years of operation	Total current assets	Total fixed assets	Liabilities	Main business	Monthly net income	Monthly investment	Reputation status	Loan category
1	1	3	3	12	4	4	1	0	Electric vehicles	2	2	Good	Pass
1	1	2	2	12	3	3	1	1	Electric vehicles	1	3	Good	Special-mention
1	2	2	2	12	2	1	2	1	Subsidiary food	1	2	Good	Pass
1	1	2	3	12	4	3	2	3	Toys	2	2	Good	Pass
2	2	2	3	6	3	2	1	0	Electric vehicles	1	3	Good	Pass
2	2	1	2	6	1	3	2	0	Motorcycles	3	3	Moderate	Pass
1	2	2	2	12	2	3	1	0	Beddings	2	3	Good	Pass

3.3 Building a decision tree

The data in the above table were all converted to a data file format (CSV Data Files) that WEKA can read. Next, WEKA was used to build the model. We first launched the Explorer interface of WEKA and load the data. Then we chose a method to build a decision tree [8]. We analyzed the experimental results of nine classifiers, BFTree, DecisionStump, J48, LMT, NBTree, RandomForest, Randomtree, REPTree, SimpleCart, and found that J48 classifier has the highest accuracy rate.

3.4 Model evaluation

Based on the established classification models and sample data, the prediction accuracy of the models was evaluated. The accuracy of the model can be expressed as the percentage of test samples that are correctly classified by the model. If the prediction accuracy of the model is acceptable, we can use it to guide the classification of customer groups. We applied the J48 classifier for classification evaluation with an accuracy of 82%, which means that out of 50 samples of data, 41 were correctly classified and 9 were incorrectly classified. This evaluation result was obtained by the default hierarchical 10-fold cross-validation.

4. Improvement

Data mining is a systematic process from source data mining, knowledge discovery to application [8], and it requires more than just having algorithms. In the classification process, the classification performance generally improves as the number of selected attributes increases. However, when the number of attributes increases to a certain level, the performance of the classification is sometimes degraded by adding more attributes, a phenomenon called Hughes Phenomenon. Therefore, although from a theoretical point of view, more attributes selected implies an increase in the amount of information, the performance is deteriorated when there are too many attributes, because it always works on samples of limited size in practical applications. Therefore, it is necessary to perform attribute reduction in the integrated design of classifiers.

The accuracy of the decision tree model constructed above using the J48 classifier is acceptable, and commercial banks can derive an estimated category for each new loan application by the model, and thus take appropriate measures for different categories of loan applications. For example, commercial banks can directly approve and pass loans that fall into the pass category, while loans below the special-mention category need to be examined more closely, or post-loan inspection of the enterprise should be strengthened, or the loan should be rejected, thus improving the security of credit assets. The risk rating of credit assets also changes with the business operation. For this reason, commercial banks need to re-analyze the current classification of each loan at certain intervals and then summarize the trend of changes in loan classification characteristics to improve the management of credit risks and reduce the loss of credit assets.

5. Conclusion

We selected and generalized 14 attribute fields closely related to the classification results in our application, and preprocessed a large amount of data to obtain the training set. Then we conducted effective data mining on the training set using the WEKA 3.5.7 mining platform. Here we chose the J48 classification algorithm, which not only improves the overall performance of the classifier but also reduces the cost of data collection and significantly improves the efficiency of commercial banks' credit work by performing attribute selection prior to classification. Thus, a simple application of data mining techniques was completed for risk-based loan classification in a classification technique based on decision trees.

References

[1] Zhao Jun, Zhang Chunhai, Li Hua Data sharding strategy of distributed database based on XML middleware [J] Computer Engineering and Design, 2006, 27 (3): 3

[2] Pan Ying Application of adaptive genetic algorithm in query optimization of distributed database [J] Journal of Inner Mongolia Normal University (Natural Science Chinese Edition), 2016, 045 (001): 94-97

[3] He Ming, Chen Guohua, Liang Wenhui, Lai Haiguang, in the early morning Research on Cloud Data Storage Security and Privacy Protection Strategy in the Internet of Things Environment [J] Computer Science, 2018, 39 (5): 62-65

[4] Dai Hongjun, Wu Guoqiang, Liu Liming Research on Industry Risk Early Warning in Credit Management of Commercial Banks [J] Journal of Changchun University of Science and Technology: Social Science Edition, 2013, 26 (8): 4

[5] Zhang Bo Measurement Methods and Empirical Research on Credit Industry Risk of Commercial Banks [D] Yunnan University, 2010

[6] Chen Guanying Problems and solutions of credit management of local commercial banks [J] two thousand and twenty-one