[1]Yunmin Zhu

Research on Intelligent Financial Statement Analysis and Anomaly Identification Techniques by Fusing Multi-source Data

Abstract: - In order to accurately assess the financial status of a company and identify potential anomalies, this paper first implements unsupervised classification of financial transaction data based on Support Vector Machines, which automatically classifies the data into normal and abnormal categories. Histograms are introduced in combination with LightGBM to quickly fuse data from multiple sources. The most suitable first layer is selected by different algorithms, and the outputs of these algorithms are combined with industry-wide common abnormal features as inputs for LightGBM's second layer identification. With this two-layer structure, the model not only takes into account the industry characteristics, but also the common anomaly features. Empirical results show that in the accuracy of smart financial statement generation, the sensitivity of this paper's model iterates to 99.99% at 41.25% specificity, and the accuracy of this paper's model is as high as 0.98 when dealing with financial private information, macroeconomic, and market information.In the identification of financial transaction anomalies, the number of anomalous weeks is identified to be 24, 29, 34, and 36, and the fusion of multi-source data effectively identifies the large amount of financial transactions, fluctuating transactions and other suspicious abnormal transactions.

Keywords: support vector machine; unsupervised classification; anomaly features; multi-source data; intelligent financial statement

1. Introduction

With the continuous development of information technology and the booming of the big data era, financial statement analysis is also gradually developing in the direction of intelligence [1-2]. The traditional financial statement analysis method is limited to a single data perspective, which cannot fully reflect the real financial status of the enterprise. In order to more accurately assess the financial status of enterprises and identify potential abnormalities in financial data, financial statement identification technology that integrates data from multiple sources is a hot topic nowadays [3]. By integrating information from different data sources, it provides enterprises with more accurate and comprehensive financial management. Intelligent financial statements can also help enterprises to discover financial anomalies and take timely measures to make adjustments to avoid financial risks [4-5].

Many scholars have already conducted in-depth research in this field. These multi-source data are processed and analyzed by applying advanced machine learning algorithms such as Support Vector Machines and LightGBM. These algorithms can help to extract useful information from the massive amount of data, providing companies with a more comprehensive financial perspective and stronger risk prevention capabilities [6-7]. A study in order

[1] [1]School of Business, Sichuan University Jinjiang College, 620860, Sichuan, China.
Email: zhuyunmin502@163.com

to reduce the default risk of p2P companies, the model of Light-Gradient Boosting Machine Algorithm was used to analyze a large amount of sample data from Renrendai, exploring the integration of the basic LightGBM model as well as a linear mixture to build an optimal default risk identification model.The LightGBM algorithmic model has a prediction accuracy of the test set of 80.25%, which can accurately identify more than 80% of users [8]. There are also studies that use big data technology, support vector machine, Logit model and information fusion to conduct in-depth research on corporate financial risk. Among them, the selection of financial risk indicators has a great impact on the monitoring results of SVM-based FRM model, Logistic regression-based FRM model can effectively categorize financial risks, and information fusion-based FRM model adopts a fusion algorithm to fuse different information sources. In practice, it can effectively manage and categorize corporate financial risks, with classification accuracy rates of 90.22% and 90.88%, respectively [9]. In addition, a series of financial intervention practices in the U.S. are investigated, combined with the loan data provided by the U.S. government's financial intervention department, and mined for data from the generalized C4.5 algorithm of the decision tree algorithm. A decision tree is generated and converted into classification rules. During the construction of the decision tree, the algorithm selects the best split attributes based on the information gain rate. Ultimately, a decision tree is generated by the C4.5 algorithm that clearly shows the relationship between various loan application characteristics and loan approval or rejection. It can help financial institutions to quickly and accurately evaluate loan applications and improve the efficiency and accuracy of financial interventions. Data mining techniques by applying data fusion and information entropy are used to better understand market trends and risks [10]. There is also research that proposes an attentive and normalized deep learning approach to predict financial distress, using multimodal data to accurately predict the financial distress of firms. Heterogeneity within and between modalities was taken into account and ratio awareness, report awareness and neighbor awareness were designed accordingly. These three mechanisms are able to intelligently extract key information from financial indicators, firms' current reports, and inter-firm networks, respectively. To address the problem that it is difficult to distinguish between complementary and redundant information between patterns, a regularization method based on conditional entropy is designed to guide the approach of focusing on complementary information while eliminating redundant information. By integrating diversified data sources and supplemented with a refined information screening mechanism, the method provides a more accurate and reliable tool for the prediction of corporate financial distress [11].

Intelligent financial statement strategy and anomaly identification technology is a development trend in the financial field, which can help enterprises better understand their financial situation, detect and respond to financial anomalies in a timely manner, and improve the accuracy and reliability of financial statements. In this paper, we propose a multi-source data fusion intelligent financial statement analysis method, which combines one-class support vector machine s and LightGBM to accomplish unsupervised classification and anomaly detection of financial data. The method first uses one-class SVMs to screen the approximate range of anomalies, followed by LightGBM to improve the accuracy. The two-layer SVM-LightGBM hybrid model, combined with industry and general features, can efficiently identify precise anomalies in financial data with good generalization ability. By fusing information from different data sources, more comprehensive and accurate financial statement data can be obtained, which in turn improves the accuracy and reliability of financial analysis. At the same time, combined with the anomaly identification technology, the anomalies in the financial statements can be detected in a timely manner, providing strong support for the enterprise's financial risk management.

2. Unsupervised scope classification of financial data

In the field of intelligent financial statement research and anomaly identification, support vector machine plays an important role with its unique advantages. Support vector machine is a new generation of statistical learning model that solves machine learning problems with the help of optimization methods based on VC dimension theory and structural risk minimization principle. In financial statement analysis, two types of classification problems are often faced, such as the distinction between normal and abnormal transactions. The traditional c-SVM and nu-SVM, as supervised learning algorithms, are able to establish optimal decision surfaces for known categories of normal transaction data, and then categorize or predict emerging transaction data. However, in anomaly identification for smart financial statements, unsupervised learning problems are often faced, e.g., anomalous financial activity identification without a priori knowledge. The one-classSVMs are then mainly applied to the unsupervised learning problem for problems such as image analysis, intrusion detection and anomaly detection [12].

The two-dimensional optimal hypersphere classification surface is shown in Fig. 1.One-elassSVMs are categorized into two types of hyperspheres and hyperplanes, with the former looking for the smallest hypersphere in the feature space that contains most of the samples. In financial statement analysis, one-class SVMs of the hypersphere type are utilized to find a hypersphere that is as small as possible, which is able to contain all the training samples of normal financial data. The center c and radius R of this hypersphere can be found by an optimization algorithm that defines the boundary of a normal financial data. The latter hyperplane separates the samples with a maximum edge distance for high-dimensional density estimation.

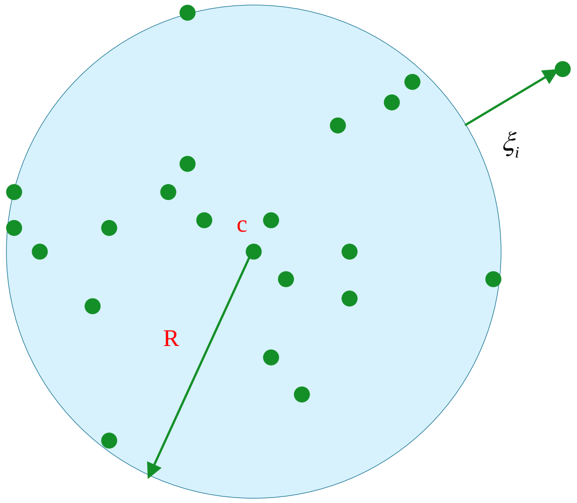

Figure 1 Two-dimensional optimal hypersphere classification surface

In d -dimensional space, n training samples are x_i, $i=1,2,...,n\ x \in R^d$, construct a hypersphere with center c and radius R that is as small as possible and contains all the training samples, and for each training sample there is $\left\| x_i - c \right\|^2 \geq R^2$, find the minimum value of R that is the minimum hypersphere to be found. Since there may be anomalies in the samples indicating noncompliance or fraud in the financial statements [13]. Then the distance between the training center c and the sample should be less than or equal to an amount slightly greater than R. For the two-dimensional problem, a new set of non-negative scalar variables $\left\{ \xi_i \right\}_{i=1}^{N}$ is introduced to the separating hypersphere:

$$\left\{ \xi_i \right\}_{i=1}^{N} \leq R^2 + \xi_i, i = 1, 2, \cdots, n \tag{1}$$

ξ_i Called the slack variable, it is a measure of how far a data point deviates from the ideal conditions of the minimum hypersphere. A larger ξ_i indicates that the point is farther away from the hypersphere. Minimization of radius R is done by constructing a generalized function:

$$\Phi\left(R, \xi\right) = R^2 + C\sum_{i=1}^{N} \xi_i \tag{2}$$

Where the penalty parameter C is a compromise between the misclassification rate and the hypersphere volume, taking a value between 0, 1. In any case, the penalty parameter C can be specified by the user or determined experimentally by using a training set.

Equation (2) is solved for the minimal value of R under the constraint equation (1), which can be solved with a Lagrangian function [14]. As follows:

$$L\left(R, \xi, c, \alpha_i, \beta_i\right) = R^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left\{ R^2 + \xi_i - \left(x_i - c\right)^2 \right\} - \sum_{i=1}^{N} \beta_i \xi_i \tag{3}$$

where the auxiliary nonnegative variable α_i, β_i is the Lagrange multiplier. For R, ξ, c with respect to the minimal value of the Lagrange function, the partial derivatives of R, ξ, c and make it equal to zero have, respectively:

$$\sum_{i=1}^{N} \alpha_i = 1 \tag{4}$$

$$c = \sum_{i=1}^{N} \alpha_i x_i \tag{5}$$

where $0 \le \alpha_i \le c$. The dyadic problem of the above problem is to find the maximum of the function under the constraints of Eq. (4) and Eq. (5):

$$L(\alpha) = \sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i x_i^T x_i - \sum_{i,j=1}^{N}\alpha_i\alpha_j x_i^T x_j \tag{6}$$

The dyadic problem defined in Eq. is expressed based on the training data, and the maximization of function $L(\alpha)$ depends only on the set of pattern dot products. In general, the support vectors are a subset of the training samples are sparse. Since it is difficult to find a suitable hypersphere for the actual samples, the kernel function $K(x_i, x_j)$ is utilized instead of the inner product operation $(x_i \cdot x_j)$ to transform the non-spherical distribution in the low-dimensional space into a spherical distribution in the high-dimensional space [15-16]. This process can be expressed as:

$$L(\alpha) = \sum_{i=1}^{N}\alpha_i K(x_i, x_j) - \sum_{i,j=1}^{N}\alpha_i\alpha_j K(x_i, x_j) \tag{7}$$

Equation (7) can be used to find the optimal solution by quadratic optimization method, if α_i^* is the optimal solution, then for most samples α_i^* is zero, and the samples that α_i^* are not zero are the support vectors, which support the hypersphere. Substitute the test samples into $\|x_i - c\|^2 \ge R^2$.

If the inequality is not satisfied, it is an outlier and finally the optimal classification function is obtained:

$$f(x) = \operatorname{sgn}\left\{K(x_i, x_j) - 2\sum_{i=1}^{N}\alpha_i K(x_i, x_j) + 2\sum_{i,j=1}^{N}\alpha_i\alpha_j K(x_i, x_j) - R^2\right\} \tag{8}$$

In this paper, the radial basis function is used as the kernel function to nonlinearly transform the input financial data into a high-dimensional feature space [17]. The results of financial data segmentation are as follows:

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right) \tag{9}$$

Based on the above, the extraction and classification of financial data is completed to facilitate subsequent work such as anomaly identification and analysis.

3. Intelligent financial statement analysis and exception identification

3.1 LightGBM algorithm for multi-source data fusion

Unsupervised classification of data using one-class SVMs. Automatically learning and recognizing normal patterns in the data without a priori knowledge can initially screen out potentially problematic transaction records. Relying on one-class SVMs alone cannot capture all anomalies. Therefore, the introduction of histogram technology through LightGBM improves the efficiency and accuracy of model training. This feature gives LightGBM a significant advantage when dealing with large-scale financial data.

The LightGBM algorithm, as an improved version of the gradient boosting decision tree algorithm, demonstrates significant advantages in data processing and analysis with its high efficiency and accuracy [18]. LightGBM not only employs partial samples for the computation of information gain, but also combines the built-in feature downscaling technique, which significantly reduces the computational cost of each information gain.The LightGBM The algorithmic flow of the model is shown in Fig. 2.In the study of corporate financial statements, this efficiency enables LightGBM to quickly fuse and analyze data from multiple sources.LightGBM employs a leaf-wise growth strategy, which involves finding the node with the largest splitting gain among the current leaf nodes to split. This strategy provides higher accuracy while maintaining the same number of splits, and is well suited for fine-grained analysis of complex data in smart financial statements. The search for the optimal splitting point may be affected by the excessive number of splitting points, samples and features, which leads to an increase in computational complexity [19]. To solve these problems, the LightGBM algorithm effectively reduces the computational burden caused by the excessive number of split points, samples and features by introducing the histogram algorithm, the gradient-based one-sided sampling algorithm, and the mutually exclusive feature bundling algorithm, which can accurately and efficiently handle multi-source data in smart financial statements.The LightGBM algorithm treats the value of the loss function's current negative |gradient as an approximation of residuals, and uses this value to progressively fit a regression to the residuals. The value is used to fit a regression tree step by step, and then sequentially incremental, generating the next tree in the decision-making process, and finally the run results are weighted and summed according to the weights to arrive at the final result.

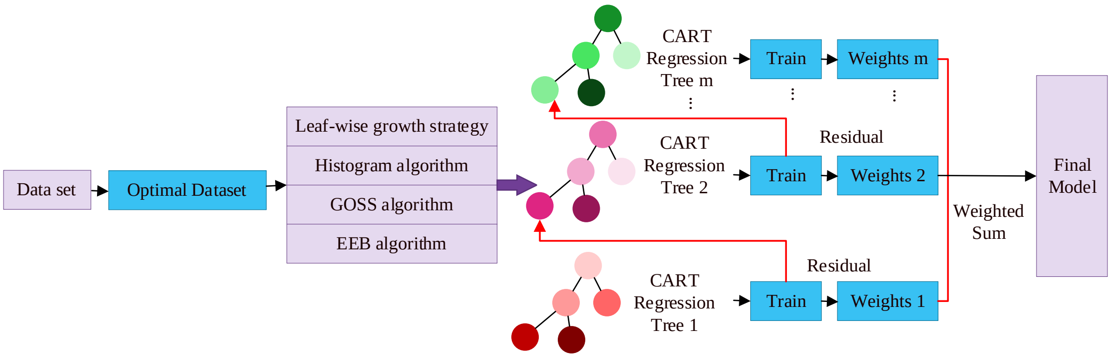

Figure 2 LightGBM algorithm flow

3.2 Financial statement anomaly identification model

At present, the identification of financial data anomalies is often based on industry-wide data using a single algorithm for modeling and identification, and the results of identification are difficult to show the differences in financial data of different industries. In fact, through the analysis of financial data, it is not difficult to conclude that the financial data anomalies of different industries are significantly different in the sensitivity performance of different indicators, which leads to the model effect cannot be further refined [20-21]. Therefore, it is necessary to establish a hybrid model of financial data anomaly identification that processes financial data by industry.

The intelligent financial statement anomaly identification program of multi-source data is shown in Figure 3, and the specific process of financial data anomaly identification model construction is to use the financial anomaly indicators of each industry to select the most suitable model for each industry using the random forest algorithm, SVM algorithm and logistic regression algorithm for the parameterization process, which constructs the first layer of the model. As the output of layer 1 is only the probability value, the number of features is too small for the overall amount of data, which is easy to cause insufficient model generalization ability and reduce the accuracy and authenticity of the model results. Therefore, the output of layer 1 is spliced with the abnormal financial indicator system common to the whole industry as the input of layer 2 of the model, and the LightGBM algorithm is used in layer 2 to identify the financial anomalies again [22]. The LightGBM integration algorithm occupies less memory space and operates faster than the previous integration algorithms, and it has not been used for the time being in the large-scale detection of abnormalities in financial data. It has a good application prospect.

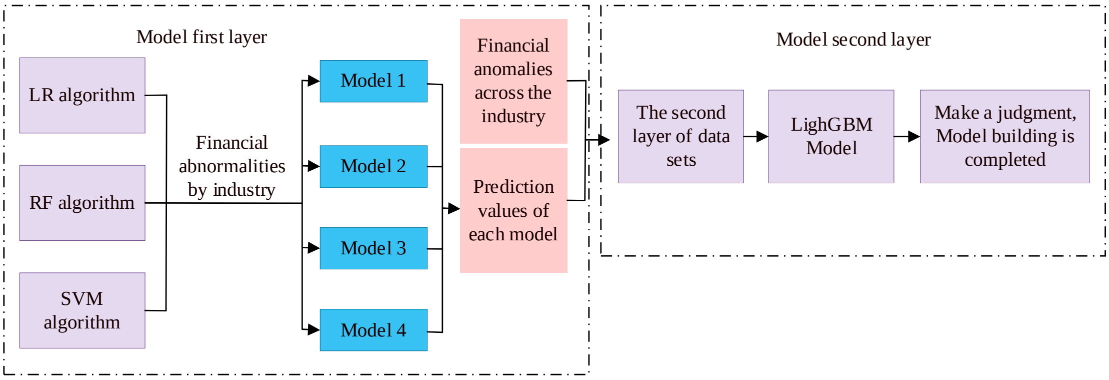
Model first layer
LR algorithm
RF algorithm
SVM algorithm
Financial abnormalities by industry
Model 1
Model 2
Model 3
Model 4
Financial anomalies across the industry
Prediction values of each model
Model second layer
The second layer of data sets
LighGBM Model
Make a judgment, Model building is completed

Figure 3 Intelligent financial statement anomaly recognition based on multi-source data

4. Empirical study of intelligent financial statement analysis and anomaly identification

4.1 Data sources and indicator design

4.1.1 Platform parameters

In order to evaluate the performance of the algorithm, the collection and cleaning of company financial data is required. The paper selects the data related to listed companies published to the public in the RESSET financial database. Among all the companies from 2010-2023, 100 listed companies were screened by excluding the relevant companies with missing data. At the same time, the paper also screened 100 companies and their financial data from the companies with financial fraud in that time period, which together form a dataset containing 200 companies. The simulation platform parameters are shown in Table 1, with 2000 training sets and 1000 test sets in Windows 7 operating system.

Table 1 Simulation platform parameters

Name	Parameter
Operating system	Windows7
Programming environment	Matlab2015b
Data analysis environment	SPSS
CPU	Intel Dual-core 3.2 GHz
Memory	128GB
Training set/pcs	2000
Test set/pcs	1800

4.1.2 Interpretation of variables

The explanatory variables in this paper are based on financial indicators, adding financial public information indicators, which are more suitable for financial anomaly identification, in order to improve the accuracy of financial anomaly identification. The construction of the core explanatory variables system is shown in Table 2, including financial private information, macroeconomics, and market, and in the part of financial private information, it is divided into three categories: solvency, operational capacity and profitability. Solvency indicators, such as current ratio, equity multiplier, gearing ratio, etc., which can reflect the ability of enterprises to repay debts in the short term, which is the key to assess the robustness of enterprises. Operational capacity indicators, such as inventory turnover ratio and current asset turnover ratio, determine the efficiency of the enterprise's asset management and operational status. Profitability indicators, such as cost of sales ratio and net sales margin, reflect the profitability and economic efficiency of the enterprise. Indicators in the macroeconomic section are in the macroeconomic environment, including Ml growth rate, CPI growth rate and so on. It reflects the overall economic activity and inflation. Finally the market section, indicators such as total market capitalization/total liabilities and rate of return increase reflect changes in the macroeconomic environment and provide companies with real-time information on market feedback and competitive dynamics. This multi-source financial information metrics table provides an all-encompassing framework for intelligent financial statement analysis and anomaly identification by integrating private and public financial information.

Table 2 Multi-source financial information indicators

Level	Category	Indicators	
Financial private information	Solvency	Current Ratio(X1)	Asset-Liability Ratio(X5)
		Equity Multiplie(X2)	Current Liabilities / Total Assets(X6)
		Quick Ratio(X3)	Operating Cash Flow / Current

			Liabilities(X7)
		Interest Coverage Ratio(X4)	Accounts Receivable Turnover,(X8)
	Operational capability	Inventory Turnove(X9)	Current Asset Turnover(X11)
		Fixed Asset Turnover(X10)	Total Asset Turnover(X12)
	Profitability	Cost of Sales Rati(X13)	Net Profit Margin on Sales(X15)
		Return on Current Assets(X14)	Return on Fixed Assets(X16)
Financial public information (macroeconomics)	Macroeconomic capacity	M1 Growth Rate(X17)	CPI Growth Rate(X18)
		Industrial Value-Added Growth Rate(X19)	Interest Rate(X20)
		M2 Growth Rate(X21)	PPI Growth Rate(X22)
		Interest Rate Growth,(X23)	Employment Rate(X24)
Financial public information (market)	Market capacity	Total Market Value / Total Liabilities(X25)	A-Shares / Total Share Capital(X26)
		Yield Growth(X27)	GDP Growth Rate(X28)

4.2 Intelligent Financial Statement Accuracy and Standard Deviation Analysis

In this paper, we test the financial data 100 times, and Fig. 4 shows the average accuracy values of the fused multi-source dataset. The generalization ability of the model in this paper is evaluated by comparing the performance on the training, validation and test sets. The average accuracy value of the training mean is at 0.87%, the average accuracy value of the validation mean is at 0.82%, and the average accuracy value of the test set is at 0.825%. The performance on these three datasets is similar, indicating that the model in this paper is able to learn the training financial data well as well as generalize to unseen financial data. By observing the distribution of the results of the 100 tests, the distribution shows a relatively concentrated trend when the concentration of the distribution of the results of the model in this paper is reflected in the small range of fluctuations in the model output in multiple tests, without excessive deviations or outliers. When dealing with complex and changing multi-source financial data, it can maintain high performance stability and provide stable and reliable financial statement analysis services for enterprises.

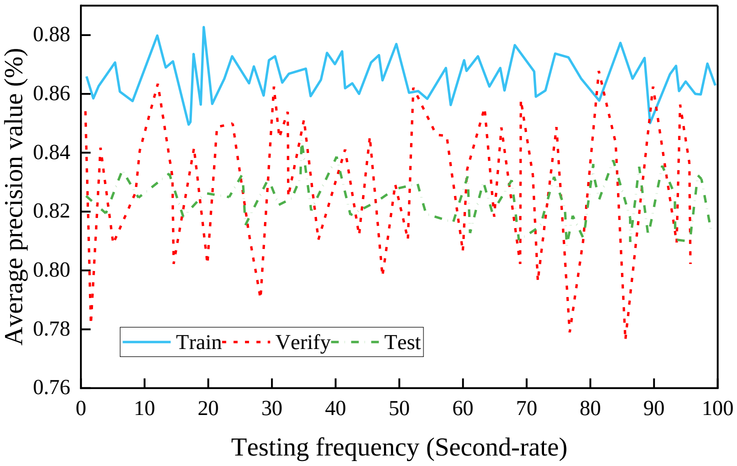

Figure 4 Average precision value of fusion multi-source data sets

The results of the standard deviation comparison are shown in Figure 5, where the standard deviation is below 0.083 in all the tests conducted.This data not only reflects the stability of the model predictions, but also the effectiveness of data fusion. In financial statement analysis, diversity of data sources is the norm, including internal company financial data, market data, and industry reports. The average precision value of the standard deviation of the training means ranges from 0.012% to 0.028%, the average precision value of the validation means ranges from 0.021% to 0.083%, and the average precision value of the test set ranges from 0.022% to 0.081%. By fusing these data, the model is able to perform well on different datasets, which proves its strong generalization ability to capture the intrinsic laws and patterns of financial data and play a key role in risk management and strategic planning.

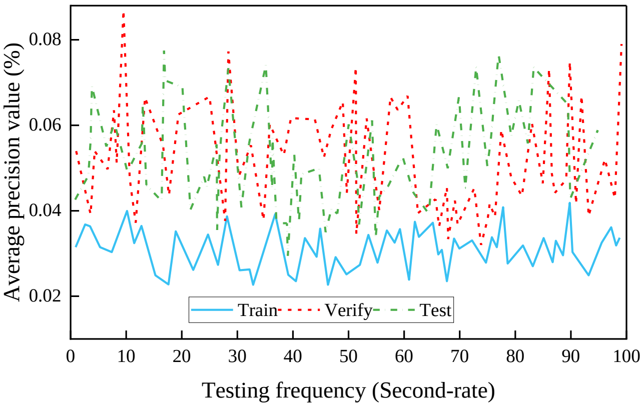

Figure 5 Standard deviation comparison results

4.3 Assessment of the accuracy of financial statement generation

The ROC curve is an effective tool to visually assess the accuracy of financial statement generation. In the analysis of intelligent financial statements fusing multi-source data, the processing results of the algorithms proposed in the paper are visualized using ROC curves, and a comparison of the ROC curves of each model is shown in Fig. 6. It can be clearly seen that the classification performance of each algorithm is roughly equivalent, but the method proposed in this paper has improved in performance. In this paper, the sensitivity of the model is iterated to 99.99% at 41.25% specificity, and the sensitivity of the multilayer perceptron MLP model is iterated to 99.99% at 78.34% specificity. The sensitivity of the XGBoost gradient boosting algorithm is iterated to 99.99% at 79.66% specificity, and the sensitivity of the Logistic Regression logistic regression algorithm is iterated to 99.99% at 79.61% specificity. This is due to the fact that the model in this paper effectively integrates financial multi-source data, which enables the model to capture and parse the financial risks of listed companies from more dimensions. Compared with other algorithms, this paper's method demonstrates higher accuracy and reliability in smart financial statement generation, providing investors and decision makers with more accurate risk identification.

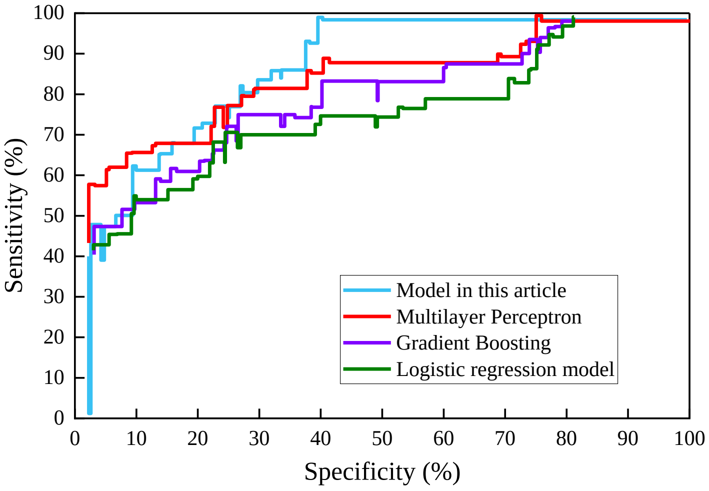

Figure 6 Comparison of ROC curves of various models

The performance of the three types of financial information generation is shown in Table 3. When it comes to financial private information indicators, this paper's model achieves the highest scores in all four indicators, which indicates that the model has excellent accuracy and stability when dealing with internal financial data of enterprises. In terms of precision rate, this paper's model scores 0.98, which is much higher than the multilayer perceptron's 0.84. In terms of macroeconomic indicators, this paper's model also performs outstandingly in this category, with a precision rate of 0.92, a recall rate of 0.91, an F1 value of 0.94, and an F2 value of 0.91. This indicates that this paper's model can effectively take macroeconomic factors into account, providing financial analysis with a more comprehensive Perspective. The model in this paper is optimal based on the data results, indicating that the model can keenly capture the market dynamics and make accurate financial forecasts accordingly. The accuracy rate is 0.95, far exceeding the 0.73 of the multilayer perceptron, 079 of the gradient boosting algorithm, and 0.80 of the logistic regression, which demonstrates its excellent ability in market data analysis. It proves the excellent performance of the model in this paper in dealing with different types of financial information. Whether it is in dealing with internal financial data, macroeconomic data or market dynamic data, it can be a comprehensive and accurate data analysis ability, which has a high application value in the field of complex financial data analysis.

Table 3 Performance of three types of financial information

Indicators	Multilayer Perceptron	Gradient Boosting Algorithm	Logistic Regression Model	Model in this paper
Financial private information precision	0.84	0.90	0.91	0.98
Financial private information recall	0.78	0.85	0.89	0.96
Financial private information F1 value	0.82	0.88	0.88	0.95
Financial private information F2 value	0.75	0.82	0.90	0.92
Macroeconomic precision	0.70	0.80	0.88	0.92
Macroeconomic recall	0.74	0.76	0.85	0.91
Macroeconomic F1 value	0.75	0.83	0.90	0.94
Macroeconomic F2 value	0.75	0.79	0.87	0.91
Market precision	0.73	0.74	0.82	0.95
Market recall	0.76	0.84	0.84	0.93
Market F1 value	0.81	0.90	0.88	0.96
Market F2 value	0.79	0.84	0.90	0.94

4.4 Effectiveness of Anomaly Recognition Technology Application

The account data was generated as a time series with a time span of weeks, with August 1, August 7, 2021 as week 1, and cycling sequentially to the end of April 1, 2023, for a total of 82 weeks. The number of transactions, the total amount of transactions, and the transaction dispersion coefficient are calculated for each week, and the three eigenvalues obtained are normalized. The penalty factor C and the parameter y of the radial basis function are set to C = 0.1 and y = 0.01, respectively, by the cross-validation method. The normalized eigenvalues are substituted into SVM-LightGBM to detect outliers.

The financial transaction anomalies are identified as shown in Figure 7, where the horizontal axis represents the transaction dispersion coefficients and the vertical axis corresponds to the dates of the financial transactions.

The model in this paper successfully identifies suspicious anomalies in the financial statement data, which are mainly concentrated in some specific reporting periods, as well as in the abnormal transaction weeks, which are the suspicious anomalies in the financial statements of the 11th, 24th, 30th, 29th, 32nd, 34th, 36th, and 82nd weeks, respectively. When the input financial statement data vectors were nonlinearly transformed into a high-dimensional feature space using the SVM-LightGBM algorithm, it was found that the very large and very small values in the input vectors fell outside of the hypersphere constructed by the algorithm, which further corroborated that there were anomalies in the transaction data for these weeks.

Special attention is paid to the extremely large outliers, which occur in weeks 24, 29, 34, and 36, which are the same weeks as the outliers identified in the first field. A peak in the total transaction amount is reached in week 29, where multiple large financial transactions occur, indicating that the model in this paper effectively monitors large transaction activity in the account. Week 24 not only has a relatively high total transaction amount, but also has the highest number of transactions, and it is a year-end week, running from December 27, 2022 to January 2, 2023, which may suggest the existence of an anomaly of concentrated year-end spending. Weeks 34 and 36 are in close proximity to each other, both located in late March 2023. During this period, there is a high level of transaction activity with a large total amount of money, as well as a high coefficient of dispersion of the transactions. By applying the SVM-LightGBM anomaly identification technique, the financial statements were successfully identified with large financial transactions as well as suspicious anomalous transactions with high transaction fluctuations. This not only improves the efficiency of financial analysis, but also provides strong support for timely detection and resolution of potential financial risks.

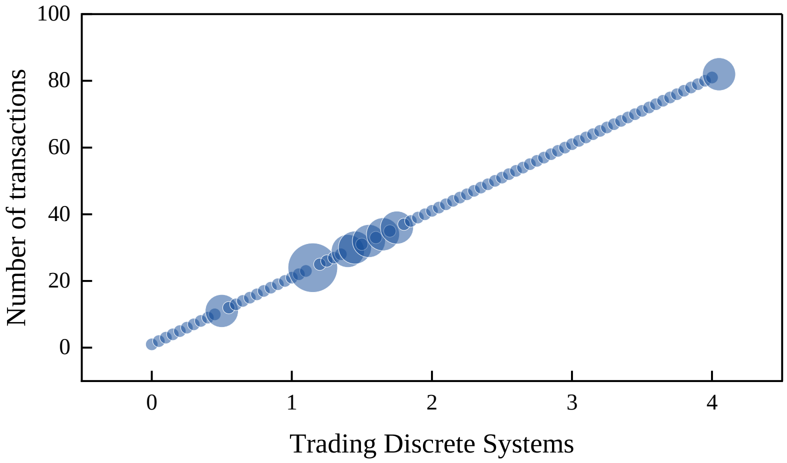

Figure 7 Identification of abnormal points in financial transactions

5. Conclusion

The core of this paper lies in combining multiple state-of-the-art algorithms in order to achieve in-depth analysis and anomaly detection of financial statements. Based on statistical learning theory, a class of support vector machines is used for unsupervised classification of financial transaction data. Histogram technique is introduced using LightGBM to discretize continuous data values and fuse multiple sources of financial statement data data. In this way, a hybrid SVM-LightGBM model with a two-layer structure is constructed to identify industry-specific anomalies through different algorithms for initial screening at the first layer. Through empirical analysis, on the fusion of multi-source data, the average precision value of the training mean of this paper's model is at 0.87.3, the average precision value of the validation mean is at 0.82.4., and the average precision value of the test set is at 0.82.5. On macroeconomic indicators, this paper's model's precision rate is 0.92, recall rate is 0.91, and the F1 value is 0.94, and the F2 value is 0.91. This multi-source data fused intelligent financial statement analysis method provides a comprehensive and efficient financial anomaly detection solution for enterprises, which helps them to detect and respond to potential financial risks in a timely manner.

References

[1] Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science*, *176*, 3037-3046.

[2] Nizioł, K. (2021). The challenges of consumer protection law connected with the development of artificial intelligence on the example of financial services (chosen legal aspects). *Procedia Computer Science*, *192*, 4103-4111.

[3] Fei, X. (2023). Application of real-time data processing system of Internet of Things based on blockchain technology in the financial field of Yangtze River Delta urban agglomeration. *Soft Computing*, *27*(14), 10121-10131.

[4] Leonov, P., Kozhina, A., Leonova, E., Epifanov, M., & Sviridenko, A. (2020). Visual analysis in identifying a typical indicators of financial statements as an element of artificial intelligence technology in audit. *Procedia Computer Science*, *169*, 710-714.

[5] Shao, M., & Fan, H. (2024). Identifying the systemic importance and systemic vulnerability of financial institutions based on portfolio similarity correlation network. *EPJ Data Science*, *13*(1), 9.

[6] Mahmood, F., Qadeer, F., Saleem, M., Han, H., & ArizaMontes, A. (2021). Corporate social responsibility and firms' financial performance: A multi-level serial analysis underpinning social identity theory. *Economic Research-Ekonomska Istraživanja*, *34*(1), 2447-2468.

[7] Ren, S., Tang, G., & Jackson, S. E. (2021). Effects of Green HRM and CEO ethical leader*ship on organizations' environmental performance.* International Journal of Manpower, 42(6), 961-983.

[9] Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. *Journal of Intelligent Systems*, *31*(1), 767-779.

[9] Yue, H., Liao, H., Li, D., & Chen, L. (2021). Enterprise financial risk management using information fusion technology and big data mining. *Wireless Communications and Mobile Computing*, *2021*(1), 3835652.

[10] Gu, C. (2022). Application of data mining technology in financial intervention based on data fusion information entropy. *Journal of Sensors*, *2022*(1), 2192186.

[11] Che, W., Wang, Z., Jiang, C., & Abedin, M. Z. (2024). Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Information Processing & Management*, *61*(4), 103703.

[12] Lesouple, J., Baudoin, C., Spigai, M., & Tourneret, J. Y. (2021). How to introduce expert feedback in one-class support vector machines for anomaly detection?. *Signal Processing*, *188*, 108197.

[13] Padhi, D. K., Padhy, N., Bhoi, A. K., Shafi, J., & Ijaz, M. F. (2021). A fusion framework for forecasting financial market direction using enhanced ensemble models and technical indicators. *Mathematics*, *9*(21), 2646.

[14] Scheiner, B., & Schmitt, M. (2021). Considerations for the modeling of the laser ablation region of ICF targets with Lagrangian simulations. *AIP Advances*, *11*(10).

[15]. Zhao, J., Ahmad, Z., Mahmoudi, E., Hafez, E. H., & Mohie El-Din, M. M. (2021). A New Class of Heavy-Tailed Distributions: Modeling and Simulating Actuarial Measures *Complexity*, *2021*(1), 5580228.

[16] Tung, Y. L., Ahmad, Z., & Mahmoudi, E. (2021). The Arcsine-X Family of Distributions with Applications to Financial Sciences. *Comput. Syst. Sci. Eng.*, *39*(3), 351-363.

[17] Li, G., Zhang, Q., Lin, Q., & Gao, W. (2021). A three-level radial basis function method for expensive optimization. *IEEE Transactions on Cybernetics*, *52*(7), 5720-5731.

[18] Zhou, L., Duan, Y., & Wei, W. (2023). Research on the Financial Data Fraud Detection of Chinese Listed Enterprises by Integrating Audit Opinions. *KSII Transactions on Internet & Information Systems*, *17*(12).

[19] Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, *167*, 113913.

[20] Zhang, H., & Luo, Y. (2022). Enterprise financial risk early warning using BP neural network under internet of things and rough set theory. *Journal of interconnection networks*, *22*(03), 2145019.

[21] He, T. (2021). Research on the fusion of enterprise financial accounting and management accounting in the new economic situation. *Finance and Market*, *6*(1), 53.

[22] Gabdrakhmanova, N., Fedin, V., Matsuta, B., & Pilgun, M. (2021). The modeling of forecasting new situations in the dynamics of the economic system on the example of several financial indicators. *Procedia Computer Science*, *186*, 512-520.

ABOUT THE AUTHOR

Yunmin Zhu was born in Ya'an，Sichuan, China, in 1987. She obtained a master's degree from Southwestern University of Finance and Economics in China. She is currently working at the School of Business, Sichuan University Jinjiang College. She main research direction is the analysis and application of financial big data.

E-mail: zhuyunmin502@163.com