Hindawi
Journal of Sensors
Volume 2022, Article ID 2192186, 10 pages
https://doi.org/10.1155/2022/2192186

Research Article

Application of Data Mining Technology in Financial Intervention Based on Data Fusion Information Entropy

Cong Gu

College of Science, Zhongyuan University of Technology, Zhengzhou, 450007 Henan, China

Correspondence should be addressed to Cong Gu; gucong@zut.edu.cn

Received 19 November 2021; Revised 7 December 2021; Accepted 13 December 2021; Published 13 January 2022

Academic Editor: Gengxin Sun

Finance, as the core of the modern economy, supports sustained economic growth through financing and distribution. With the continuous development of the market economy, finance plays an increasingly important role in economic development. A new economic and financial phenomenon, known as financial intervention, has emerged in recent years, which has created a series of new problems, promoting the rapid increase both in credit and investment and causing many problems on normal operation of financial bodies. In the long run, it will inevitably affect the stability and soundness of the entire economic and financial system. In order to maximize the effect of financial intervention, in response to the above problems, this article uses a series of US practices in financial intervention as the survey content, combined with the loan data provided by the US government financial intervention department, and mines the data of the general C4.5 algorithm of the decision tree algorithm. Generate a decision tree and convert it into classification rules. Next, we will discover the laws hidden behind the loan data, further discover information that may violate relevant financial policies, provide a reliable basis for financial intervention, and improve the efficiency of financial intervention. Experiments show that the method used in this article can effectively solve the above problems and has certain practicability in fiscal intervention. With stratified sampling, the risky accuracy rate increased by 10%, probably because stratified sampling increased the number of high-risk samples.

1. Introduction

The United States is considered to be the world's most free market [1], but no country in the world has a free market economy that is completely laissez-faire and free from government regulation, and the United States is no exception [2]. In fact, American government intervention in the economy is to ensure that the market can operate more healthily and that market players can compete more fairly and freely [3]. The financial system of the United States is a financial system dominated by the capital market. Because of the normative system, well-developed financial institutions and financial instruments have formed a developed capital market by virtue of the world's leading international monetary status of the United States dollar [4]. The market should be determined by the laws of the market, not determined and controlled by administrative orders [5]. The free competition of market entities under equal conditions is very important. Antimonopoly is because monopoly harms free market competition [6]. Unified financial legislation is because fraud and misleading harm the free market competition. The way the United States handles the economic crisis shows that [7], administrative intervention is an effective way for the country to emerge from the crisis, and practice has proved this. Government intervention in the economy is inevitable for the development of market economy [8]. It is also a good remedy for "market failure" and "market self-defeat" in the process of the development of market economy [9–10]. Market mechanism and government intervention have their own time and space [11–12], which cannot be ignored and replaced [13]. A government should perform its coercive intervention management function in the economy during dramatic market changes and economic crises [14] with the purpose of curbing the damage to society caused by harmful behaviors resulting from dramatic market changes and economic crises [15].

The most basic characteristics of data mining include a large amount of data [16], which is to discover unknown

and hidden information, extract valuable information, and use this kind of information to make important decisions [17–20]. Data mining is the process of extracting useful information from data and using it to make more appropriate decisions. The key to data mining can be divided into three parts: data, information, and decision-making [21]. Data is the basis of all mining [22], but it is only when we mobilize them or convert them into useful information that they are most valuable [23–25]. It is not enough to simply obtain information [26], and it is not what data mining requires [27]. The information obtained in the decision-making application is the ultimate goal of obtaining information. Therefore, the ultimate goal of data mining is to extract useful information from data to improve the efficiency of decision-making and make more appropriate decisions. In the past few years, data mining has been used in many industries to help senior managers make important and appropriate decisions. For example, different data mining methods can be used in the banking industry to solve and help the difficulties encountered in the business process of bank cards, credit, etc. Use these advanced computer technologies to enhance or improve their decision-making security and efficiency.

The financial market is producing huge amounts of data. Analyzing these data, explaining valuable information and helping to make financial decisions are great opportunities and challenges for data mining. The essence of many financial theories is to study how to construct a prediction model which is in line with the reality and minimize the prediction error. However, traditional financial analysis and theory, the prediction models used are often established on some harsh assumptions, and the form is a model of some simple mathematical expressions. Although this model is simple, it has good interpretability and comprehensibility, but it damages the accuracy of prediction to some extent. Data mining technology has broken this limitation in some respects. Through the analysis of the characteristics of financial data, we can see its advantages more clearly. Data mining technology is produced under the background that the database cannot predict the development trend of data, and its concept was first proposed at the 1989 International Joint Conference on Artificial Intelligence (IJCAI). Its significance is the process of extracting hidden and potentially useful information and knowledge from a large amount of incomplete and noisy, ambiguous, and random practical application data. Data mining is a new information processing technology. Its main function is to extract, transform, analyze, and model a large amount of data in the database. The process of data mining is also called the process of knowledge discovery. This is a broad academic subject.

In this paper, when studying the problem of financial intervention strategies, the existence of various irregular noises in the data can cause serious interference to the experiment. In order to avoid this situation and realize the hidden laws of data, this paper adopts an effective two-way cohesive information entropy data analysis method to establish a relevant model, which can discover the hidden information and patterns in financial data and help government financial departments to make correct intervention decisions. Under the support of information entropy theory, a simulation model based on two-way clustering is proposed for simulation. After extensive analysis and theoretical demonstration, the results show that the multichannel clustering algorithm has obvious effect on improving the accuracy of data analysis, which provides a strong scientific basis for the formulation of the financial intervention policy of the modern American government. In view of the fact that traditional clustering algorithms can only deal with single attribute data and cannot deal with the clustering problem of mixed attribute data well, and that most of the current clustering algorithms of mixed attribute data are sensitive to initialization and cannot deal with arbitrary shape data, a spectral clustering algorithm of mixed attribute data based on information entropy is proposed to deal with mixed type data. Firstly, a new similarity measurement method is proposed. The traditional similarity matrix is replaced by the combination of the Gaussian kernel function matrix composed of numerical data in spectral clustering algorithm and the influence factor matrix composed of new information entropy-based classification data. The new similarity matrix avoids the conversion and parameter adjustment between numerical attribute and classification attribute data. Then, the new similarity matrix is applied to spectral clustering algorithm to process arbitrary shape data, and finally, the clustering results are obtained.

2. Proposed Method

2.1. Basic Technology of Data Mining. Following years of development, it has been gradually matured the data mining technology. There are commonly used data mining techniques and algorithms such as decision trees, neural networks, rough sets, association rules, cluster analysis, regression analysis, genetic algorithms, and rough set algorithms. Here, the focus will be on clustering analysis, association rules, and regression analysis algorithms in line with the application area of this paper. Figure 1 is a display of several common data mining methods.

2.1.1. Cluster Analysis. Among them, cluster analysis plays a role in data mining in the following aspects: First, preprocessing steps for other algorithms, and then, these algorithms are generated into new clusters and processed; second, to analyze each cluster, mainly to analyze specific clusters; and third, explore and process some relatively independent data. However, it is often ignored when mining some relatively independent data.

(1) Split Method. If a database containing n data objects or tuples is provided, the analytical method can construct c data partitions, and each partition has its own representative cluster "$c < n$". As a general rule, divisive criteria (such as distance) are used to make objects in the same cluster "similar" and to make objects in different clusters "different." It is mainly used to find spherical clusters. These are mostly used for small- and medium-sized databases. For the purpose of better management and processing of data in clusters, some new partitioning methods are urgently needed.

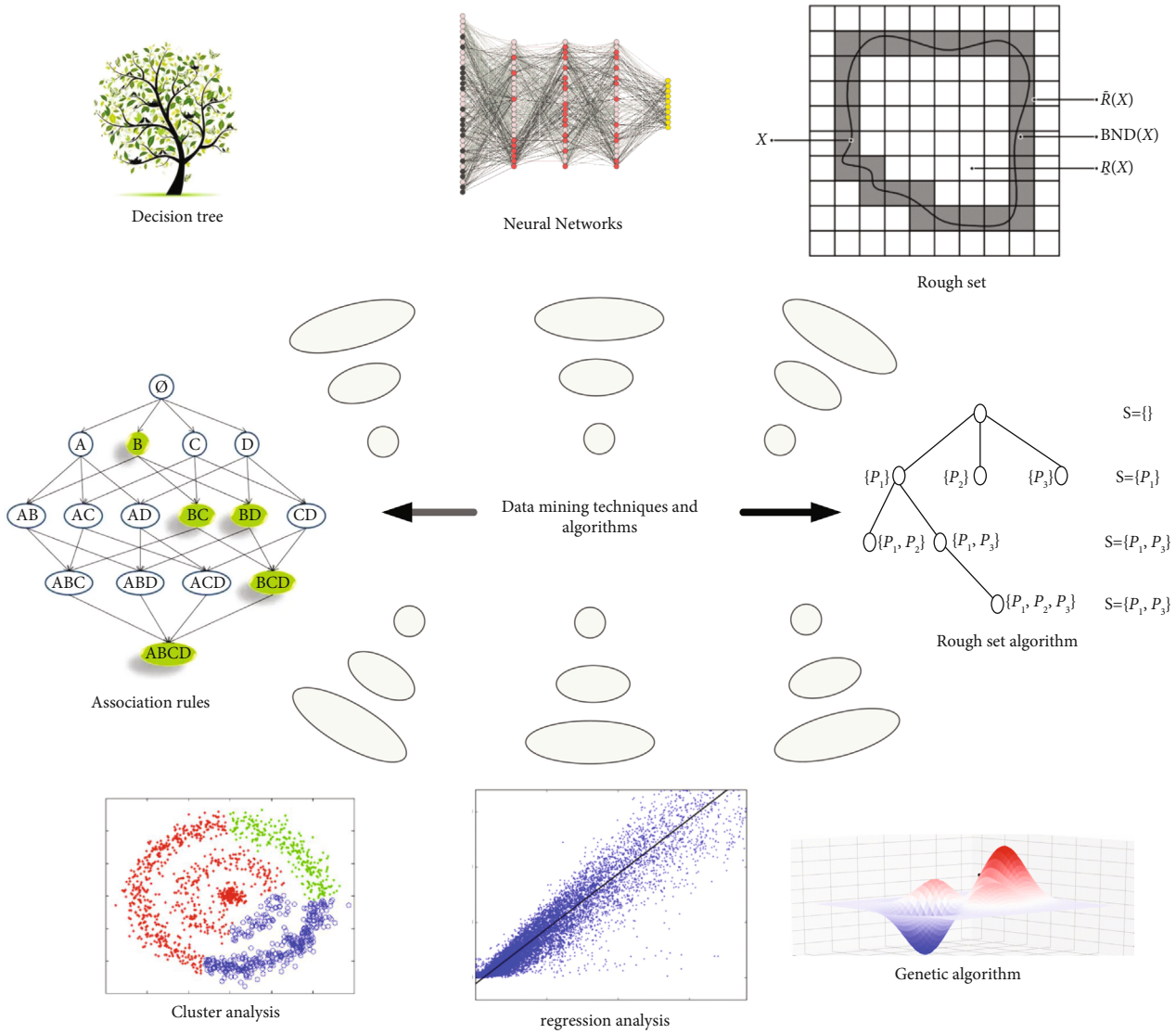

Figure 1: Several common data mining techniques.

(2) Stratification. The hierarchical method decomposes the collection of specific data objects hierarchically. According to whether the hierarchical decomposition is bottom-up or bottom-up, the hierarchical clustering technique can be divided into agglutination and segmentation. The disadvantage of hierarchical clustering is that it cannot be restored after the steps are completed, so the errors are corrected.

2.1.2. Association Rules. Association can be divided into simple association and time association. The most commonly used association rule algorithm is the Apriori algorithm proposed by R. Agrawal. Even using candidate itemsets to search for frequent itemsets, mining itemset with frequent Boolean correlation rules is the most influential algorithm.

(1) Find all frequency sets that are at least the same as the predefined minimum supported frequency

(2) Use the frequency set found in the first step to generate the target rule, and generate all the rules that only include the setting items. There is only one correct part of each rule. The definition of the intermediate rule is used here

Apriori algorithm will generate more candidate sets and may need to scan the database repeatedly. This is where the Apriori algorithm is insufficient.

2.1.3. Regression Analysis

(1) Simple Linear Regression Analysis. It is possible to determine the linear equation with a high correlation between the dependent vector a and the independent variable B if they are found to be highly correlated, with a view to making all data points as close in approximation to a straight line as possible. The model can be expressed as follows.

$$A = x + yB. \tag{1}$$

(2) Multivariate Linear Regression Analysis. What we usually see more often is that a single dependent variable corresponds to multiple independent variables. This corresponding mode is called regression. Its performance is as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k. \quad (2)$$

a represents the intercept, and $b_1, b_2, b_3, \cdots, b_k$ represents the correlation coefficient.

(3) Analysis of Nonlinear Regression Data. For linear regression problems, the sample points fall on or near a straight line in space, so a linear function can be used to represent the corresponding relationship between independent variables and dependent variables. However, in some applications, the relationship between variables is in the form of curve, so it is impossible to express the corresponding relationship between independent variables and dependent variables by linear functions, but it needs to be expressed by nonlinear functions.

2.2. Decision Tree Algorithms. It is closer to the objective function. Both leaf node classification and instance classification are performed mainly at the basis of the arrangement of nodes. On each node corresponding to one possible case, a root of a tree node is started; its attributes are measured; then, the node is changed according to its corresponding value.

2.2.1. ID3 Algorithm. On the basis of the ID3 algorithm that the attribute selection metric is the information gains when selecting on the best attribute as each node. And the measure is based on the pioneering work of C.E. in the study of information value or information theory by scientists of C.E. the Shannon:

We first compared the growth of each type of information. To choose the attribute from which the highest information is gaining (for example, maximum extraction compression) one of the tree points.

The second step is to branch according to the different values of the root node and then establish the lower nodes and branches for each branch.

The third step is to repeat the first and second steps and stop branching when the data contained in the subset are of the same category.

In this way, a decision tree can be obtained and used to classify test samples.

For the calculation description of information gain value, let D be a set of training data and define m different $C_i ni = 1n2n \cdots, mn$. The expected information for a given training data classification is given by the following formula:

$$\inf\ o(D) = -\sum_{i=1}^{m} P_i \log_2(P_i). \quad (3)$$

Note that the logarithmic function bottoms 2 because of information binary encoding.

Now, suppose you want to divide the tuples in D by attribute A, where attribute A has $V\{a_1, a_2, a_3, \cdots, a_V\}$ values according to the observation of training data. Therefore, attribute A divides D into v subsets $\{D_1, D_2, D_3, \cdots, D_V\}$, where the tuples in D_j have the same value a_j on attribute A. However, these partitions may contain tuples not from the same class but from different classes, that is, impure. After this partition, how much information is needed for the accurate classification of the generated tuples, which can be measured by the following formula:

$$\mathrm{Info}(A) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} \mathrm{Info}(D_j). \quad (4)$$

Among them, item $|D_j|/|D|$ denotes the weight of the j th partition, and $\mathrm{Info}(A)$ denotes the expected information needed to classify the components of D by attribute A. The information gain obtained by branch on attribute A can be described as:

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{info}(A). \quad (5)$$

The advantages of ID3 algorithm are as follows:

(1) The basic principle of the algorithm is clear

(2) The classification speed is faster

(3) Practical example learning algorithm

Its shortcomings are as follows:

(1) There is a bias problem. The number of feature attributes affects the amount of information

(2) A problem with training data will make the results different and more sensitive to noise

(3) The probability of error is proportional to the increase of category

2.2.2. C4.5 Algorithm. An early machine learning algorithm and a common algorithm for constructing decision tree classifiers became the basis of many decision tree algorithms later.

(1) The information gain rate is used as attribute selection measure to solve the problem of bias

(2) It can discretize attributes with continuous values and deal with incomplete data

(3) Pruning at the same time in the process of tree construction

With the extension of information gain, benefit ratio can solve the drawback of ID3. In the assumption that a variable is selected as a partitioning attribute, with a higher information gain of the variable than the information gain of its other variables is needed. The definition formula of

segmentation information is as follows:

$$\text{SplitInfo}(A) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \quad (6)$$

The ratio of the increase in information is mainly compared with the total amount of information in some segments. The formula is as follows:

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}. \quad (7)$$

2.2.3. CART Algorithm. Classification and Regression Tree (CART) is a technique for generating binary decision trees. In fact, its principle is dichotomy recursive segmentation technology. In order to produce subnodes, it divides two sample subsets; that is, only two subnodes are generated, so finally, a simple binary decision tree is obtained. Unlike ID3 and C4.5, which are based on information entropy splitting technology, CART chooses the best grouping variables and splitting points based on gini coefficient and variance and chooses the attributes with the minimum gini coefficient as the current test attributes. If the gini coefficient value is smaller, the more reasonable the segmentation is, and the higher the purity of the sample set is.

If the training tuple set D contains records of m categories, then the gini index is determined as follows:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} P_i^2. \quad (8)$$

Calculate the sum of m classes, where P_i is the probability that any record in D belongs to C_i class and is expressed by $|Ci, D|/|D|$. If D is divided into D1 and D2, the gini coefficient of this division is

$$\text{Gini}(D) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2), \quad (9)$$

where $|D|$ is the number of samples in D and $|D_1||D_2|$ is the number of samples in D1 and D2, respectively.

The CART algorithm terminates splitting and stops constructing decision tree if the following conditions exist.

(1) The data records contained in leaf nodes belong to the same category

(2) The number of samples covered by a branch is less than a threshold set by the user in advance

3. Experiments

3.1. Selection of Experimental Platform. Through this paper, SPSSC lementine 12.0 is elected as the data mining platform in conjunction for the actual research work. For the mining platform, the selection of the platform is mainly based on the following six aspects:

(1) Clementine has the functions of classification and prediction, association analysis, time series analysis, and clustering. It provides a variety of methods, such as neural network, decision tree and regression tree, linear regression, logistic regression, self-organizing network, and fast clustering

(2) Clementine has an interactive and visual user interface, which combines intuitive user graphics interface with a variety of analysis techniques. It is a very easy software for users to build models by connecting nodes, and data mining model can be built without programming. So that users can put more energy into the application of data mining to solve specific business problems, rather than the use of software

(3) Clementine has an open database interface that provides rich data access capabilities for access to files and relational databases. It also provides the ability to input data processing and output data settings

(4) Clementine provides two ways to build models. In the simple mode, the user does not need to make any settings; the system will build the model according to the default settings; in the expert mode, the user can adjust the parameters in the model according to his own needs, so that the model achieves the best results

(5) Provide powerful publishing capabilities to export data mining models or entire data mining processes to embedded systems

(6) Provide complete data flow management and project management functions. The former can effectively manage the data flow, data mining model, and mining results in the work area. The latter can effectively manage the entire project; users can manage related project files according to different stages of data mining and can effectively manage data mining projects according to data flow, nodes, data mining models, results, and other methods

3.2. Data Acquisition. With this paper, the data are obtained from the financial data of 1500 relevant firm clients of a commercial bank, averaged over the years 2015 to 2018. The attributions in the financial information data tables provided by the bank are in the transaction database based attributes, so conversion of attributes is performed to form 18 attributes that reflect the financial indicators for a firm, as shown in Table 1. Firstly, according to the relevant indicators of enterprises and the actual situation of enterprises in 2018, the experts of financial institutions define the risks of enterprises as high, higher, medium, and low. Among them, the enterprises with high risk are those that will fail from 2017 to 2018; the enterprises with high risk are those that will produce credit default; the

Table 1: Randomly selected with 12 attributes from 18 alternatives.

Number	Attribute	Calculation formula
1	Asset-liability ratio	Total liabilities/total assets
2	Net profit margin of operating income	Major business profit/major business income
3	Return on assets	Net profit/(total shareholders' equity + total shareholders' equity in the previous period)*2
4	Fixed asset ratio	Total fixed assets/total assets
5	Liquidity ratio	Total current assets/total current liabilities
6	Quick ratio	(Total current assets - net inventory)/total current liabilities
7	Interest guarantee multiple	(Net profit + income tax + financial expenses)/financial expenses
8	Total asset turnover rate	Main business income/(total assets + total previous assets)*2
9	Inventory turnover	Main cost/(net inventory + net previous inventory)*2
10	Receivable turnover rate	Main business cost/(accounts receivable + last accounts receivable)*2
11	Receivable turnover rate	Main business income/(accounts receivable + last period accounts receivable)*2
12	Cash ratio of main business income	Cash flow/main business income from operational activities
13	Inventory current liability ratio	Net inventory/total liquidity liabilities

Table 2: Comparison table of classification accuracy of multiple random decision trees.

Verification times	High risk%	Higher risk%	Medium risk%	Low risk%
1	53.21	71.66	78.23	88.79
2	54.43	72.12	77.98	89.01
3	48.67	74.98	81.61	89.37
4	58.65	75.45	78.92	86.02
5	51.36	73.37	78.76	88.09
Average	53.34	73.46	78.94	88.57

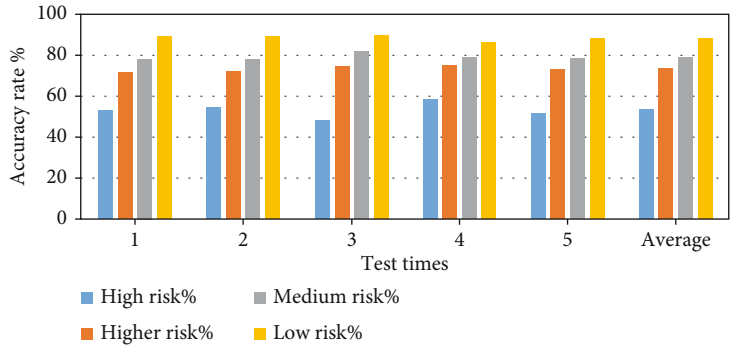

Figure 2: Comparison of classification accuracy of multiple random decision trees.

enterprises with medium risk are those that have no default but have deteriorating financial situation, and the enterprises with low risk have good financial situation and no credit default. At each tree construction, a randomized method was used. For verification of the stability of a decision tree classification, a total of 5 experiments were conducted. At each training dataset, 1200 data were randomly selected with the tree from the original dataset as training data, which was randomly selected with 12 attributes from 18 alternatives.

Table 3: Comparing table of accuracy of C4.5 algorithms.

Verification times	High risk%	Higher risk%	Medium risk%	Low risk%
1	35.24	60.62	65.29	72.75
2	37.43	62.12	66.98	73.01
3	34.67	64.98	71.61	76.37
4	38.65	65.45	70.92	76.02
5	31.36	63.37	68.76	78.09
Average	35.34	63.46	68.94	74.57

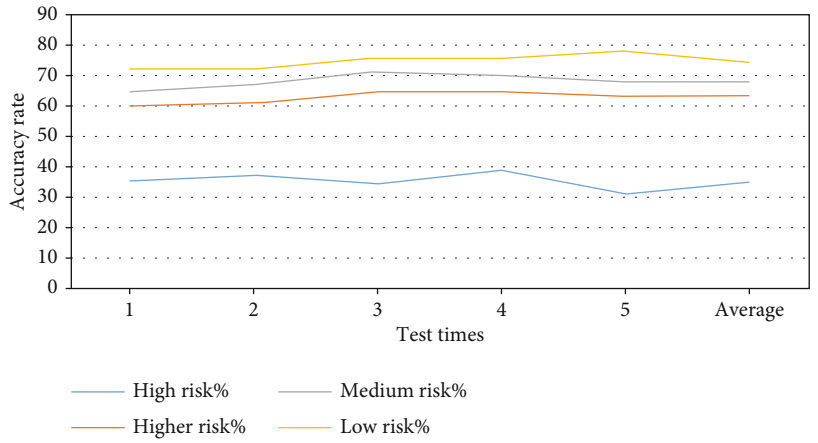

Figure 3: Comparisons of accuracy of C4.5 algorithms.

Table 4: Accuracy comparison between random decision tree algorithm and C4.5 algorithm.

	High risk%	Higher risk%	Medium risk%	Low risk%
Stochastic decision tree algorithm	52.63%	70.15%	78.89%	82.51%
C4.5 algorithm	37.26%	60.94%	65.75%	68.21%

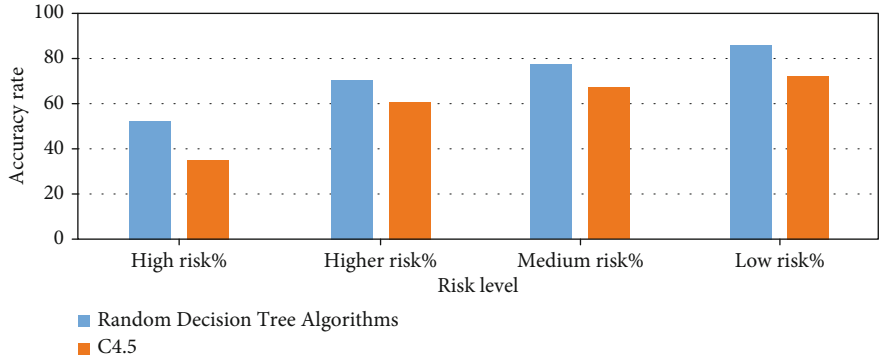

Figure 4: Comparing the accuracy of stochastic decision tree algorithm with that of C4.5 algorithm.

Table 5: Comparison table of stratified sampling accuracy of multiple random decision trees.

Verification times	High risk%	Higher risk%	Medium risk%	Low risk%
1	71.41	77.32	83.29	89.78
2	72.43	76.12	84.98	90.01
3	71.67	78.98	87.61	90.37
4	70.65	79.45	78.92	89.02
5	72.36	79.37	85.76	88.09
Average	71.34	78.46	85.94	89.57

Table 6: Accuracy comparison table of stratified sampling and random sampling.

	High risk%	Higher risk%	Medium risk%	Low risk%
Stratified sampling	71.25%	78.59%	85.12%	88.91%
Random sampling	51.69%	71.20%	79.21%	86.57%

4. Discussion

4.1. Accuracy Comparison

(1) In order to compare the data, we counted the progress of a large number of random decisions, making the data comparison between them obvious, and the comparison results are shown in Table 2 and Figure 2.

We can see that the classification accuracy is informative for bank risk prediction by the confirmation of bank personnel and based on the data presented through the graph. However, the algorithm has relatively low classification accuracy for high risks. The main reason is that the number

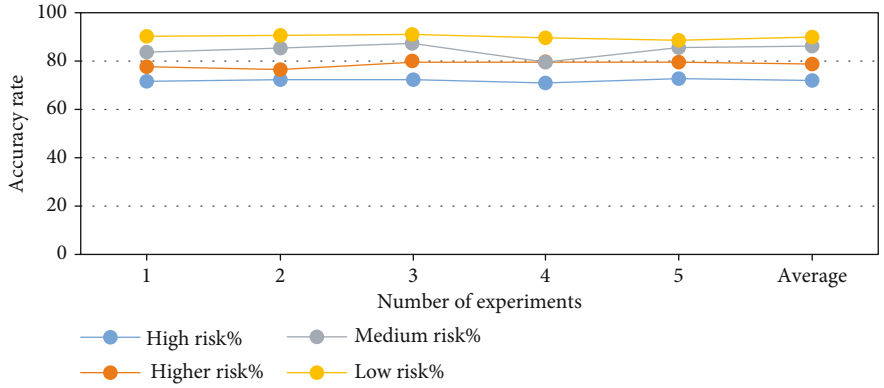

Figure 5: Comparisons of stratified sampling accuracy for multiple random decision trees.

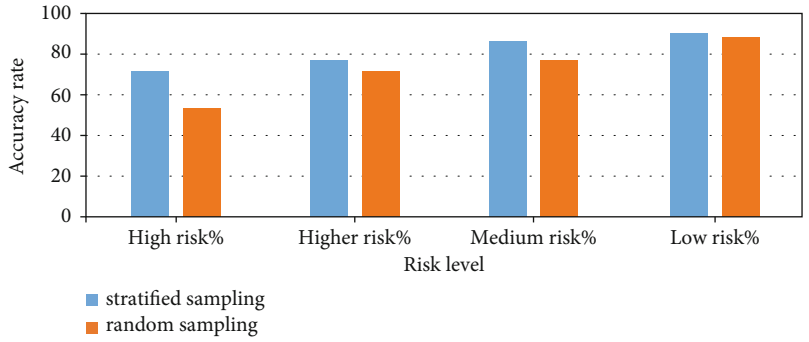

Figure 6: Comparison of stratified sampling and random sampling accuracy.

of data with high risk in the training data set is small, resulting in insufficient training of this kind of branch.

(2) Accuracy analysis of C4.5 algorithm, as shown in Table 3 and Figure 3.

The classification accuracy of the algorithm for high risk is relatively low. The main reason is that the number of data with high risk in the training data set is small, which results in insufficient training of this kind of branch. Table 4 is the accuracy comparison table between the random decision tree algorithm and the C4.5 algorithm.

According to Table 4, the accuracy comparison chart between the random decision tree algorithm and the C4.5 algorithm is obtained, as shown in Figure 4.

From Figure 4, we can see that the accuracy of random decision tree method is about 10% higher than that of C4.5. In order to improve the accuracy of high risk, 300 high risk data were added to the training data set. The original random sampling is replaced by stratified sampling. The original data are stratified according to the high, higher, medium, and low risk. Random sampling is used for each level to ensure the number of training data with high risk. The following Tables 5 and 6 and Figures 5 and 6, respectively, show the stratified sampling accuracy comparison table of multiple random decision trees, the comparison table of stratified sampling and random sampling accuracy, the comparison of stratified sampling accuracy of multiple random decision trees, and the stratified sampling.

We can see that the accuracy of high-risk increases to 10% after stratified sampling, which is mainly because stratified sampling increases the number of high-risk samples. Then, the accuracy of decision tree classification is related to the number of training data samples. By having a larger sample size, the more accurate the decision tree of classification.

5. Conclusions

With the continuous progress of computer theory and technology, more and more computer data processing and analysis methods are combined with financial intervention work efficiently and organically, which has brought revolutionary innovation to the theory, mode, and method of financial intervention work. Especially the introduction of data mining technology, it brings new ideas for financial analysts, improves the efficiency and quality of financial intervention, and plays an increasingly important role.

(1) On the basis of introducing the background of topic selection, process steps, and application fields of data mining and focuses on the commonly used algorithms of data mining

(2) Based on the theory of information entropy and through theoretical proof, this paper proposes an objective and fair method to evaluate the clustering effect and applies this method to solve practical

problems and achieves better practical results. Due to incomplete data and partial distortion in raw data acquisition, the accuracy of the model is affected to a certain extent. Further work is to increase the number of experimental samples, fully tap the potential useful information; add some derivative variables to make the results of analysis more objective and convincing; the results of analysis are more comprehensive and have greater practical value

(3) This paper analyses and studies the classification technology of decision tree in data mining, especially the application of C4.5 algorithm to loan data of a credit cooperative, establishes decision tree and classification rules, builds audit analysis model, and facilitates financial analysts to find problems and find clues to financial problems

Data Availability

This article does not cover data research. No data were used to support this study.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the Humanities and Social Sciences Project of Henan Provincial Education Department (No. 2022-ZZJH-098), the Independent Innovation Application Research Project of Zhongyuan University of Technology (No. K2018YY023), and the Graduate Quality Engineering Project of Zhongyuan University of Technology (No. QY202102).

References

[1] C. R. Tryon, *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities*, Edwards Brothers, Ann Arbor, 1939.

[2] R. B. Cattell, "The description of personality: basic traits resolved into clusters," *Journal of Abnormal and Social Psychology*, vol. 38, no. 4, pp. 476–506, 1943.

[3] D. Swingley, "Statistical clustering and the contents of the infant vocabulary," *Cognitive Psychology*, vol. 50, no. 1, pp. 86–132, 2005.

[4] U. Maulik, "Medical image segmentation using genetic algorithms," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 166–173, 2009.

[5] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4324–4338, 2015.

[6] D. Liu, M. H. Jiang, X. F. Yang, and H. Li, "Analyzing documents with quantum clustering: a novel pattern recognition algorithm based on quantum mechanics," *Pattern Recognition Letters*, vol. 77, pp. 8–13, 2016.

[7] L. A. Lopes, V. P. Machado, R. A. L. Rabêlo, R. A. S. Fernandes, and B. V. A. Lima, "Automatic labelling of clusters of discrete and continuous data with supervised machine learning," *Knowledge-Based Systems*, vol. 106, pp. 231–241, 2016.

[8] R. J. Kuo, C. H. Mei, F. E. Zulvia, and C. Y. Tsai, "An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation," *Neurocomputing*, vol. 205, pp. 116–129, 2016.

[9] Hung-Leng Chen, Kun-Ta Chuang, and Ming-Syan Chen, "On data labeling for clustering categorical data," *IEEE Transactions on Knowledge and Data Engineer*, vol. 20, no. 11, pp. 1458–1472, 2008.

[10] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Satistics and Probability*, vol. 1, no. 14, 1967.

[11] I. V. Zhihan and H. Song, "Trust mechanism of feedback trust weight in multimedia network," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 4, pp. 1–26, 2021.

[12] X. Li, H. Liu, W. Wang, Y. Zheng, H. Lv, and Z. Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167–177, 2021.

[13] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[14] Z. X. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 1997.

[15] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transaction on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[16] J. Y. Yeh and C. H. Chen, "A machine learning approach to predict the success of crowdfunding fintech project," *Journal of Enterprise Information Management, Accepted Manuscript*, 2020.

[17] S. Guha, R. Rastogi, and K. Shim, "Rock: a robust clustering algorithm for categorical attributes," *Information Systems*, vol. 25, no. 5, pp. 345–366, 2000.

[18] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS-clustering categorical data using summaries," *Knowledge Discovery and Data Mining*, pp. 73–81, 2000.

[19] D. Barbará, Y. Li, and J. Couto, "COOLCAT," in *Proceedings of the eleventh international conference on Information and knowledge management - CIKM '02*, New York, 2002.

[20] J. F. Brendan and D. Delbert, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[21] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Transactions on Pattern Analysis and Machin*, vol. 29, no. 3, pp. 503–507, 2007.

[22] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, pp. 120–127, 2012.

[23] G. Chen, Y. Lu, B. Li, K. Tan, and T. Moscibroda, "MP-RDMA: enabling RDMA with multi-path transport in datacenters," *IEEE/ACM Transactions on Networking*, vol. 27, no. 6, pp. 2308–2323, 2019.

[24] G. Chen, Y. Lu, Y. Meng et al., "FUSO: fast multi-path loss recovery for data center networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1376–1389, 2018.

[25] Y. Zeng, G. Chen, K. Li, Y. Zhou, X. Zhou, and K. Li, "M-skyline: taking sunk cost and alternative recommendation in consideration for skyline query on uncertain data," *Knowledge Based Systems*, vol. 163, pp. 204–213, 2019.

[26] Z. Y. He, X. F. Xu, and S. C. Deng, "Squeezer: an efficient algorithm for clustering categorical data," *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611–624, 2002.

[27] D. W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1263–1271, 2004.