Design of Enterprise Financial and Economic Data Accurate Classification Management System Based on Random Forest

Junlin Li[1(✉)] and Haonan Chu[2]

[1] Beijing Union University, Beijing 100101, China
lijunlll1512@163.com
[2] School of Labor Relations and Human Resource, China University of Labor Relations, Beijing 100048, China

Abstract. The conventional system has the problems of low recall rate, accuracy rate and high false positive rate of financial data classification. Therefore, an accurate classification management system of enterprise financial and economic data based on random forest is proposed. In the hardware, the front end, middle layer, server end and enterprise financial and economic data display end are used to form the overall architecture of the system, optimize the data memory of the server end, and transform the serial communication circuit of the development board; In the software design, abnormal financial data are filtered, a decision tree is established for each sample data, the utility function value is learned through the membership of the decision tree, and the optimal classification category is selected for the data through the random forest classifier. The experimental results show that the designed system improves the recall and accuracy of data classification, reduces the false positive rate, and the financial data classification results are more accurate and reliable.

Keywords: Random forest · Financial data · Management system · Data classification

1 Introduction

Reasonable classification is the basis for in-depth mining and analysis of financial data. However, due to the large scale and weak regularity of enterprise financial and economic data, the problem of data imbalance has become the key problem of financial data mining. Unbalanced data leads to the decline of the accuracy of financial data classification and the actual effect of data. Therefore, it is necessary to accurately classify financial data [1]. The research on the accurate classification management system of financial data is conducive to improve the classification management and information analysis ability of financial data, and is of great significance to the safe operation of enterprises [2].

With the rapid development of big data and information technology, foreign data classification management systems have achieved good development. Through the feature extraction and fusion clustering processing of financial data, the internal association

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 287–299, 2022.
https://doi.org/10.1007/978-3-031-18123-8_22

rule feature information of financial data is extracted, and then the automatic classification and identification of financial data are carried out according to the distribution of feature information, so as to realize the adaptive fusion processing of financial data, It can improve the business process management ability of financial data [3]. The domestic data classification management system has also made great progress. The software is designed in the embedded environment, the multiple regression analysis method is used to analyze the statistical characteristics of financial data, the retrieval structure model of financial database is constructed, and the statistical analysis method is used for automatic statistics of financial data to realize the optimal classification of financial data [4].

Liu et al. Proposed to design a text classification method of distributed financial data based on multi neural network fusion. This method designs a multi-element neural network path including word embedding layer, convolution layer, bidirectional gating cycle unit layer, attention mechanism layer and softmax layer; On this basis, the demand effect resource classification strategy is adopted to complete the mapping transformation from the demand of qualitative science and technology resources to the solution of quantitative resource service effect, and then to the output of qualitative science and technology resources. It focuses on solving the significant long-distance dependence characteristics of distributed text, obtaining effect knowledge quickly and accurately, and improving the analysis effectiveness of financial data. Liu et al. Proposed to design a financial data classification method based on width learning system. This method extracts the deep features of financial data through simple structure to speed up the classification speed The input data is constructed by using the time series of voxel mean of the region of interest in the data, the shallow and deep features of the financial data are extracted respectively, mapped into feature nodes and enhancement nodes of width learning, and the model framework is constructed. The connection weight of the classification model is calculated by inverse ridge regression to classify the financial data and greatly reduce the training time.

However, the data classification of conventional systems mostly presupposes the uniform distribution of the number of samples in different categories, resulting in insufficient data classification. In response to this problem, combined with existing research theories, a random forest-based accurate classification management system for corporate financial and economic data is proposed.. Random forest is to artificially synthesize new minority samples to reduce the imbalance of data categories, and perform linear difference between neighboring minority samples to synthesize new minority samples.

2 Design of an Accurate Classification Management System for Enterprise Financial and Economic Data Based on Random Forest

2.1 Hardware Design of Accurate Classification Management System for Enterprise Financial and Economic Data

The Overall System Architecture Design

The overall architecture of the system is composed of four parts: front end, middle layer,

server end and enterprise financial and economic data display end. The server side adopts micro server, uses database to transmit information, provides business interface through middle tier architecture, and connects the front end and server side. The front end and server side exchange information with the middle tier through interfaces respectively, uses the front end to control enterprise financial and economic data, and outputs enterprise financial and economic data at the display end. The overall architecture of the system is shown in Fig. 1:

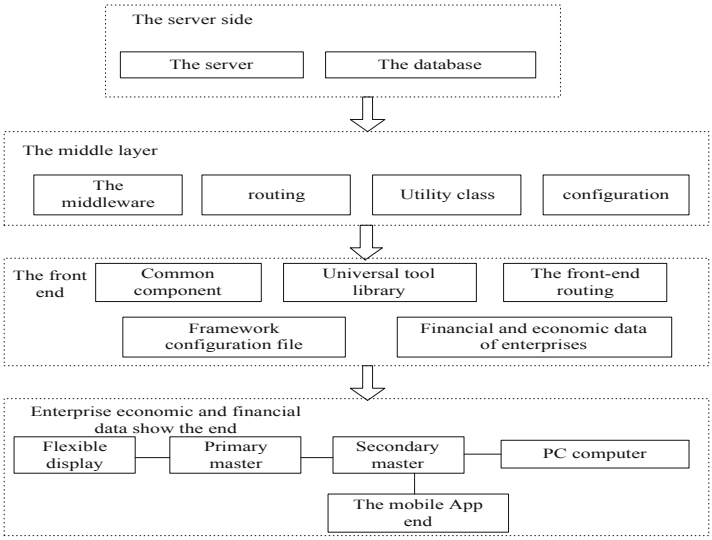

Fig. 1. The overall architecture of the accurate classification management system for corporate financial and economic data

The front-end adopts a component-based development method, using chart customization, file segmentation and uploading and other components to form reusable universal components, abstract processing of element event-related libraries to form a universal tool library, and the front-end routing layout components adopt a skeleton structure. All business components are embedded in the layout components to ensure that the business can be reused. According to different business functions, switch different business components, select front-end routing configuration, front-end environment configuration, packaging tool configuration, multi-national language configuration, and form a framework configuration file, Corporate financial and economic data includes static resources, image resources, etc. When classifying corporate financial and economic data, each business component independently references label resources internally. The middle-tier architecture uses the middleware organizational framework to add user authentication, log, and exception handling middleware to the middle-tier to check the login status requested by the interface, record the corporate financial and economic data classification program behavior, and handle the wrong program behavior in a timely manner. The interface request is mapped to the tool module through routing, the encapsulated business component is used to forward the request, and the configuration of the

framework environment is formed through the server address, deployment environment, and external device address [5]. The server adopts a B/S structure, including multiple controllers and filters, as well as multiple business entities and database entities, to check and process different interface requests and business logic. The enterprise financial and economic data display terminal adopts the secondary micro-control method. The single-chip microcomputer and the development board are respectively used as the primary and secondary master controllers. The development board is regarded as the corporate financial and economic data motherboard, equipped with Wi-Fi modules, The power conversion module, etc., convert the communication protocol to realize the communication between the primary control node and the PC end and the APP end. The PC end uses serial communication and the APP end uses wireless communication. At this point, the overall system architecture design is completed.

Optimize the Storage of Corporate Financial and Economic Data

Optimize the enterprise financial and economic data memory on the server side of the system, and expand the storage capacity of the memory for enterprise financial and economic data. Firstly, the input and output interfaces of the memory are optimized. The optimized interfaces are shown in Fig. 2:

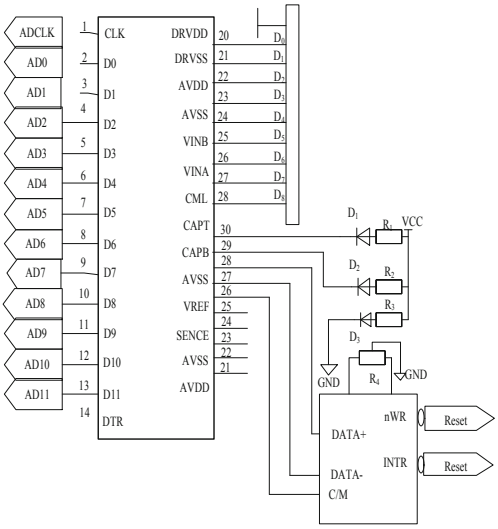

Fig. 2. Schematic diagram of enterprise financial and economic data storage interface

LDPC decoder, Q hard decision message memory, buffer buffer, decoding result memory and cyclic shift coefficient memory are selected as the configuration devices of server-side memory. Firstly, the enterprise financial and economic data meeting the security level is written into the buffer, the buffer is full, the data is read out, it is input into the Q hard decision message memory, the corresponding column address of the network information is read, the corresponding cyclic shift value is taken out from the address, the information is cyclically shifted left by the same digit according to the value, and the left

shift results are transmitted to the LDPC decoder in turn, Update the information node repeatedly, obtain the latest decoding message, output the decoding results of enterprise financial and economic data, and finally store the decoded information in the memory [6]. Define the interface form of memory input and output, as shown in Table 1:

Table 1. Memory input and output interface

Port	Input/Output	Bit width	Explain
DRVDD	Enter	1	Input from outside, system clock
AVDD	Enter	2	Information input enable terminal, input the quantized data to be decoded
VINA	Enter	1	Reset terminal, ensure low level is valid
CAPT	Enter	1	Information input terminal, to ensure high level effective
AVSS	Enter	1	Bit rate selection input
SENCE	Output	8	Decode data output terminal to ensure high level is effective
AVDD	Output	2	Output enable indicator

The 0.5 code rate represents the low level of the code rate selection input, and the 0.8 code rate represents the high level. For the output enable indicator, the parallel output mode is adopted, the decoding result is set to 1 bit, and 8 decoding results of information are output at one time, so as to increase the size of system memory and improve the maximum processing number of user access requests. So far, the optimization of enterprise financial and economic data memory has been completed.

Optimize the Main Board Structure of Corporate Financial and Economic Data

Optimize the main board structure of the display end of the financial and economic data of the enterprise, and transform the serial communication circuit of the financial data development board to reduce the interference of data collection, update, and communication. The development board uses EEPROM chip as the core chip, and the main board uses STM32F103ZET6 signal microprocessor, equipped with 512 KB Flash memory, provides high-density code instructions, efficiently stores business tag data, and integrates comparators and timers in the peripheral modules of the microprocessor., Power supply, etc., the external power supply adopts AMS1117 power conversion core, and provides 3.3 V DC voltage for the main board through the positive and negative interface. Configure the network access point inside the development board, use the 433 MHz wireless communication module, and use the TCP communication protocol to communicate with the APP and the first-level main controller, set the wireless communication module parameters, and maintain the APP-side communication channel and the first-level main controller The communication channels are the same, and the 433 MHz frequency band is selected [7]. Use serial communication to connect the corporate financial economic

data mainboard and PC upper computer, and the optimized serial communication circuit is shown in Fig. 3:

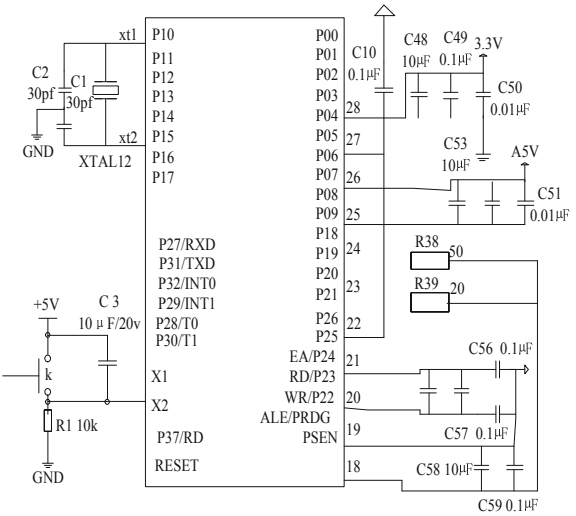

Fig. 3. The serial communication circuit of the main board of enterprise financial economic data

The serial communication circuit of the development board is divided into amplification circuit and filter circuit. AD620 differential amplification chip is selected as the amplification circuit, and the amplification factor of the chip is 7 times. Set the zero point of the output end of the amplifier, ground the 6 pins, and connect with the polar capacitor and ceramic capacitor to reduce the signal noise. AD8676 amplifier is selected as the filter circuit to connect the 6 pins of the amplification circuit, The circuit is disassembled into low-pass filter and high pass filter, and the voltage following method is used to realize the main stage amplification and band-pass filter of the signal. The optimized serial communication circuit includes data collector signal processing circuit, power supply circuit and serial port circuit. The financial data development board adopts 3.3 V power supply, configures the power monitoring reset chip for the power supply circuit, integrates a certain capacity of serial memory, and connects the reset signal to the reset pin of the reset chip, so that the power supply circuit has the function of power down protection. In order to keep the output signals of each data collector consistent, a differential processing chip is configured for the signal processing circuit of the collector, the original signal is input into the differential processing chip, and the three differential electrical signals are converted into three single ended electrical signals, in which the two signals of x-axis and y-axis are processed into single ended signals, the signal of z-axis is processed into zero clearing signals, and then the three signals are inverted, Ensure that the signal matches the voltage of the differential processing chip pin. After the differential processing chip outputs three logic electrical signals, it transmits the signals to the serial port circuit, configures a voltage stabilizing chip for the serial port circuit, performs pulse counting processing on the three logic electrical signals, installs

a static register and a clock backup register at the serial port, and uses the two registers to obtain voltage from the pins of the voltage stabilizing chip to provide reliable power supply for the serial port circuit, Parallel capacitors are connected to the output of the core controller to filter the three signals, so as to improve the transient response and stability of the signal. So far, the optimization of the structure of the main board of financial data has been completed, and the hardware design of the accurate classification management system of enterprise financial and economic data has been realized.

2.2 Software Design of Precise Classification Management System for Enterprise Financial and Economic Data

Preprocessing Enterprise Financial and Economic Data

Through abnormal identification, find the source of weak data, and correct the abnormal financial and economic data of the enterprise. Obtain the financial and economic data files of the system enterprise, simplify and merge the data, and keep the data format consistent. Traverse all data files, treat each sample data as a two-tuple, including sample observations and sample labels, and check data integrity. Through time-effect correlation, test the data relevance of different time series, consider the existence of small-scale statistical samples in the system, and use statistic Z for testing. The statistic calculation formulas for normal samples and test samples are:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{\sum u_1^2 + \sum u_2^3}{W_1 + W_2 - 2} \times \frac{W_1 + W_2}{W_1 W_2}}} \tag{1}$$

Among them, z_1 and z_2 are the mean value of the data, $u_1,\ u_2$ is the standard deviation of the data, and $W_1,\ W_2$ is the data capacity. The larger the Z value, the greater the degree of difference between the normal sample and the test sample. Through the completeness and correlation test, the abnormality of the financial and economic data of the enterprise can be obtained [8].

Identify the sample data that is significantly different from the normal sample as the abnormal data in the financial and economic data of the enterprise. Using a clustering algorithm based on the density of clustered data groups, clustering analysis of emergency repair data, for the multi-dimensional big data of the power distribution system, the normal sample is taken as the core point, the array to be tested is $V = \{v_0,\ v_1,\ \cdots,\ v_n\}$, the core point is v_0, and the inspection data is $v_1,\ \cdots,\ v_n$, n is the number of arrays, and the formula for the distance between the normal sample and the test sample is:

$$U\left(v_j,\ v_0\right) = \sqrt{\left(s_j - s_0\right)^2 + \left(w_j - w_0\right)^2} \tag{2}$$

Among them, $U\left(v_j,\ v_0\right)$ is the core distance of the two sets of data, s_j is the observation value of the inspection data, $j = 1,\ 2,\ \cdots,\ n$ and s_0 are the observation values of the core points, and w_j and w_0 are the types of the inspection data and the core points respectively. Set the ideal core distance S and the reachable distance T, when $U\left(v_j,\ v_0\right)$ is less than the ideal core distance, it is judged that the test sample is in the same cluster

as the normal sample, and when $U(v_j,\ v_0)$ is greater than the ideal core distance, it is judged that the two sets of samples are of different clusters. Regarding the core distance and the reachable distance as the neighborhood and density of the array, respectively, they are used as the measurement standard for the clustering of the sample data, and the clustering constraint conditions are obtained. The expression is:

$$\begin{cases} S(v_j,\ v_0) = \{U(v_j,\ v_0) \le S\} \\ S(v_j,\ v_0) \ge T \end{cases} \tag{3}$$

where $S(v_j)$ represents the data in the cluster with v_0 as the cluster center. After the enterprise financial and economic data is classified into clusters, calculate the statistics $Z(v_j,\ v_0)$ of each test data and the normal sample, sort the statistics in the same cluster, and use the sample data with the largest $Z(v_j,\ v_0)$ value as the abnormal data, and filter the abnormal data to obtain the revised data set. At this point, the preprocessing of corporate financial and economic data has been completed.

Accurate Classification of Corporate Financial and Economic Data Based on Random Forest

For the preprocessed corporate financial and economic data, a random forest classifier composed of multiple decision trees is used to accurately classify data categories. The classification results of the random forest classifier include quick ratio, current ratio, accounts receivable turnover ratio, market-sales ratio, price-to-book ratio, cost-to-interest ratio, total asset growth rate, asset-liability ratio, shareholder equity ratio, equity ratio, etc., Establish a decision tree for each financial and economic sample data, and select the best classification result for each decision tree among all the classification results. Analyze the membership degree of each decision tree to the data category, set the membership degree of the i decision tree to the b data category as B_{ib}, and the replacement accuracy of the membership degree B_{ib} as $C(B_i)$, calculate the node purity $d(B_{ib})$ of the membership degree B_{ib}, and the formula is:

$$d(B_{ib}) = D(B_{ib}) log \frac{E}{e_i} \tag{4}$$

where $D(B_{ib})$ is the ratio of the sum of membership degrees of B_{ib} and i decision tree to all data categories, E is the number of categories of the random forest classifier, $b \in [1, E]$, e_i are the personalized parameters of the i decision tree, and the actual sample data Semantic correlation [9]. Calculate the membership reliability c_i of the i decision tree, the formula is:

$$c_i = \frac{C(B_{ib})}{A} d(B_{ib}) \tag{5}$$

where A is the total number of financial and economic sample data, and $i \in [1, A]$. The larger the value of c_i, the higher the credibility of the membership B_{ib} of the i decision tree to the b data category. According to the credibility c_i, the decision trees of all sample data are arranged in descending order. When the system when adding new enterprise financial and economic data, the membership reliability of each decision tree is dynamically adjusted, and the new decision tree is ranked in descending order.

Arrange the results in descending order of the decision tree, match the weights of the weight set in turn, set the weight set to $\{k_1, k_2, \cdots, k_A\}$, and assign a weight value k_i to the membership B_{ib} of the i decision tree. Use the learning utility function K to evaluate the sample set, make each enterprise financial and economic data learn a utility function, assign a utility score for each data, and keep the constraint conditions of all the effective scores consistent. The expression of the learning utility function Q is:

$$Q = m \sum_{i=1}^{A} \frac{C(B_{ib})}{A} M(B_{ib}) d(B_{ib}) k_i \tag{6}$$

Among them, m is the utility score, and $M(B_{ib})$ is the priority of membership B_{ib} [10]. The Q value is closely related to the accuracy of the membership degree of the sample data. Set the high and low thresholds of the function Q. When Q is greater than the high threshold, the decision tree's membership of the data category must be accurate. When Q is less than the low threshold, the decision is determined the membership of the tree to the data category must be inaccurate. If Q is less than the high threshold and greater than the low threshold, the membership of the decision tree to the data category is judged to be a fuzzy number. When the learning utility function value of the membership degree of the decision tree is inaccurate or fuzzy value, the membership degree of the decision tree to the data category is re-selected, and the above process is repeated until the Q value of the membership degree is greater than the high threshold. Count the accurate membership degrees of each decision tree to all categories of the random forest classifier, and select the classification category with the highest membership degree as the classification result of the financial data. So far, the accurate classification of enterprise financial and economic data based on random forest is completed, the system software design is completed, and the hardware design and software design are combined to realize the accurate classification and management system design of enterprise financial and economic data based on random forest.

3 Experiment and Analysis

The designed system is compared with two conventional enterprise financial and economic data accurate classification management systems to compare the recall rate, false positive rate and accuracy of the three systems.

3.1 Experimental Data

The experiment is built on the Matlab platform. In order to verify the effectiveness of the system, the data set uses the quarterly financial statements of all listed companies from 2015 to 2020, including income statements, cash flow statements, and balance sheets. The statistical information sampling scale of financial data is 1000 MBit, the data set contains abnormal samples of abnormal financial conditions of enterprises, and normal samples of normal financial conditions. The random forest classifier starts from six different dimensions of capital structure, cash flow, operating capacity, development capacity, debt clearing capacity, and profitability. The classification categories of all financial and economic data included are shown in Table 2:

Table 2. Classification categories of financial and economic data

Type	Primary coverage
Capital structure	Asset liability ratio, long-term applicability of assets, dynamic liabilities, shareholders' equity, proportion of fixed assets and current assets
Cash flow	Cash recovery rate of total assets, sales revenue, cash proportion of operating income, and capital expenditure
Service power	Total asset turnover rate, fixed asset turnover rate, current asset turnover rate, accounts payable turnover rate, accounts receivable turnover rate, inventory goods turnover rate
Development ability	Growth rate of total assets, growth rate of net assets, net amount of cash held by enterprise operation, growth rate of net profit, total profit, growth rate of operating profit, growth rate of income from single share
Debt clearing capacity	Current cash-liability ratio, net operating current cash, net operating current cash, profit before amortization, depreciation and tax, property right ratio, overspeed moving ratio, current ratio
Profitability	Cost profit margin of capital, operating profit margin, asset impairment loss, total operating cost, net profit, net interest rate of assets, return on assets, return on equity

3.2 Analysis of Experimental Results

The three systems divide the categories of each sample data, and manage enterprise financial and economic data sets. Compare the data classification behavior recall rate R of the three systems to measure the proportion of normal samples that are correctly classified. The formula for calculating the R value is:

$$R = \frac{x_1}{x_1 + x_2} \tag{7}$$

Among them, x_1 is the number of normal financial data classified correctly, and x_2 is the number of normal financial data classified incorrectly. The comparison result of the recall rate experiment is shown in Fig. 4:

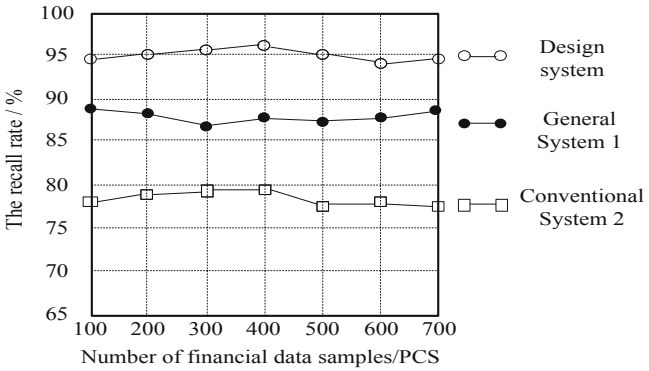

Fig. 4. Comparison results of the financial data classification recall rate experiment

It can be seen from the above figure that the average recall rate of data classification of the design system is 95.0%, the average recall rate of data classification of conventional system 1 is 86.8%, and the average recall rate of data classification of conventional system 2 is 78.9%. The recall rate of the design system has increased by 8.2% and 16.1% respectively, which improves the proportion of normal samples correctly classified.

Compare the misreport rate ξ of data classification behaviors of the three systems to measure the misclassification ratio of abnormal samples. The formula for calculating the value of ξ is:

$$\xi = \frac{x_3}{x_3 + x_4} \tag{8}$$

where x_3 is the number of abnormal financial data classified incorrectly, and x_4 is the number of abnormal financial data classified correctly. The experimental comparison result of false alarm rate is shown in Fig. 5:

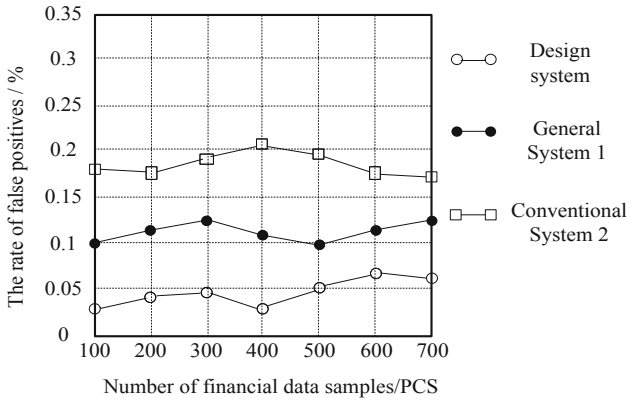

Fig. 5. Experimental comparison results of the false alarm rate of financial data classification

It can be seen from the above figure that the average false alarm rate of the design system is 0.05%, and the average false alarm rates of the other two systems are 0.11% and 0.19% respectively. The false alarm rates of the data classification behavior of the design system are reduced by 0.06% and 0.14% respectively, reducing the misclassification proportion of abnormal samples.

Compare the accuracy rate β of the data classification behavior of the three systems, and measure the proportion of the normal sample and the abnormal sample that are correctly classified. The calculation formula of the β value is:

$$\beta = \frac{x_1 + x_4}{x_1 + x_2 + x_3 + x_4} \tag{9}$$

The comparison result of the accuracy rate experiment is shown in Fig. 6:

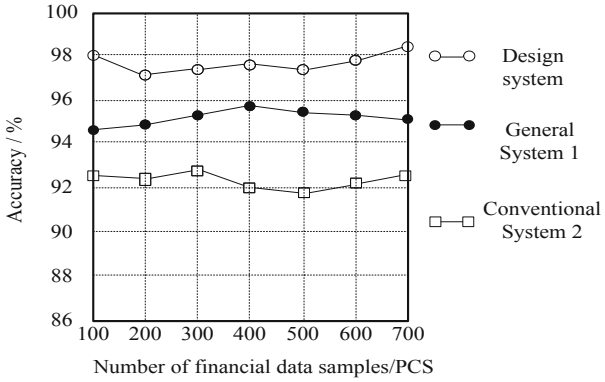

Fig. 6. Experimental comparison results of the accuracy of financial data classification

It can be seen from the above figure that the average accuracy of the design system is 97.8%, the average accuracy of the other two systems are 95.1% and 92.3% respectively, and the accuracy of data classification behavior of the design system is increased by 2.7% and 5.5% respectively, which improves the proportion of correct classification of normal data and abnormal data.

4 Conclusion

In order to improve the classification effect of financial data, this paper designs an accurate classification management system of enterprise financial and economic data. Classify the financial and economic data of enterprises, and the classification results of financial data are more accurate and reliable. In the hardware, the front end, middle layer, server end and enterprise financial and economic data display end are used to form the overall architecture of the system, optimize the data memory of the server end, and transform the serial communication circuit of the development board; In the software design, abnormal financial data are filtered, a decision tree is established for each sample

data, the utility function value is learned through the membership of the decision tree, and the optimal classification category is selected for the data through the random forest classifier. The experimental results show that the designed system improves the recall and accuracy of data classification, reduces the false positive rate, and the financial data classification results are more accurate and reliable. However, there are still some deficiencies in this design system. In the future research, the characteristic parameters of financial data will be filled in the thing characteristic table. Through the intelligent filling of characteristic parameters, the data of classification management will be more accurate.

References

1. Cui, B., Gao, J., Tong, Y., et al.: Progress and trend in novel data management system. J. Softw. **30**(1), 164–193 (2019)
2. Zhao, Z., Shen, Z.: An interactive analysis framework for multivariate heterogeneous graph data management system. Data Anal. Knowl. Discov. **3**(10), 37–46 (2019)
3. Ma, L., Wang, J., Chen, H.: Business data security of system wide information management based on content mining. J. Comput. Appl. **39**(2), 488–493 (2019)
4. Zang, H., Zhao, Q., Li, G., et al.: Design and implementation of data acquisition and management system of agronomic trait for maize. J. South. Agric. **50**(11), 2606–2613 (2019)
5. Li, S., Li, Z., He, Y., et al.: Design and implementation of information management system for ballastless track monitoring data. Railway Stand. Des. **63**(9), 28–33 (2019)
6. Zhu, F., Guo, J.-F., Cao, L.: Hierarchical recognition of data multi label features based on classification rule mining. Comput. Simul. **38**(4), 310–314 (2021)
7. Li, W., Zhao, F.: Application of data statistics management system in hospital performance management. Bull. Sci. Technol. **35**(2), 178–182 (2019)
8. Li, T., Qiu, W., Liu, Y.: Development and application of refined management system of tunnel geological information based on data drive. Tunnel Constr. **39**(1), 68–74 (2019)
9. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9
10. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)