Association Rules Algorithm in Bank Risk Assessment

Guorong Xiao

Department of Computer Science and Technology
GuangDong University of Finance Guangzhou, China
newducky@126.com

Abstract. Domestic financial enterprise data management applications generally brings together the vast amounts of data, but can not find the relationship and business rules exists in the data to do risk prediction assessment. Therefore, the domestic financial companies need to accelerate the pace of information technology in regions of integration of customer resources, business analysis and investment decisions. This paper analyzes the risk assessment approach of banks, mainly focuses on the analysis of association rules data mining in bank risk assessment, and discusses the working principle of improved association rules algorithm genetic algorithm in commercial bank risk assessment. We described the methods and processes of system application. We select the matrix form, only scan the database once, and use the method of selecting assumption frequent items and numbers, find the frequent item sets through high end item sets, minimize the number of candidate data sets, greatly improve the efficiency of the algorithm.

Keywords: bank; risk assessment; association rules; algorithm.

1 Introduction

With the extensive application of database and computer networks, demands on banking sector increased continuously. In the face of three key factors, change, competition and customer, which impact and determine the development of bank, if there is no information technology, it will become increasingly difficult for banks to understand, grasp and respond to change and increasingly difficult to integrate resources, planning restructuring to cope with competition, also difficult to achieve their business process reengineering and strategic tasks of intelligent marketing and management decisions. Comprehensive information management is the real core competitiveness of bank. Therefore, domestic commercial banks all consider the development of entering informatization as an important strategic move, established a relatively perfect system of financial informatization. Since the end of 2006, China opens the full domestic financial markets to WTO. Foreign banks are allowed to engage in RMB totally, the competition between domestic and foreign banks become more intensive. In this global financial crisis, although China's banking sector was not seriously impacted itself, the market turbulence remains intense, and China's banking sector also needs to seize the opportunity, make great efforts to promote the process of informatization of domestic banking sector.

J. Lee (Ed.): Advanced Electrical and Electronics Engineering, LNEE 87, pp. 675–681.
springerlink.com

For credit card, here exist the issues of malicious overdraft and fraud which bring a great deal of risk to bank. Therefore, card issuers should take effective measures to prevent risks in advance, to quantify the credit rating to applicants' qualifications and credit, and then assist decision-makers to decide whether to give the credit card to the applicant. Usually bank judge the credit of applications through statistical techniques and experience, however, with rapid increase of credit card users and the trading volume, Experience alone is not enough to effectively make the right judgments, therefore, we need to introduce intelligent information processing technology to provide decision support to decision-makers. In this paper, we focus on the association rules of intelligent algorithm, discussed and researched upon risk assessment of banks.

2 Bank Risk Assessment Methods

Bank risk refers to the possibility of encountering economic loss during operating by various factors, or just the possibility for banks to meet assets and income loss. According to the cause of risk, the risks include credit risk, market risk, interest rate risk and legal risk, credit risk is the current key risk facing banking sector and also the main research of this paper. Since the 30's of 20th century, bank credit risk assessment method has mainly gone through 3 stages, judge according to experience, statistical analysis and artificial intelligence.

Mainly includes the following methods.

- Expert Judgment, in the initial phase of the credit rating, as the historical data information of trading partners is not enough; the level of trading partners' credit is entirely based on subjective experience of credit experts. This method is not efficient, costs high, and often have inconsistent conclusions.
- Scoring method, Banks and credit rating companies based on a pre-designed set of standardized indicators system, to rate each indicators of risk status about trading partners and customers, then average the rate according to the importance, and make the totally score the main judgment to customer risk rating. This method requires risk management experts to set indicator and importance according to their experience. The scoring of each indicator also need experts to use their experience and feelings, therefore, the level and experience of experts has a great impact on the effectiveness of ratings.
- Model approach, the method of risk rating system is based on trading partners or customers' historical database. Built the probability statistical model on historical data, including the discriminate analysis model, probability of default measurement model and loss given default rate measurement model. This method has the advantage of high efficiency; low cost, high accuracy measurement of default risk factors, inadequacies is difficult to directly enter the model of qualitative indicators, making it difficult to reflect the qualitative indicators of information.

3 Data Mining

With the rapid development of database technology and the widely application of management systems, people accumulated more and more data. Many important

information are hidden behind data, people want to have higher level of analysis to make better use the data.

The current database system can efficiently implement data entry, modification, statistics, query and other functions, but can not find the relationships and rules that exist among data, thus can not predict the future trends according to the current data. As currently lack of means to detect the knowledge behind the data mining, this led to the "data explosion but lack of knowledge" phenomenon. Therefore, intelligently and automatically valuable knowledge and information research among large amounts of data, know as data mining, is of great practical significance and wide application prospects.

From the technical point of view, Data mining is the process of the extraction of implicit, and unknown but potentially useful information and knowledge among a lot of, incomplete, noisy, fuzzy, random, real data. From the perspective of business applications, data mining is a new business information processing technology. Its main feature is a commercial database or data warehouse large amounts of data extraction, transformation, analysis and pattern processing, to extract the key of knowledge supporting business decisions, from a database or data warehouse model to automatically find related business.

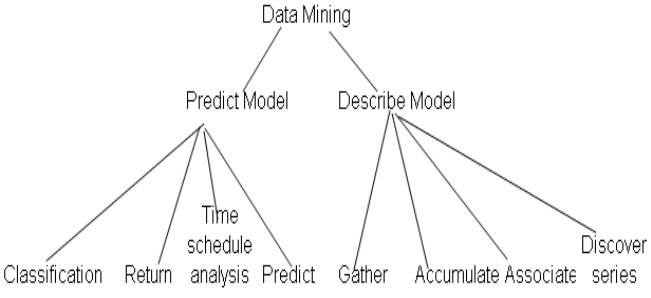

Fig. 1. Classification of data mining.

In short, data mining is actually a class of in-depth data analysis method. Data analysis itself has many years of history, only that the data collection and analysis in the past targeted at scientific research. But data mining is to do exploration and analysis on a large number of enterprise data in accordance with corporate business objectives. It's the effective ways of reveal the hidden, unknown or the verify regularity, and further the model.

The predictive model will predict the value of the data from results known from different data. It can accomplish data mining tasks, including classification. Map data to the predefined group or class. Return (the data item is mapped to a real predictor variable). Time series analysis, the behavior of determines the series according to the distance and structure. Forecast data value based on the historical event sequence diagram. Predict (based on past and current data to predict the future state of the data). Descriptive model identify the data in patterns or relationships. Different from the predictive models, descriptive models provide the method of data nature analysis, rather than predict the new nature. It usually consists of Gather (unsupervised learning or

partition), Accumulate (to be accompanied by a brief description of the data mapped to the subset), association rules (reveal the relationship between the data) and the series discovery (identifying data related to time between the sequence modes).

4 Association Rules in Bank Risk Assessment

At present, the association rule mining technology has been widely used in western financial industries and enterprises; it can successfully forecast demand for bank customers. Once got this information, banks can improve their marketing. The current banks are developing new methods of exploring new way of communication with customer every day. All banks bundled product information that may be interesting to customer in their ATM machines. To benefit customer who use their ATM and what to understand the products. If the database shows that a high credit customer change the address, the client is likely to buy a new bigger house, so there may need a higher credit limit, the higher end of the new credit card, or need a housing improvement loans, these products can be mailed to the customer through the credit card bill. When customers call to consult, the database can effectively assist telephone sales representative. Sales representative's computer screen can show the characteristics of clients, and what products the customer would be interested in.

4.1 Association Rules

Agrawal first proposed in 1993, equivalent to mining a database of customer transactions association rules between sets of items, and designed a basic algorithm, its core is based on the frequency of the recursive method of set theory, which is based on the frequency set of two-stage method of thinking , the design of association rules is decomposed into two sub-problems: ① found that the frequency set. This sub-issue is most important, the most expensive, therefore, focused on various algorithms to improve the efficiency of frequent item set discovery. ② According to the obtained frequent item sets to generate strong association rules. The algorithm uses the following two basic properties. The nature of a subset of any frequency band must be set. The nature of any non-frequent item sets 2 superset of a non-frequent item sets.

Apriori Algorithm is one of the classic data mining algorithms of association rules. One important step of Apriori is pruning, which matches every subset of the candidates with the frequent sets of previous layer, and then those infrequent sets will be removed according to one character of Apriori. This operation becomes the fact of time-cost of Apriori. A new algorithm named NPA (No Pruning Apriori) referred in this article is based on Apriori and, by means of modifying the JOIN operation; the pruning operation has been canceled. Such improvement enhances the speed of the algorithm and it brings practice application value in a certain degree. Apriori specific algorithm is as follows:

(1) L1 = {large 1-itemsets};
(2) for (k = 2; Lk-1 ≠; k + +) do begin
(3) Ck = apriori_gen (Lk-1);
(4) for all transactions t ∈ D do begin
(5) Ct = subset (Ck, t);

(6) for all candidates c ∈ Ct do
(7) C.count + +;
(8) End;
(9) Lk = {c ∈ Ck | c.count ≥ minsup}
(10) End;
(11) Answer = UkLk;

The basic idea of this algorithm is: first find all frequent sets, such as frequent item sets occur at least a predefined minimum support the same. Generated by the frequency of collection and strong association rules, these rules must satisfy minimum support and minimum confidence. Then use Step 1 to find the desired frequency set of rules generated, resulting in only a collection of items containing all the rules, in which the right side of each rule is only one, here is the rules used in the definition. Once these rules are generated, then only those greater than the minimum confidence given by the user was only to stay the rules. In order to generate all frequency sets, using the recursive method.

For many applications, the dispersion of the data distribution, it is difficult in the most detail level data find strong association rules. Although the rules drawn on a higher level may be general information, however, it is common for a user's information, but not necessarily so for another user. Therefore, data mining should provide a dig at multiple levels of functionality. Multi-level association rules mining are generally two ways: one is the single-level association rule mining algorithms directly applied to multi-level; the other is applied at different levels of different support threshold and confidence threshold. Existing multi-level association rule mining algorithm is mainly Improved Association Rules Algorithm.

4.2 Improved Association Rules in Bank Risk Assessment

Set each item I as an example, each transaction is a line in database D, together constructed incidence matrix M, M is equal to m * n, n is the total number of I, m for the database D contains the total number of Project Services. In the association matrix that contains the items that each firm, also contains a similar transaction.

Algorithm ideas, for the sum of each row of the matrix associated, the number of items calculated Affairs. Calculated to support the number of items removed is less than the number of sets to support the remaining items to our collection to find.

Input, transactional date T, the minimum support minsup count digital. Out put, the maximum frequency item set L.

- 1. C [n] = 0; / / C [n] n the maximum number of items
- 2. For each ti is Tdo {
- 3. I = | ti |
- 4. C [i] = C [i] +1
- 5.}
- 6. For i = n to 1 {
- 7. If (c [i] is greater than minsup) then {
- 8. K = i
- 9. Break

- 10.}
- 11.}
- 12. For i = k to 1 do {
- 13. Ck = {select k - item sets}
- 14. For each Ci is Ck do {
- 15. Lk = {Ci count Ci is greater than minsup}
- 16.}
- 17. If Lk is not equal to return Lk
- 18.}

From the point of example calculation, the use of matrix, only scan the database once, reducing the operation, and break the conventional approach using the first hypothesis, set out to find items from the high set, to minimize the number of data sets to improve work efficiency.

4.3 Improved Association Rules in Bank Risk Assessment

Select a month credit card records and customer information as credit card data, raw data, set the item sets C1, C2 ... Ck represent clients in T1, T2 ... Tk business consumer behavior.

First, scan personal credit card data, to obtain a matrix set, select the previous data sets.

Second, calculate the number of frequent item sets, the maximum was 7. Based on business experience, book the minimum support is 50.

According to Apriori algorithm, calculate the various supports. Assume that the most frequent item sets of 6, to obtain {C1, C2, C5, C8, C12, C13} support for the 67, {C1, C2, C5, C9, C12, C14} support for the 23, {C1, C2, C4, C10, C12, C13} support 82, which said that {C1, C2, C5, C8, C12, C13} and {C1, C2, C4, C10, C12, C13} for the maximum frequent item sets. Compared to the classic Apriori algorithm, the improved method only scans the database once, and improves work efficiency. Based on this data, you can engage in business promotion with the proposed business alliances {T1, T2, T5, T8, T12, T13} and {T1, T2, T4, T10, T12, T13}, and expand consumer groups, the maximum boost consumer spending, meaning a positive win-win business.

5 Conclusion

Domestic financial enterprise data management applications generally brings together the vast amounts of data, but can not find the relationship and business rules exists in the data to do risk prediction assessment. Therefore, the domestic financial companies need to accelerate the pace of information technology in regions of integration of customer resources, business analysis and investment decisions. This paper analyzes the risk assessment approach of banks, mainly focuses on the analysis of association rules data mining in bank risk assessment, and discusses the working principle of improved association rules algorithm genetic algorithm in commercial bank risk assessment.

We described the methods and processes of system application. We select the matrix form, only scan the database once, and use the method of selecting assumption

frequent items and numbers, find the frequent item sets through high end item sets, minimize the number of candidate data sets, greatly improve the efficiency of the algorithm.

References

[1] Fayyad, U.M., Piatetsky_Shapiro, G., Smyth, P., Uthurusamy: Advances in Knowledge Discovery and Data Mining (1996)
[2] Houtsma, M., Swami, A.: Set-oriented mining of associateion rules. IBM Almaden Research Center (1993)
[3] Keyun, H., Yu, C., Chunyi, S.: International Discovering Association Rules (1999)
[4] Agrawal, R., Yu, P.S.: Online generation of association rules. In: Proceedings of the 14th International Conference on Data Engineering, Orlando, Florida, USA (1993)
[5] Jiawei, H.: Data mining concept
[6] Btin, S.: Dynamic item counting and implication rules for market basket data (1997)
[7] Savasere, A.: An efficent algorithm for mining accosiation rules (1995)