Hindawi
Mobile Information Systems
Volume 2022, Article ID 6851616, 7 pages
https://doi.org/10.1155/2022/6851616

Review Article

Data Mining Optimization Software and Its Application in Financial Audit Data Analysis

Rui Shan,[1,2] **Xianfei Xiao**,[3] **Junwen Che**,[1] **Jingfang Du**,[1] **and Yuwu Li**[1]

[1]*Business School of Yantai Nanshan University, Longkou, Shandong 265713, China*
[2]*Future Intelligent Financial Engineering Laboratory of Shandong, Yantai, Shandong 264026, China*
[3]*Human Resources Headquarters of Nanshan Holdings, Longkou, Shandong 265706, China*

Correspondence should be addressed to Xianfei Xiao; 201823252501037@zcmu.edu.cn

Received 6 May 2022; Revised 8 June 2022; Accepted 30 June 2022; Published 13 July 2022

Academic Editor: Muhammad Muzammal

The scope of finance is very wide, data also plays a very important role in the financial industry, a small data change, and it may have a great impact on the economy. Therefore, the author proposes data mining optimization software and its application in financial audit data analysis. First, discuss the decision tree method, the main function module design of the system software, the financial analysis software method of weighted multiple random decision trees is described. To conduct verification experiments, the decision-making effect of constructing 10 random decision trees is the best. So, the author constructed a total of 10 random decision trees to analyze the data, since the tree is constructed using a random method, in order to verify the stability of decision tree classification, a total of 5 experiments were carried out, the training data set for each experiment, randomly select 1200 pieces of data from the original data set as training data, the tree is constructed by randomly selecting 12 attributes from 24 attributes. The remaining 300 pieces of data are used as verification data. As can be seen from the results, the accuracy of the random decision tree method is about 10% higher than that of C4.5. In order to improve the accuracy rate of high risk, 300 pieces of high-risk data were added to the training data set. To change the original random sampling into stratified sampling, according to the high, medium, and low risk, the original data is stratified; random sampling is used for each layer, thereby ensuring the amount of training data with high risk. The accuracy of decision tree classification is related to the number of samples of the training data, the larger the number of samples, the more accurate the classification of the constructed decision tree.

1. Introduction

The 21st century world has many characteristics such as economic globalization, data informatization, and financial internationalization, the informatization of financial data plays an increasingly important role in life, people through large-scale analysis of informatized financial data, to find important information [1]. To facilitate the processing of related financial services, data mining is a technology closely related to the future development. Through data mining, it can effectively help people get useful information in advance. Basic overview of data mining: data mining is the processing of data in large databases, from a large amount of random data, the process of extracting hidden and potentially useful information [2]. The object that this process needs to face is a large amount of business data, so it needs the help of artificial intelligence, statistics, automation, summarize the massive data, summarize the effective information, and apply the information. Data mining can effectively help people find directions, occupy market opportunities, and maximize the profits [3]. Due to the complexity of the financial industry, it needs to involve a lot of collecting and processing data [4]. Most financial banks and financial institutions provide financial services, such as personal deposit, credit card, loan business, and investment business, the complexity of these transactions and the asymmetry of information, coupled with the large number of people doing related business every day, so it will generate a lot of data and these large amounts of financial data [5]. In this massive amount of information, it contains very few valid information, and through data

mining, you can dig out the effective information that exists, the technical architecture of the bank's unified data mining analysis platform is shown in Figure 1.

2. Literature Review

In response to this research question, Kaffash and Marra and others in modern audit work, when traditional manual auditing methods do not meet the requirements, proposed modern auditing to provide new methods and ideas, improved audit quality, and resolved the audit risk [6]. Lausch et al. and others creatively proposed an audit application model based on data mining, pointed out the interrelationship, and workflow of data mining and audit work [7]. Zhang et al. and others proposed the feasibility of two commonly used data mining techniques (association analysis and cluster analysis) in auditing. Due to the lack of specific actual data verification, even though the examples cited are representative, they are only theoretical studies [8]. Chaovalitwongse et al. and others under the modern auditing technology and information environment conducted exploratory research. The author first studied the audit work, in the context of the development of computer technology, then he proposed that when faced with massive amounts of raw data for review and analysis, can use a combination of real-time online auditing methods and data mining techniques, this can improve audit efficiency, resolve audit risks, and ensure audit quality [9]. Hu et al. and others pointed out that in the field of auditing, the point of intervention in engineering thinking, using engineering techniques, created a new type of audit service, some special problems in the audit are dealt with well [10]. Beiles et al. and others comprehensively analyzed data mining technology, when faced with massive amounts of audited data, and the process of its realization [11]. In the research of classification technology, Syrimi and Hiwarkar in order to process a large amount of high-dimensional data, attempt to construct its set theory system [12]. Ratcliffe et al. and others in the process of knowledge discovery, combine rough set and fuzzy set theory [13]. Lehmann et al. constructed fuzzy system identification methods and fuzzy system knowledge models and constructed an intelligent expert system [14]. When Aitken studied the data acquisition of financial websites, same as general data capture, through the XMLHTTP object provided by Microsoft, get the overall data of financial webpages. The grabbing process is shown in Figure 2 [15]. On the basis of current research, the author proposes data mining optimization software and its application in financial audit data analysis, according to relevant corporate indicators, a company with a higher risk has a credit default. The risky enterprise is that although there is no default, but companies that are at risk of deteriorating financial conditions. Low-risk companies have good financial status, and there is no credit default. Use training data to build a random decision tree, use the verification data to verify the built decision tree, finally, the classification correctness of the decision tree for each type of data is recorded. It can be seen from the experimental results that, the random decision tree algorithm has a higher accuracy rate for classification with low risk, medium risk, and high risk, through the confirmation of the personnel of the banking institution, the correct rate of this classification has certain reference significance for the prediction of bank risk. However, the algorithm has a relatively low accuracy rate for classification with high risk, the main reasons for analysis are: in the training data set, the amount of data with high risk is small, the training of this type of branch is not sufficient. Effectively improve the existing distance-based method to improve its matching efficiency, improve its forecast accuracy.

3. Methods

3.1. Overview of Decision Tree Method. Decision tree algorithm is the most commonly used algorithm in classification data mining. Decision tree algorithm is a classification process of fitting problems through a tree structure. In the constructed tree, each level corresponds to a classification attribute, call it a split attribute, the nodes in this layer correspond to different values of the attribute, the corresponding data under the value of the attribute is stored in the node, each leaf node saves different types of label attributes, the probability distribution under this branch, when performing classification, the predicted value of the class label attribute of the data falling into a certain leaf node, it is the attribute value of the class label with the highest probability in the leaf node. The most widely used algorithms in decision-making algorithms are ID3 algorithm and C4.5 algorithm [16]. A decisive factor that affects the construction of the decision tree, it is the choice of classification attributes for each layer, ID3 algorithm before the construction of each layer, calculate the information entropy of different classification attributes, and then calculate its information gain, the classification attribute with the largest information gain is selected as the split attribute of this layer. PE and NE, respectively, represent the positive example set and the negative example set, which together form the training set. 'PE, PE' and 'NE, NE,' respectively, represent a subset of the positive example set and the negative example set [17]. Among:

Information entropy:

$$H(U) = -\sum P(u_i)\log_2 P(u_i). \tag{1}$$

Probability of category:

$$P(u_i) - \frac{|u_i|}{|S|}. \tag{2}$$

$|S|$ represents the total number of example sets S, $|ui|$ represents the number of examples of category u_i.

Calculation of conditional entropy:

$$H(U|V) = -\sum_j P(V_j) \sum_i P\left(\frac{U_i}{V_j}\right)\log_2 P\left(\frac{U_i}{V_j}\right). \tag{3}$$

When the attribute Al takes the value v_j, the conditional probability of category u_i is

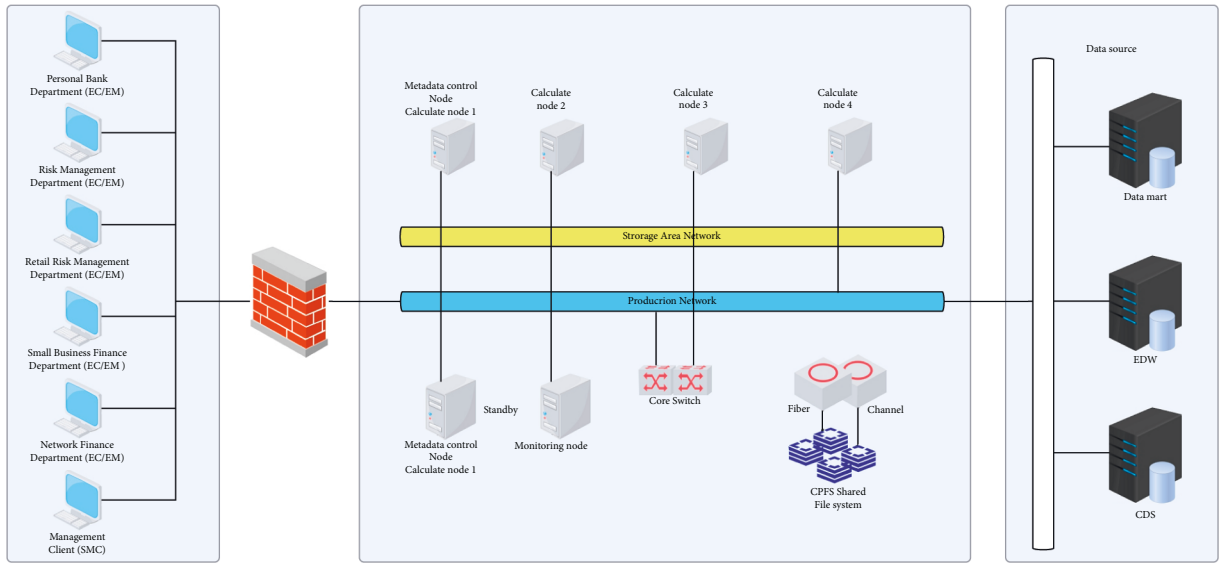

Figure 1: The technical architecture of the bank's unified data mining analysis platform.

$$P\left(\frac{U_i}{V_j}\right) = \frac{|u_i|}{|V_j|}. \tag{4}$$

ID3 algorithm introduces information entropy into the tree construction process, quantify the splitting properties, the method is simple and easy to implement, the tree construction is very simple, and the constructed tree is small in scale, the classification effect is good, without any relevant domain knowledge, strong versatility, and adaptability. So, the current D3 algorithm has been widely used in different fields.

3.2. Design of Main Functional Modules of System Software. According to the results of the preliminary demand analysis, the main functional modules of the system software are: task management module, data analysis module, and data management module [18]. The main functions of each module are designed as follows:

(1) The main functions of the task management module: create new tasks, execute tasks, and task acceptance [19].

New task: financial institution users submit task requests through the system, system administrators based on requests and the results of communication with users of financial institutions, query the evaluation data of each analyst, and assign the task to the most suitable analyst, the analyst determines to accept the task, and then starts the execution of the analysis task [20].

Perform tasks: after the analyst receives the user's request, the task starts to execute, the analyst calls the analysis method of the data analysis module to analyze the task, after the analysis is completed, submit the analysis results to the financial institution user, and according to the feedback results of users of financial institutions, modify the execution of tasks. Financial institution users can apply to change analysts during the task execution process, the system administrator confirms whether to replace according to the execution of the task. Every information exchange of task execution must be recorded [21].

Task acceptance: if users of financial institutions are satisfied with the analysis results, they can submit tasks for acceptance, and evaluate the analysis, the system saves related data interfaces and analysis methods and environmental information, for the next analysis.

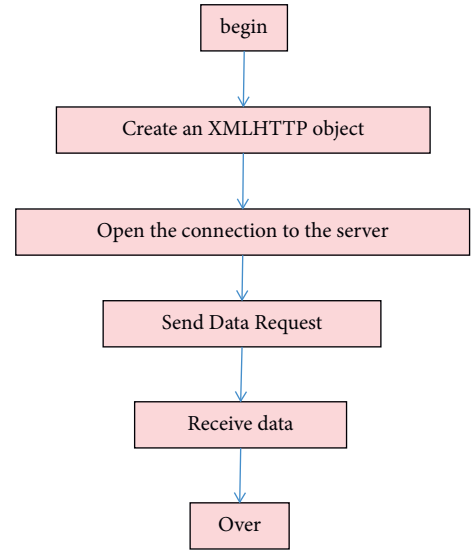

Figure 2: Flow chart of data acquisition algorithm for financial websites.

(2) Data analysis module, which is the core module of the department

Financial data analysis, including two levels of individual analysis and regional analysis, among them, individual analysis analyzes the individual operating data and external environment of a single financial institution, derive the potential financial risk factors of the institution, analyze the operating data and external environment of all financial institutions in a certain area, and combined with the results of individual analysis, the regional financial risk assessment software in a certain area can be obtained [22].

(3) System software data management module

The data administrator is responsible for the maintenance of the data in the system, the method of collecting system data is done by the system analyst, but the analyst is not responsible for maintaining the data, this part of the work is done by the data administrator, its main work includes the use of related interfaces, perform operations such as data cleaning, data backup and recovery, and data dumping.

3.3. Financial Analysis Software Method Based on Weighted Multiple Random Decision Trees. Financial data analysis contains many classification and prediction problems, such as: when a credit customer makes a loan application, financial institutions need to classify according to their financial indicators and past credit conditions, then decide whether to approve its credit request. In addition, for financial risk prediction, classification methods can be used in the determination of violations, at present, the commonly used classification method is the decision tree method, among which, compared with traditional decision trees, random decision trees have the advantages of fast analysis speed, high accuracy rate, and strong robustness. Based on this, the method of random decision tree is introduced into the model, realizing the classification problem of financial data [23].

3.3.1. Attribute Weight Calculation. For the attributes in the financial data warehouse, under different mining goals, the degree of importance is different, so before building the decision tree, the importance of each attribute needs to be quantitatively analyzed, at present, the commonly used methods for determining the importance of attributes are: the method based on the discernibility matrix and the method based on information entropy. The author uses the discernibility matrix method to analyze the importance of attributes. In addition, due to the high professionalism of financial data, only rely on the discrimination matrix to analyze the importance of attributes, cannot fully fit the true importance of the attribute, therefore, artificial weights are introduced to modify and intervene the weights of the discernibility matrix, in order to further increase the accuracy of attribute weight calculation.

Define the discrimination matrix: the discrimination matrix of an information system is a diagonal matrix of $|U| \times |U|$. Each of these is defined as:

$$C_{ij} = \begin{cases} \{a \in A | a(x_i) \neq a(x_j)\}, & d(x_i) \neq d(x_j), d(x) \in D, \\ \varphi, d(x_i) = d(x_j), d(x) \in D. \end{cases} \tag{5}$$

The more the attribute appears in the discernibility matrix, the more important the attribute. The shorter the data item that contains the attribute, the more important the attribute is.

3.3.2. Calculation of Financial Data Attribute Weights. Order $w(a_i) = 0$ for all $a_i \in A$ at the beginning.

Calculate $w(a_i) + |c_{jk}|$ for each item c_{jk} of the lower diagonal matrix in the discernibility matrix, $a_i \in c_{jk}, 0 < k < j <= |U|$. Where $|A|$ is the cardinality of all attributes, and $|c_{jk}|$ is the cardinality of c_{jk} in the discernibility matrix. After the system gives the weight, the weight of the system can be modified manually, and a correction coefficient is introduced for this purpose, $w'(a_i)$, $-1 <= w'(a_i) <= 1$, in order to increase the weight of a_i, set $W'(a_i)$ to a positive value, otherwise, set it to a negative value, then the weight $w_{a_i} = w(a_i) + w'(a_i)$ of attribute a_i.

4. Results and Analysis

4.1. Verification Experiment. The verification data comes from the financial data of 1500 corporate customers of a commercial bank, due to the attributes in the financial information data sheet provided by the bank, is based on the properties of the transaction database, so, perform attribute transformation on it, form 24 attributes that can reflect corporate financial indicators.

First of all, according to the relevant indicators of the company, a company with a higher risk has a credit default. The risky enterprise is that although there is no default, but companies that are at risk of deteriorating financial conditions. Companies with low risk have good financial status and no credit default. Research shows that the decision-making effect of constructing 10 random decision trees is the best [24]. So, the author constructed a total of 10 random decision trees to analyze the data, since the tree is constructed using a random method, in order to verify the stability of decision tree classification, a total of 5 experiments were carried out, in the training data set of each experiment, 1200 pieces of data are randomly selected from the original data set as the training data, a tree is constructed by randomly selecting 12 attributes from 24 attributes [25]. The remaining 300 pieces of data are used as verification data [26]. Use the training data to build a random decision tree, and use the verification data to verify the built decision tree, finally, the classification correctness of the decision tree for each type of data is recorded. The experimental results are shown in Figure 3:

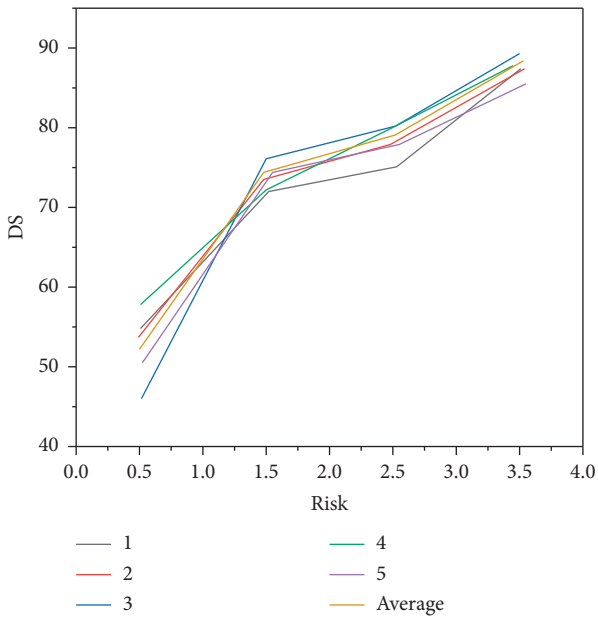

Figure 3: Comparison of classification accuracy of multiple random decision trees.

Table 1: C4.5 classification accuracy comparison table.

Number of verifications	High risk	Higher risk	At risk	Low-risk
1	35.51	60.12	65.73	72.65
2	37.01	62.35	66.17	73.52
3	34.12	66.41	66.37	74.35
4	42.15	62.99	65.82	73.09
5	33.75	65.78	66.72	75.96

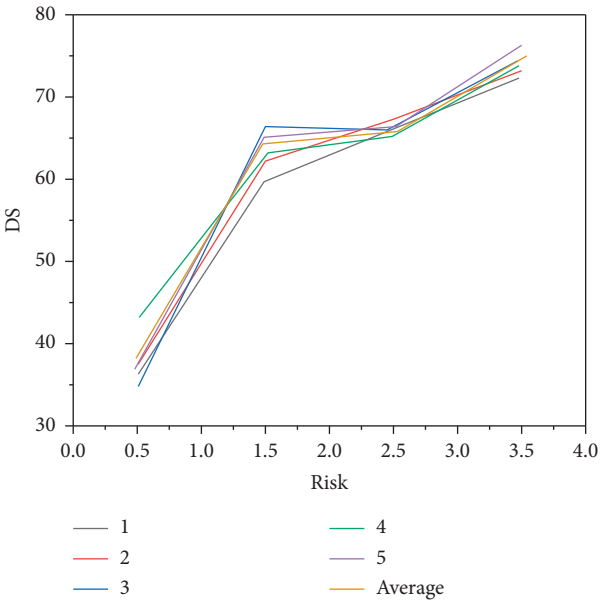

Figure 4: C4.5 algorithm classification accuracy comparison chart.

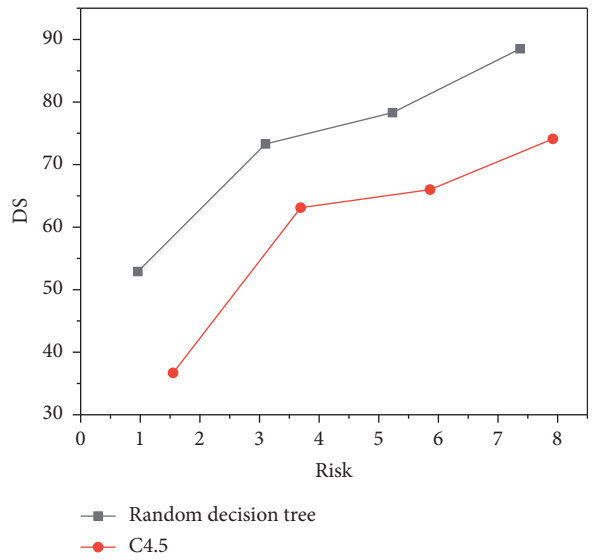

Figure 5: Comparison of classification accuracy of two algorithms.

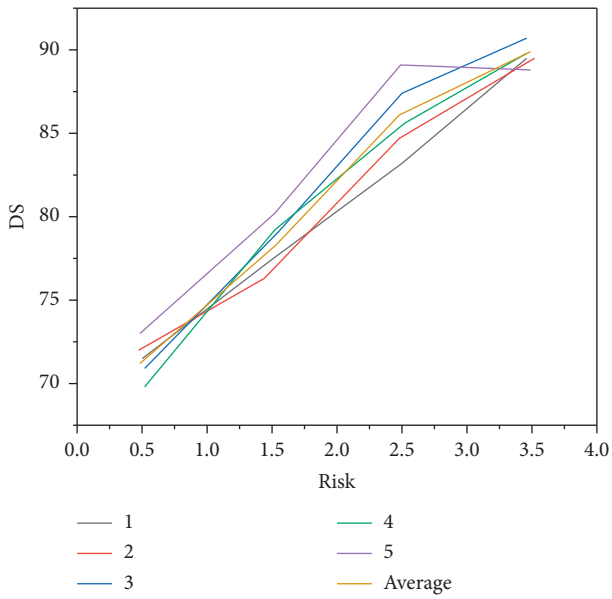

Figure 6: Comparison of accuracy rate of stratified sampling classification with multiple random decision trees.

It can be seen from the experimental results that, the random decision tree algorithm has a higher accuracy rate for classification with low risk, medium risk, and high risk, through the confirmation of the personnel of the banking institution, the classification accuracy rate, the prediction of bank risk has certain reference significance. However, the algorithm has a relatively low accuracy rate for classification with high risk, the main reason for the analysis is: the amount of data with high risk in the training data set is small, the training of this type of branch is not sufficient. Using the same training and testing data every time, the results of classification using the C4.5 algorithm are shown in Table 1 and Figure 4:

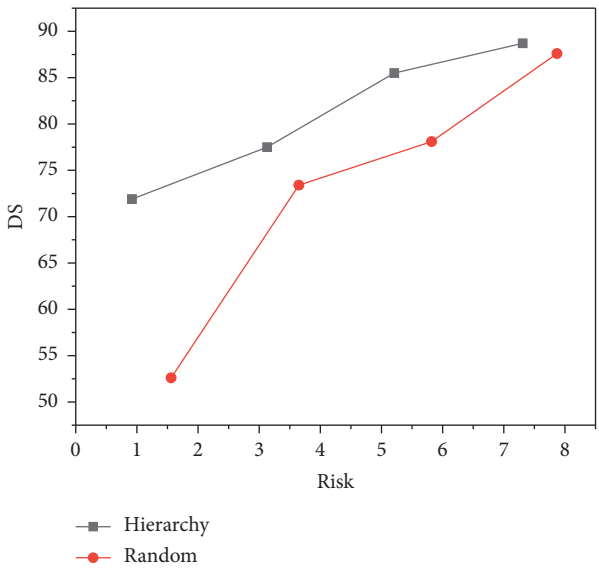

Figure 7: Comparison of classification accuracy between stratified sampling and random sampling.

It can be seen from the experimental results that, the random decision tree algorithm has a higher accuracy rate for classification with low risk, medium risk, and high risk. Similarly, the algorithm has a relatively low accuracy rate for classification with high risk, the main reason is also: the amount of data with high risk in the training data set is small; this results in insufficient training in this type of branch, as shown in Figure 5:

As can be seen from Figure 5, the accuracy of the random decision tree method is about 10% higher than that of C4.5. In order to improve the accuracy rate of high risk, 300 pieces of high-risk data were added to the training data set. Change the original random sampling into stratified sampling, stratify the original data according to the high, medium, and low risk, random sampling is used for each layer, thereby ensuring the amount of training data with high risk. The classification results after stratified random sampling are shown in Figures 6 and 7:

It can be seen that after the use of stratified sampling, the accuracy rate of risk is increased by 10%, mainly because stratified sampling increases the number of samples with high risks. Therefore, the accuracy of decision tree classification is related to the number of samples of the training data, the larger the number of samples, the more accurate the classification of the constructed decision tree.

5. Conclusion

The article proposes data mining optimization software and its application in financial audit data analysis, introduce the multi-random decision tree method into financial data analysis, and verified its validity with the operating data of a certain bank, the results show that this method can effectively analyze financial data, but the training data has a certain influence on the result of decision tree training, so to ensure the classification results of this method, it is necessary to use stratified sampling and other methods to deal with the training data. Understand the basic theory of data mining in the future, carry out relevant analysis according to the applicable theories of different data, do a good job in demand analysis and system design. Combining the specific application of data mining in financial data analysis, carry out effective thinking and reference, realize the long-term scientific development of the financial industry.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Yang, H. Cheng, H. Wu, and T. Wang, "Electromagnetic optimization of the integrated magnetorheological jet polishing tool and its application in millimeter-scale discontinuous structure processing," *Applied Optics*, vol. 56, no. 11, pp. 3162–3170, 2017.

[2] B. T. Russell, D. S. Cooley, W. C. Porter, B. J. Reich, and C. L. Heald, "Data mining to investigate the meteorological drivers for extreme ground level ozone events," *Annals of Applied Statistics*, vol. 10, no. 3, pp. 1673–1698, 2016.

[3] Mach, "Reduction of optimization problem by combination of optimization algorithm and sensitivity analysis," *IEEE Transactions on Magnetics*, vol. 52, no. 3, pp. 1–4, 2016.

[4] N. Chen, Z. Ren, D. Li, E. Y. Lam, and G. Situ, "Analysis of the noise in backprojection light field acquisition and its optimization," *Applied Optics*, vol. 56, no. 13, pp. F20–F26, 2017.

[5] Y. Zeng, Z. Zhang, and A. Kusiak, "Predictive modeling and optimization of a multi-zone hvac system with data mining and firefly algorithms," *Energy*, vol. 86, pp. 393–402, 2015.

[6] S. Kaffash and M. Marra, "Data envelopment analysis in financial services: a citations network analysis of banks, insurance companies and money market funds," *Annals of Operations Research*, vol. 253, no. 1, pp. 307–344, 2017.

[7] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data - new perspectives for data analysis in environmental research," *Ecological Modelling*, vol. 295, pp. 5–17, 2015.

[8] Z. Zhang, A. Kusiak, Y. Zeng, and X. Wei, "Modeling and optimization of a wastewater pumping system with data-mining methods," *Applied Energy*, vol. 164, pp. 303–311, 2016.

[9] W. A. Chaovalitwongse, C. A. Chou, Z. Liang, and S. Wang, "Applied optimization and data mining," *Annals of Operations Research*, vol. 249, pp. 1–3, 2017.

[10] J. Hu, J. Fang, Y. Du, Z. Liu, and P. Ji, "Application of pls algorithm in discriminant analysis in multidimensional data mining," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 6004–6020, 2019.

[11] C. B. Beiles, C. Retegan, and G. J. Maddern, "Victorian audit of surgical mortality is associated with improved clinical outcomes," *ANZ Journal of Surgery*, vol. 85, no. 11, pp. 803–807, 2015.

[12] E. Syrimi and P. Hiwarkar, "G547(p)the use of rasburicase for patients at risk for tumour lysis syndrome," *Archives of Disease in Childhood*, vol. 101, no. 1, pp. A324–A325, 2016.

[13] C. Ratcliffe, A. Abelian, and R. Reynolds, "G537(p)prospective re-audit of central line associated bloodstream infections on the neonatal unit following guideline implementation," *Archives of Disease in Childhood*, vol. 100, no. 3, pp. A236–A237, 2015.

[14] J. Lehmann, N. Miri, P. Vial et al., "MO-D-213-08: remote dosimetric credentialing for clinical trials with the virtual EPID standard phantom audit (VESPA)," *Medical Physics*, vol. 42, no. 6, 2015.

[15] R. J. Aitken, "Western australian audit of surgical mortality," *Medical Journal of Australia*, vol. 183, no. 10, pp. 504–508, 2015.

[16] R. W. Lynch, A. M. D. Churchhouse, A. Protheroe, and I. D. R. Arnott, "Predicting outcome in acute severe ulcerative colitis: comparison of the Travis and Ho scores using UK IBD audit data," *Alimentary Pharmacology & Therapeutics*, vol. 43, no. 11, pp. 1132–1141, 2016.

[17] D. Krušič, D. Brilej, C. Currie, and R. Komadina, "Audit of geriatric hip fracture care – a slovenian trauma center analysis," *Wiener Klinische Wochenschrift*, vol. 128, no. S7, pp. 527–534, 2016.

[18] E. A. Knapp, C. Nau, S. Brandau et al., "Community audit of social, civil, and activity domains in diverse environments (cascadde)," *American Journal of Preventive Medicine*, vol. 52, no. 4, pp. 530–540, 2017.

[19] P. Beckett, A. Khakwani, R. Hubbard et al., "P104results of the first analysis of national lung cancer audit data based on cancer registration data," *Thorax*, vol. 71, pp. A139–A140, 2016.

[20] J. L. Vincent, U. Jaschinski, X. Wittebole, J-. Y. R. Lefrant, and Y. Sakr, "Worldwide audit of blood transfusion practice in critically ill patients," *Critical Care*, vol. 22, no. 1, p. 102, 2018.

[21] L. Torcelpagnon, V. Bauchau, P. Mahy et al., "Guidance for the governance of public-private collaborations in vaccine post-marketing settings in europe," *Vaccine*, vol. 37, no. 25, pp. 3278–3289, 2019.

[22] G. P. Rubin, C. L. Saunders, G. A. Abel, S. McPhail, G. Lyratzopoulos, and R. D. Neal, "Impact of investigations in general practice on timeliness of referral for patients subsequently diagnosed with cancer: analysis of national primary care audit data," *British Journal of Cancer*, vol. 112, no. 4, pp. 676–687, 2015.

[23] G. Buston, J. Nicholson, and H. Satish, "G454(p)quantity of patient contact with a paediatric diabetes service – is there correlation with hba1c?" *Archives of Disease in Childhood*, vol. 101, p. A270, 2016.

[24] S. Baxter, K. Sanderson, A. Venn, P. Otahal, and A. J. Palmer, "Construct validity of sf-6d health state utility values in an employed population," *Quality of Life Research*, vol. 24, no. 4, pp. 851–870, 2015.

[25] D. C. Howlett, K. J. Drinkwater, N. Mahmood, J. Illes, J. Griffin, and K. Javaid, "Radiology reporting of osteoporotic vertebral fragility fractures on computed tomography studies: results of a UK national audit," *European Radiology*, vol. 30, no. 9, pp. 4713–4723, 2020.

[26] J. L. M-. Ocuin, M. S. Zenati, L. M. Ocuin et al., "Failure to treat: audit of an institutional cancer registry database at a large comprehensive cancer center reveals factors affecting the treatment of pancreatic cancer," *Annals of Surgical Oncology*, vol. 24, no. 8, pp. 2387–2396, 2017.