In-Depth Research and Exploration of Financial Data Mining Technology Based on the Background of Big Data

Xiaoyu Zhou [1], Boling Wu [2]

Email: 18021214359@163.com

Reading Academ, Nanjing University of Information Science and Technology, Nanjing Jiangsu 210044, China[1]
School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong, 266100, China [2]

Abstract: Along with the large-scale application of financial institutions' management information systems, financial institutions have generated massive amounts of data in their daily use. However, this huge amount of data is not used efficiently, and financial managers only get some superficial information in their daily management by simple statistics and sorting and filtering potential methods. These financial data are accumulated from the history of daily work and management practices, so there is a large amount of useful financial data hidden in them, and it is a problem that needs to be studied in depth to find out the risks and control and prevent the occurrence of risks by mining this resource to a deeper extent.

Keywords: Big data; Data mining; Financial data analysis; System Module

1 INTRODUCTION

The world economy of the 21st century is characterized by economic globalization, data informatization, and international financialization. The role played by the informationization of financial data in daily life is also increasing, and people can conveniently conduct financial operations through the large-scale application of informationized financial data [1]. In this context financial data mining and and analysis are highly valued. With the further development of networking, the financial industry has put forward higher requirements for financial data in terms of accuracy and time effect, and urgently needs a more efficient way to obtain financial data [2]. How to process financial data quickly and accurately from the huge and complex financial information resources like the Internet is one of the dilemmas that people encounter when dealing with financial business. Therefore, it is necessary to combine data collection and networking to improve data analysis and processing capabilities [3].

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334431

2 ANALYSIS OF THE REASONING PROCESS OF THE CASE

2.1 Case Extraction

Suppose there is a case S that needs to be added to the case library, S contains N inputs $IN1, IN2, \ldots, Inn$ and M processes $P1, P2, \ldots, Pm$ and K outputs $O1, O2, \ldots, OK$. I is the information entropy of the initial case-base, while IS is the information entropy obtained after adding case S in. The expression for the information gain after adding S in is:

$$E(S) = I.IS \tag{1}$$

The greater the degree of information gain, the greater the degree of difference between the added cases and the original cases. A threshold Y is set at the top, and S is added to the case pool as long as E(S) exceeds Y. If E(s) is smaller than Y, S is dropped. The flow of the case addition is illustrated in Figure 1:

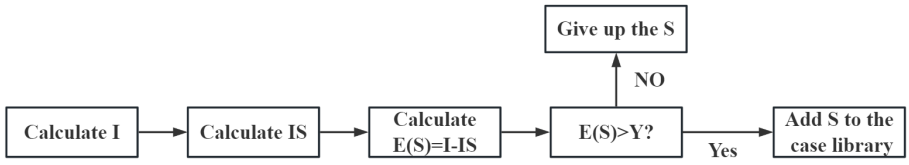

Figure 1: Case add flow chart

2.2 Case experiment verification

This project collects ten days of daily business data of a commercial bank and adds it to the case database, which contains many types of business such as asset business, liability business [4].,off-balance sheet business, etc [5]. The selected cases have very clear output in multiple business environments [6]. Matching tests were performed on five randomly taken input cases using the unoptimized algorithm as well as the optimized algorithm to test the matching time and the accuracy of the prediction, and its results are presented in Figure 2 as well as Table 1.

Table 1: Case matching time comparison table

Improved algorithm (MS)	352	638	429	389	386
Traditional Algorithm (MS)	1214	1025	1233	1017	1179

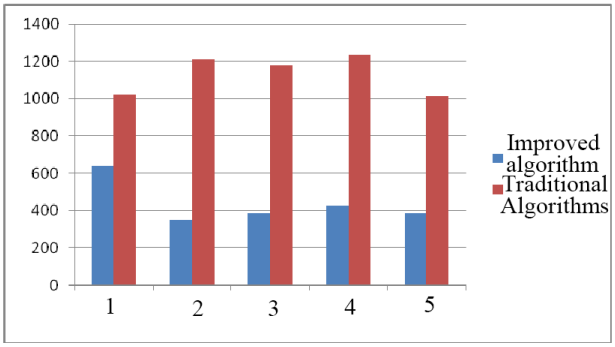

Figure 2: Case Matching Time Comparison Chart

By looking at the above graph, we can see that the optimized algorithm reduces the matching time by 60% compared to the unoptimized algorithm [7].

We randomly took 100 cases of this commercial bank as test cases for matching, in order to compare the prediction accuracy between the optimized algorithm and the unoptimized algorithm, and the test results showed that the average prediction accuracy of the optimized algorithm was 89.47% while the average prediction accuracy of the unoptimized algorithm was 84.35%.As show in figure 3.

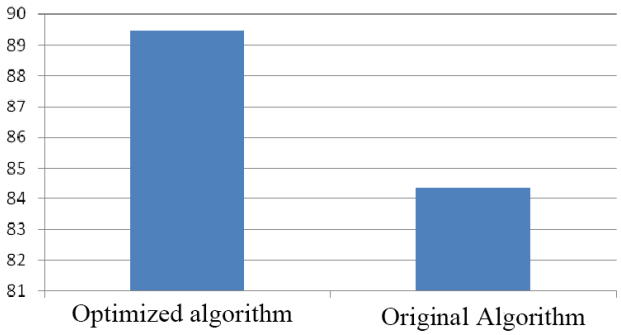

Figure 3: Case Matching Accuracy Comparison Chart

By observing the above figure, it can be concluded that the optimized algorithm has significantly reduced the matching time but also improved the prediction accuracy, which is because the algorithm has optimized the matching process by weighing too much the output of similar cases in the cluster, which improves the prediction accuracy and degree of accuracy [8].

3 A FINANCIAL ANALYSIS METHOD BASED ON WEIGHTED MULTI STOCHASTIC DECISION TREES

3.1 Calculation of attribute weights

The criticality of the attributes in the financial data warehouse varies under different mining objectives, so the criticality of each attribute should be quantitatively analysed when setting up a decision tree. This topic uses the scheme of discriminant matrix to assess the importance of attributes. In addition, since financial data are highly demanding in terms of professionalism, it is not possible to reflect the actual importance of attributes by relying only on the discriminant matrix to assess the importance of the attributes.

Define the resolution matrix 1: a diagonal matrix of $|u| \times |u|$. The definition of each of these items is $C_{ij} = \begin{cases} \{\alpha \in A | \alpha(x_i) \neq \alpha(x_j)\} d(x_i) \neq d(x_j), d(x) \in D \\ \emptyset, d(x_i) = d(x_j), d(x) \in D \end{cases}$

There is a positive correlation between the number of occurrences and the importance of the attributes in the resolution matrix; and the shorter the data item for which an attribute exists, the more critical the attribute is.

Algorithm 1 calculates the weights of financial data attributes

Initialize all $a_i \in A$ make $w(a_i) = 0$.

For each term of the diagonal array in the resolution matrix C_{jk} is calculated $w(a_i) = w(a_i) + |C_{jk}| \; a_i \in C_{jk}, 0 < k < j \leq |U|$.

3.2 Experimental verification

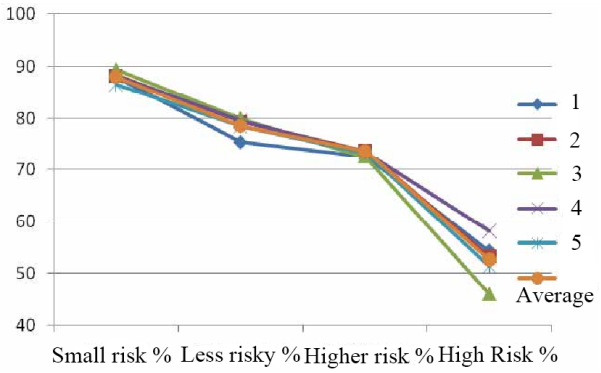

Figure 4: Multi-random decision tree classification correct rate comparison chart

As show in figure 4.The results of the experiment show that this randomized decision tree algorithm classifies companies with large risk, large risk, small risk, and small risk in terms of classification accuracy, and has been determined by bank staff to be a practical reference for predicting bank risk using this classification accuracy. However, because the number of data with a large risk level in the training dataset is relatively small, this class of branches is not

sufficiently trained, making the randomized decision tree algorithm less accurate in classifying those with a large risk level compared to other branches [9].

In each case, the same training and validation data were used to classify the risk level using the C4.5 algorithm, and the final results are presented in Table 2 below and Figure 5 below:

Table 2: C4.5 Classification correct rate comparison table

Number of verification	Small risk %	Less risky %	Higher risk %	High Risk %
1	72.56	65.37	60.28	35.47
2	73.25	66.71	62.14	37.02
3	74.53	66.37	66.21	34.28
4	73.90	65.28	63.43	42.67
5	75.69	66.27	65.71	33.83
Average	73.99	66.00	63.55	36.65

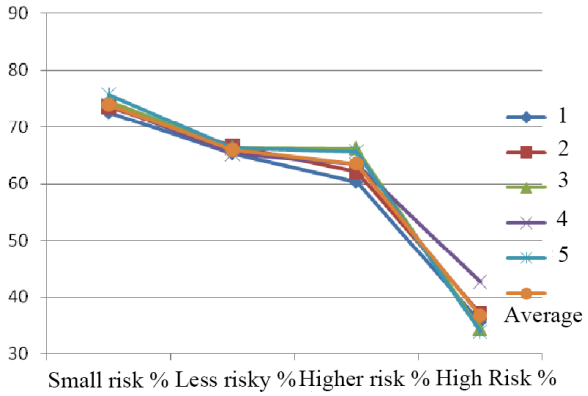

Figure 5: C4.5 algorithm classification correct rate comparison chart

The results of the experiment show that this randomized decision tree algorithm improves the classification accuracy for risk level of large, risk level of large, risk level of small, and risk level of small by a considerable amount. Similarly, because the number of risky data in the training dataset is relatively small, this class of branches is not sufficiently trained, which makes the classification accuracy of C4.5 algorithm for risky data significantly lower compared to other branches. As presented in Figure 6 below:

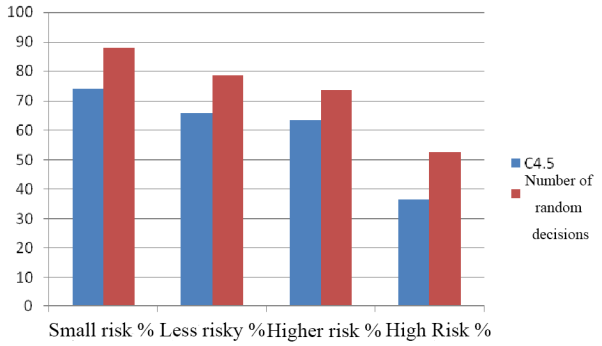

Figure 6: Comparison of the correct classification rate of the two algorithms

From Figure 6 above, it can be found that the accuracy of the randomized decision tree algorithm is even higher than that of the C4.5 algorithm, about 10% higher.

Since the training for the risk level of large is not sufficient, 300 data with the risk level of large are added to the training data set in order to improve its correctness [10]. The original random sampling is replaced with stratified sampling, and the initial data are stratified by small, small, large, and large risk level, and then random sampling is used for each stratum to ensure the number of training data with large risk level. The results of the classification using the stratified random sampling method are presented in Table 3 and Figures 7below:

Table 3: Comparison table of the correct rate of stratified sampling with multiple random decision trees

Number of verification	Small risk %	Less risky %	Higher risk %	High Risk %
1	89.13	86.32	77.35	71.31
2	90.28	84.29	76.51	72.03
3	91.37	87.16	78.67	71.58
4	89.81	85.35	79.20	70.00
5	88.90	88.62	79.67	72.69
Average	89.75	88.62	78.28	71.52

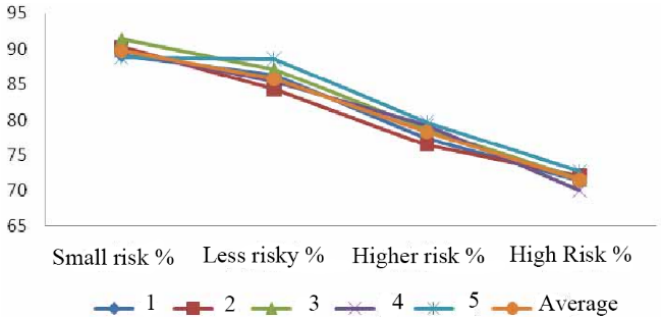

Figure 7: Multi-random decision tree stratified sampling classification correct rate comparison chart

3.3 On-page financial data extraction

Financial data presented in tabular form can be extracted from tables by pattern matching <table>, from rows by pattern matching <tr>, and from table headers and content words by pattern matching <trl> and <th>. When data segmentation is performed, the data collector needs to judge and analyze the page text features and define the segmentation mark for each level in advance. The flow of the implementation algorithm is as follows figure 8:

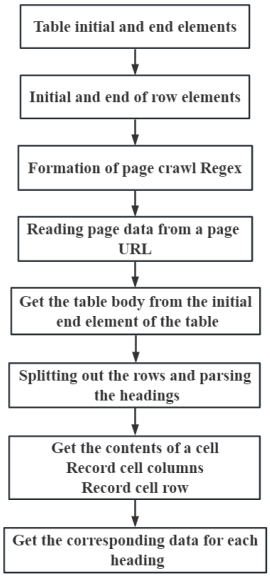

Figure 8: Algorithm implementation flow

4 CONCLUSIONS

Traditional financial data can no longer meet the existing process of globalization, diversification, and refinement of the world economy to advance, so the vast majority of data analyzed is also presented by way of reports. Along with the further increase in the volume of financial business and increasingly complex levels, this part of the financial data contains a large number of high-value data that is difficult to find through human hands, in order to collect this part of high-value financial data from the huge amount of financial data, financial data mining in the urgent need to add intelligent financial data analysis solutions [11]. In summary, this topic proposes a comprehensive class of financial data analysis solutions.

REFERENCES

[1] Dong Xinge. Analysis of the application of data mining technology in banks in the context of big data[J]. Wealth Today,2021(20):31-33.

[2] He-Ran,Huang Jin-Hui. Algorithms for data mining techniques in the context of big data[J]. Electronic Technology and Software Engineering,2019(20):141-142.

[3] Lu Jishu. Research on the application of data mining technology in customer relationship management of commercial banks in the context of big data--ZG Bank as an example[J]. Industrial Science and Technology Innovation,2023,5(01):71-74.
[4] Liu Ying. An introduction to the improvement of data mining technology application in the context of big data[J]. Technology and Innovation,2022(18):176-178.DOI:10.15913/j.cnki.kjycx.2022.18.052.
[5] Men, Xue-Lin. The use of data mining technology in management accounting in the context of big data[J]. Contemporary Accounting,2020(11):3-4.
[6] Shi Sen. Research on the application of data mining technology in the context of big data[J]. Electronic World,2020(20):126-127.DOI:10.19353/j.cnki.dzsj.2020.20.055.
[7] Tian Chuyun,Yang Suan. Research on the application of data mining technology in insurance industry in the context of big data[J]. Electronic World,2020(07):15-16.DOI:10.19353/j.cnki.dzsj.2020.07.006.
[8] Wang Lili. Application of data mining technology in the context of big data[J]. Computers and Networks,2021,47(20):45-47.
[9] Xu Mengxin,Gao Deli,Liu Jing. The use of data mining technology in file management system in the context of big data[J]. Information and Computer(Theory Edition),2021,33(02):28-30.
[10] Xie Shengjia. Research on the application of data mining technology in the context of big data era[J]. Computer Products and Circulation,2020(05):128.
[11] Yang, Yan-Ye, Liang, Xiao-Qing. Application of data mining technology in the context of big data[J]. China New Communication,2020,22(06):105.