Financial Fraud Identification Model of Listed Companies based on Time-Series Information

LILI WANG
Lyceum of the Philippines University Manila Campus,
Manila 1002,
PHILIPPINES

Abstract: - The aim of this research is to establish a high-precision financial fraud identification model for listed companies, which is mainly based on the financial indicators of time series. Support vector machine and K-means clustering algorithm are especially used in the research process. Firstly, local linear embedding is used to reduce the dimensionality of the selected financial indicators to extract the low-dimensional characteristics. Then the samples are classified into financial fraud and non-fraud by support vector machine, and the recognition model is constructed. At the same time, the research also uses K-means clustering algorithm to analyze the pattern of financial fraud. The experiment of dimensionality reduction proves that the model has a high effect on the processing of financial data, and the error between the data after dimensionality reduction and the original data is small. In addition, the clustering effect of the model also shows a clear pattern of fraud. In practical application, the accuracy rate of this model is as high as 94.89%, showing high accuracy and recall rate, and its F1 value is 87.08%, showing its feasibility and effectiveness in practice. The results highly prove that the performance of the financial fraud identification model proposed in this study is excellent, and it has a wide application prospect in the future.

Key-Words: - Temporality, Finance, Fraud, Local linear embedding, Recognition clustering, Time-Series Information, Financial index.

Received: March 14, 2023. Revised: October 19, 2023. Accepted: December 21, 2023. Published: February 21, 2024.

1 Introduction

Identifying financial fraud in listed companies has been a significant area of research in the field of finance. Financial fraud has a serious impact on investors, companies and the market as a whole. Therefore, establishing an accurate and reliable financial fraud identification (FFI) model is crucial to protect investors' interests and maintain market stability. The study aims to investigate the identification model of financial fraud in listed companies grounded on time-series information indicators. Time-series information indicators include financial indicators, financial statements and other financial data involving time and sequence, providing a detailed description of the evolution of a company's financial position over time. [1], explored the impact of financial literacy on the identification of financial fraud in China through statistical models and machine learning models. The main methods include logistic regression, decision tree, random forest, etc. The advantage is that powerful machine learning technology can effectively mine the nonlinear relationship in the data and find the complex pattern of financial fraud. [2], used the empirical research method to study the impact of digitalization on financial inclusion through questionnaire survey and interview of BOP population, combined with economic model. Its advantage is that it combines both qualitative and quantitative research methods to obtain comprehensive and in-depth insights. Through the combination of temporal information indicators, the study hopes to understand and identify financial fraud more comprehensively. The innovation of the study lies in the adoption of Support Vector Machine (SVM) classifier for classification of financial frauds and the combination of K-means algorithm for cluster analysis. SVM is a powerful classification algorithm that can construct decision boundaries in high-dimensional space, so as to effectively distinguish between financial fraud and non-fraud samples. Some studies have established multiple linear regression models and support vector machine classification models, combined with financial intermediary measures to study the factors affecting the ratio of loans to deposits, and verified the performance advantages of support vector machines in the financial field, [3]. In this study, the potential patterns and clusters are also found through the cluster analysis of K-means algorithm, which helps to reveal the characteristics and rules of financial fraud.

The advantage of the research method is to use the financial indicators of time series to investigate the dynamic characteristics of the company's financial data over time, which can more accurately reflect the high performance of the model. Using K-means algorithm, the pattern of financial fraud can be identified, and some inherent characteristics and laws of fraud can be further revealed. The combination algorithm is used to improve the effectiveness of the model, enrich the explanation of the model, and enhance the recognition ability of complex financial fraud.

This research has the following main contributions, one of which is the introduction of temporal information indexes, by introducing temporal information indexes, the research provides an in-depth understanding of the evolution of financial fraud over time, which makes the model more accurate and generalizable; two of which is the adoption of SVM classifier as the core algorithm of the FFI model, which has strong learning and generalization capabilities, and helps to better classify and predict; The third is that by combining the K-means algorithm for clustering analysis, the study can identify potential patterns and clusters of financial fraud behavior, providing reference for further analysis. The research is mainly composed of four parts. The first part is to summarize the advantages and disadvantages of domestic and foreign scholars on financial fraud; the second part is to select the relevant time-series information indicators, combined with SVM classifier and K-means algorithm to construct a FFI model for listed companies; the third part is to validate the data downgrading of the established model through the dataset, respectively, as well as the identification performance; the fourth part is to summarize and analyze the results of the study and put forward the current study also has shortcomings and gives the future research. Deficiencies and gives future research directions.

2 Experimental Methods

The research selects the financial statements and other financial data of 250 listed companies as the experimental data set. The data set contains a total of 400 samples, which covers 16 types of violations. The fraud of financial statements includes four types of violations: fictitious profits, fictitious assets, false records and general accounting improper treatment. Non-financial statement fraud includes 12 kinds of behaviors such as delayed disclosure, investment violation, insider trading, illegal guarantee and so on. The data set is divided into training set and test set, in which the training set accounts for 70%, and the test set accounts for 30%. The performance of the constructed model was verified through dimensionality reduction experiment and classification experiment, and the advanced nature of the constructed financial fraud model was verified by comparative experiment.

3 Literature Survey

Financial statements of listed companies play an important role in business as a communication tool to convey the financial position of a company for a specific accounting period to convey information to the users of the statements. Many scholars have proposed different models and methods to identify financial fraud by analyzing financial data and other related information. [4], proposed a new fraud risk assessment model for listed companies that integrates evidence from multiple internal and external sources. Internal evidence is collected through machine learning methods and external evidence is obtained through web crawlers. An evidence theory-based approach was used to integrate the multi-source evidence and establish a fraud risk assessment model. The model was applied to the fraud risk assessment of listed companies in China, and the results showed that the model can effectively assess the fraud risk of listed companies and has higher accuracy, recall, and F-1 measure compared to the traditional probabilistic model.

The purpose of the study by, [5], team was to evaluate the influence of intellectual capital (IC) on financial statement fraud of listed companies in Tehran Stock Exchange (TSE). Therefore, the research hypotheses were tested using logistic regression model with integrated data technique. In addition, some robustness checks were used to guarantee the correctness of the results. The outcomes show that there is an obvious negative relationship between IC and its components including HC, SC, RC and CC efficiency and financial statement fraud. This denotes that by investing in ICs and their components, fraud in the financial statements of commercial companies is reduced.

The study by, [6], team aimed to analyze the obligations of companies listed on the Athens Stock Exchange and to identify and manage the business risks, as well as disclose these risks to investors. Effective risk identification and management protects the business and adds value to shareholders and other interested parties. Over the past few years, many organizations have failed because of irregularities and fraud, with adverse effects on stakeholders. The failure of these organizations was attributed to the incompetence of senior management and the failure of the board of directors to identify and disclose problems and risks. This study assesses the risk of disclosure through a content analysis of the 2005-2011 annual reports of non-financial companies listed on the Athens Stock Exchange.

[7], study aims to integrate two streams of literature related to fraud, namely, relationship and overconfidence, to construct a fraud triangulation framework that explains the mechanisms of fraud commissioning and detection. A binary probit model was used for the analysis. The study found that overconfidence triggers fraudulent behavior and exacerbates fraud detection; overconfidence mediates the relationship between fraud and relationships; the "white side" of relationships comes from alumni networks, and the "black side" comes from kinship networks; and overconfidence leads to accounting and disclosure fraud and facilitates the detection of disclosure fraud. It facilitates the detection of disclosure fraud. Relationships inhibit fraud in management and disclosure but exacerbate fraud detection in the presence of fraud; overconfidence leads to fraud in state-owned and non-state-owned firms and facilitates fraud detection in state-owned firms.

[8], study aimed to explore managers' willingness to manage corporate surplus through business activity intervention. Using a sample of companies listed on the Stock Exchange of Thailand, the real activity manipulation of discretionary expense manipulation, sales manipulation, and production manipulation was examined. The research identified companies reporting small earnings or small earnings growth as suspect companies. The findings suggest that executives of the suspected firms engage in real activity manipulation to avoid losses or to smooth the firm's earnings. Given that business activity interventions are difficult to detect, market regulators in emerging countries need to monitor such practices and introduce legislation relating to this form of corporate fraud.

[9], argue that financial fraud can be devastating for victims. The research team conducted a survey experiment and found that a brief online educational intervention was effective in reducing susceptibility to fraud for at least three months. The educational intervention did not reduce willingness to invest, but rather increased participants' knowledge. The beneficial effects were centered on the financially savvy. Findings suggest that a brief financial education intervention is effective in reducing susceptibility to financial fraud.

In summary, experiment on FFI modeling for listed companies based on time-series information has made some progress. However, there are still some challenges and problems, such as the reliability of the data and the accuracy of the model. Therefore, future research needs to be further explored and improved to enhance the effectiveness and accuracy of FFI.

4 Model Construction of FFI for Listed Companies Based on Time Series Information

In the listed company's finance, there are often financial fraud, and the basis for FFI is the selection of abnormal financial indicators. Indicator selection of good and bad degree can determine the effectiveness of the FFI model, so the conditions for the selection of indicators should be satisfied with the fraud problem itself related indicators, indicators that can sensitively reflect fraudulent behavior, indicators that can comprehensively reflect the various types of fraud, indicators that are operable, as well as the frequency of the relevant indicators.

4.1 Selection of Time-Series Information Indicators

Financial statements are a structured description of a company's financial position, operating performance and sustainability, and are the main reference document for decision-making by investors, shareholders, creditors, employees and other stakeholders. Currently, the truthfulness of financial statements relies mainly on the ethical standards of managers, robust audits of financial statements, and audit reports and opinions issued by auditors, [10], [11], [12]. However, most financial statement frauds are committed with the awareness or consent of management. As the speed growth of computer technology, various fields have entered the era of big data and artificial intelligence, and machine learning is broadly used because it can process large amounts of data quickly and efficiently. Building a FFI model based on machine learning algorithms can improve the shortcomings of the traditional financial statement fraud in the principle of indicator selection, and the identification method is overly dependent on human labor. Therefore, the study uses machine learning methods to construct a FFI model based on temporal information, and the specific flow of the model is shown in Figure 1.

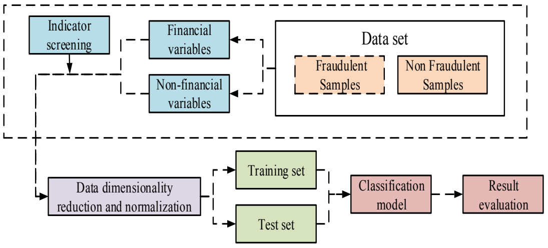


Fig. 1: FFI model construction based on time series information

The study has selected 26 indicators among the financial and non-financial aspects of listed companies, which include many aspects such as profitability, operating capacity, solvency, growth capacity, cash flow capacity, organizational and management capacity. The first-level indicators are expressed through X and the second-level indicators are expressed through numbers. Among them, the indicators selected for the study are shown in Table 1.

Table 1. Selection of financial and non-financial indicators of listed companies

Primary indicators	Code	Secondary indicators	Primary indicators	Code	Secondary indicators
Solvency	X 1	Current ratio	Business capability	X 14	Inventory turnover
	X 2	Quick ratio		X 15	Accounts payable turnover rate
	X 3	Long-term loans to total assets ratio		X 16	Accounts receivable turnover rate
	X 4	Long term debt to capital ratio		X 17	Accounts receivable to income ratio
Development capability	X 5	Total asset growth rate		X 18	Total asset turnover rate
	X 6	Fixed asset growth rate		X 19	Inventory to revenue ratio
	X 7	Growth rate of total operating revenue	Profitability	X 20	Equity Turnover
	X 8	Growth rate of total operating costs		X 21	Return on assets
	X 9	Growth rate of net assets per share		X 22	Return on capital
Governance	X 10	Proportion of independent directors		X 23	Operating gross margin
	X 11	Shareholding ratio of the board of directors		X 24	Net profit margin of total assets
	X 12	Shareholding ratio of the supervisory board		X 25	Sales expense rate
	X 13	Shareholding ratio of the top ten shareholders		X 26	Long term return on capital

When it comes specifically to the identification of financial fraud, the indicators should be able to reflect time-series anomalies in the derived variables and must be "time-series specific". The study could construct incremental forms of time-series indicators to reflect abnormal year-to-year movements in financial data to identify fraud. Taking this idea further, it may be possible to construct absolute, ratio, relative, and other forms of time-series indicators to identify fraud. Time series indicators in the form of differences are available at $X_i - X_i^{'}$; Time series indicators in the form of ratios are available at $X_i / X_i^{'}$; And time series indicators in the form of relative values are available at $(X_i - X_i^{'})/X_i^{'}$ and $(X_i - X_i^{'})/\left|(X_i + X_i^{'})/2\right|$.

4.2 Locally Linear Embedding

Locally Linear Embedding (LLE) is an algorithm capable of obtaining a low-dimensional representation from a locally linear manifold of any dimension, [13]. This algorithm has two uncertain parameters: one is the diameter of the identified neighbors and the other is the reduced dimension. The LLE method, on the other hand, mainly utilizes the method of solving for low dimensions, i.e., solving for the eigenvalues of sparse matrices, to obtain an optimized result, [14]. This method does not require several iterative operations and has a low complexity compared to other learning algorithms. The LLE method is used to find the adjacency between the sampling points and then using the LLE method with restriction, the minimum neighboring right between the sampling points is found and then the minimum neighboring right between the sampling points is found. The LLE algorithm is shown in Figure 2, [15].

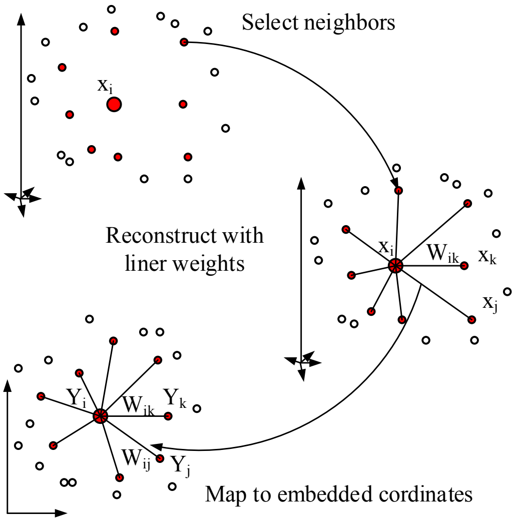


Fig. 2: Specific steps of the LLE algorithm

In Figure 2, the data are represented by the vector $\vec{X}_i$ and the dimension of the vector is set to D. In addition, there are two constraints in the solution process of this method, the first one is: that when the hollow circle is not a point adjacent to the solid circle, it has a weight value of 0; the second constraint is that the weighted sum of the elements of each line in the weighting matrix is equal to 1. When the data supply is the case of no missing, the study expects that each data point and its neighboring points fall in or close to the locally linear blob region, and then it describes the local geometrical structure of those blobs by building the linear coefficients for that data point and its neighboring points to characterize the local geometric structure of these pockets. The number of rows with reconstruction errors is measured by the cost function, which is formulated as shown in formula (1), [16], [17], [18].

$$\varepsilon(W) = \sum_i \left| \vec{X}_i - \sum\nolimits_j W_{ij} \vec{X}_j \right|^2 \tag{1}$$

In formula (1), ε denotes the cost function; W denotes the data point path weights. If the corresponding indices are to be weighted, the cost function must be minimized by converting formula (1) into a constrained least squares optimization problem and expressing it in formula (2).

$$\varepsilon(W) = \min \sum_i \left| \vec{X}_i - \sum\nolimits_j W_{ij} \vec{X}_j \right|^2 \tag{2}$$

After obtaining formula (2) the low dimensional embedding needs to be computed. The observations of the high dimensional vectors will be mapped to the low dimensional vectors and denoted as $\vec{Y}_i$. Subsequently the d dimensional vector $\vec{Y}_i$ is obtained by minimizing the embedding cost function and its expression is shown in formula (3).

$$\phi(Y) = \sum_i \left| \vec{Y}_i - \sum\nolimits_j W_{ij} \vec{Y}_j \right|^2 \tag{3}$$

In formula (3), $\phi(Y)$ denotes the low-dimensional data features. The sparse matrix is used to solve the feature vectors to get their minima, and finally d non-zero feature vectors are obtained, which is the result of feature extraction in low-dimensional space.

4.3 SVM Classification based on K-means Optimization

Support vector machines are based on the Vapnik-Chervonenkis (VC) dimension theory of statistical learning theory and the structural risk minimization principle. Support vector machines utilize a small number of samples for training and use them to forecast problems arising in real applications. The VC dimension is an important measure of the complexity of a functional problem, and the larger the dimension, the more complex the problem, [19]. By improving the support vector machine, the solution of the support vector machine is made independent of the data dimensionality, [20].

The way it is categorized is shown in Figure 3.

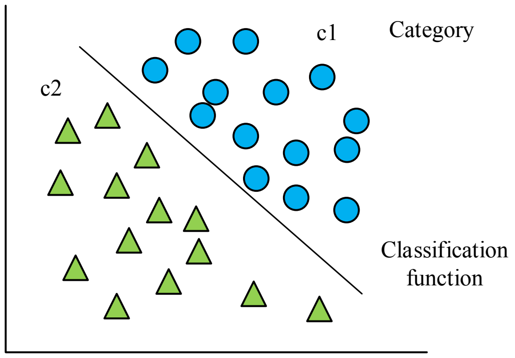


Fig. 3: SVM classifier classification

In Figure 3, $c1$ and $c2$ are in a two-dimensional plane, mainly to identify the class, and a line in the middle is a classification function that completely separates the two types of samples, which is a linear classification function called "hyperplane". If a linear function can accurately separate the samples, then the samples are linearly separated, and vice versa, the samples are nonlinearly separated. However, not every sample is linearly separable, so the data needs to be transformed to a higher dimension so that they are also better linearly separable. The core of the conversion is to find a way to map between a low dimensional vector and a higher dimensional vector, and that way is to give it a kernel function which is also the core of the Support Vector Machine. The SVM classification process starts with the need to give the training set (x_i, y_i), and the optimization problem for the SVM is expressed through formula (4).

$$\begin{cases} \min \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{1}\xi_i \\ s.t. y_i(\omega^T z_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{4}$$

In formula (4), $z_i = \phi(x_i)$ denotes the mapping function, which maps the training data x_i in high dimensions; ξ_i denotes the slack variable; C denotes the penalty factor for error loss, which takes a fixed value rather than a variable; ω denotes the normal vector; z_i denotes the sample eigenvector; and b denotes the model bias term. The study transforms the optimization problem of formula (4) into a dyadic problem to be solved as shown in formula (5).

$$\begin{cases} \min F(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T \\ y^T\alpha \end{cases} \tag{5}$$

In formula (5), α denotes the Lagrange multiplier; e denotes all vectors; Q denotes the positive matrix; and $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function used in the study. The final decision function obtained is shown in formula (6).

$$\text{sgn}(\omega^T\phi(x) + b) = \text{sgn}\left\{\sum_{i=1}^{1}\alpha_i y_i K(x_i, x) + b\right\} \tag{6}$$

In classifying the data, the study validates the estimation of classifier accuracy by using k-fold cross-validation (K-CV). First, the study divides these samples into K samples, tests each sample, and then tests the remaining K-1 samples to obtain K samples. Under K-CV, we use the average recognition accuracy of the test set of K samples to measure the classification accuracy under K-CV, where the value of K is usually 3. The K-CV algorithm can effectively prevent the phenomenon of "over-learning" or "under-learning", and the resulting conclusions have better confidence. K-CV algorithm can effectively prevent the phenomenon of "over-learning" or "under-learning", and the conclusions obtained are more reliable. It is assumed that there is a sample set $V = \{v_1, v_2, ..., v_n\}$, and the algorithm is trained in the sample set to get a prediction function, in which the prediction can be represented by $\hat{y}_i$. K-fold cross-validation is to divide the sample data into almost equal and disjoint K-folds, and each of the K-folds is taken as the verification set, and the remaining K-1 folds are taken as the training set. The prediction results of the data in the sample set can be expressed by formula (7).

$$\hat{y}_{ij} = \frac{1}{t}\sum_{v=1}^{t} pred_v^{(i)}(x_j) \tag{7}$$

In formula (7), $pred_v^{(i)}(x_j)$ represents the predicted value of the sample point, and $\hat{y}_i$ represents the average predicted value of the sample point. In document classification, this paper proposes a classification algorithm based on K-means. Using the k-mean method, multiple targets are divided into multiple clusters, and the resulting clusters have high cluster similarity and small cluster features. Cluster similarity is a measure of the average distance between objects in a cluster, which can be used as both the center of mass and the center of the cluster. The method is used in the clustering process by randomly selecting k targets that represent the mean or center of the clusters, respectively. For each remaining cluster, the cluster is assigned to the closest cluster based on the magnitude of the average distance of the other clusters. On this basis, a new average is applied to each cluster. This is usually defined using the squared error criterion, the expression of which is shown in formula (8).

$$\begin{cases} E = \sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \\ m_i = \frac{1}{n}(x_{1i}+\ldots+x_{ni}) \end{cases} \tag{8}$$

In formula (8), E denotes the sum of squared errors of all objects in the dataset; p denotes the points of a given object in space; and m_i denotes the mean of the cluster. formula (8) shows to find the square of the distance of the sample points in each cluster to its cluster center and then sum it.

5 Performance Parameter

Conduct a simple analysis of the parameter indicators used in this study.

5.1 Cross Validation

Divide the dataset into a training set and a testing set in a certain proportion. Train the model on the training set and test it on the testing set to evaluate the model's generalization ability on unknown data. The study adopts K-fold cross validation, which divides the dataset into k subsets. Each time, one subset is used as the test set, and the remaining k-1 subset is used as the training set. Repeat k times, and the final model evaluation result is the average of k results.

5.2 Precision

Accuracy is the proportion of correctly predicted positive instances to predicted positive examples. Its calculation is shown in formula (9).

$$precision = \frac{True\ Positives(TP)}{TP + False\ Positives(FP)} \tag{9}$$

In formula (9), TP represents the number of samples with a positive real category and a positive prediction; FP represents the number of samples with a negative real category and a positive prediction.

5.3 Accuracy

Accuracy is the proportion of the number of correctly predicted samples in the model to the total number of samples, calculated as shown in formula (10).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

In formula (10), TN represents the number of samples with negative real category and negative prediction; FN represents the number of samples with a positive real category and a negative predicted category.

5.4 Recall Value

Recall value is the proportion of correctly predicted positive instances, calculated as shown in formula (11).

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

5.5 F1 Value

The F1 value is the harmonic mean of recall and precision, used to consider both recall and precision, rather than prioritizing a particular indicator. The calculation formula is formula (12).

$$F1 = \frac{2precision \cdot recall}{precision + recall} \tag{12}$$

6 Performance Analysis

In the establishment of financial fraud identification models, cross validation is used to reduce the probability of overfitting and improve the predictive performance of the model in new data. Due to the possibility of data skewness in financial fraud data, the use of cross validation can ensure that both fraudulent and non fraudulent situations are fully considered when training and testing the model. The Precision indicator predicts how much of the financial fraud is true, and its high accuracy means that the model has more credibility when issuing warnings. The Accuracy indicator shows the model's ability to accurately predict. The Recall indicator focuses on the proportion of true financial fraud detected by the model to all financial fraud, and a high recall rate can identify the majority of financial fraud. The F1 value can balance the importance of recall and precision while also improving the performance of the model. In financial fraud detection, a higher F1 value indicates that the model can better identify financial fraud and maintain a low false alarm rate.

7 Performance Analysis of FFI Model for Listed Companies based on Temporal Sequence Information

The study selected the financial data of 250 listed companies to validate the LLE algorithm. In this step, we use the original data without normalized processing for dimensionality reduction, and the results are shown in Figure 4.

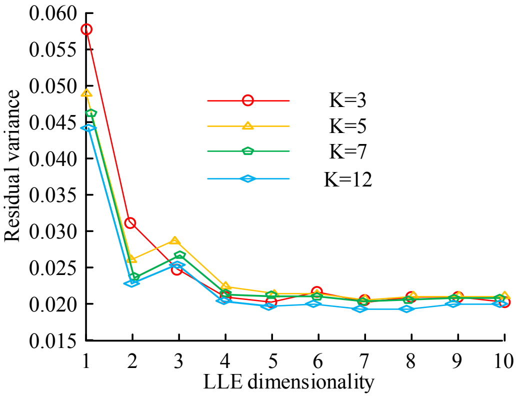
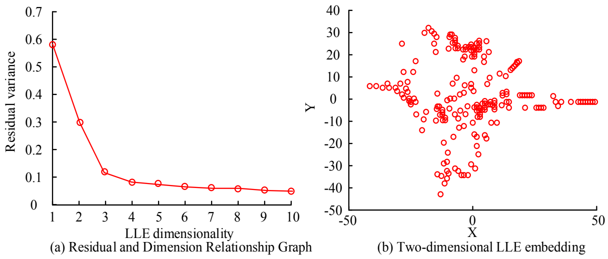


Fig. 4: Unnormalized data downscaling

In Figure 4, the most critical parameter is the number of neighboring points k. The experiment obtains different results by changing the parameter k. When the number of neighbor points takes the value of 3, the curve dimension decreases smoothly, and there is no prominent inflection point. When the value of the number of neighboring points is increased, the curve dimension will have 1 or 2 inflection points, and at this time, it is impossible to judge the true dimension. Experiment with z-score normalized preprocessed data z-score data brought into the LLE algorithm for dimensionality reduction, the results are shown in Figure 5.



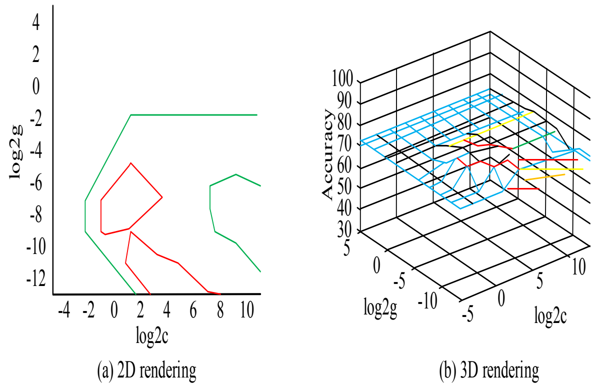
(a) Residual and Dimension Relationship Graph

(b) Two-dimensional LLE embedding

Fig. 5: Dimensionality reduction of normalized data

In Figure 5(a), the curve decreases with increasing dimensionality, indicating that the financial data used for the experiment were successfully downgraded, and that the residual values of the data become smaller as the dimensionality increases, and the minimum value of the residuals is about 0.06. The "intrinsic" dimensionality of the data can be searched for by searching for a turning point that abruptly stops the apparent decline. Before the turning point, there is a large difference between the raw dimension and the true intrinsic dimension; after the turning point, the deviation between the raw dimension and the true intrinsic dimension is very small. From Figure 5(b), it can be seen that the residual curves become flat and the residuals are almost the same at dimensions greater than 3.

The selected 250 listed companies were categorized into two main groups according to their financial status: 155 with good creditworthiness and 95 with credit default risk. The study therefore set a label "1" for companies with good creditworthiness; and a label "-1" for companies with poor creditworthiness. Once the data variables were collected, the study used cross-validation to validate the classification function of the model. The SVM parameters were chosen as shown in Figure 6.



(a) 2D rendering

(b) 3D rendering

Fig. 6: Graph of the results of SVM parameter selection

In Figure 6, the experiment obtains the contour plot of $(\log_2 c, \log_2 g)$ and a 3D view of the prediction accuracy under the combination of RBF and the parameter E by the parameter search method of parameter optimization, which achieves an accuracy of up to 82.254% by 5-fold cross-validation. The experiments visualize the output of SVM classification as a 2D problem and the visualization results are shown in Figure 7.

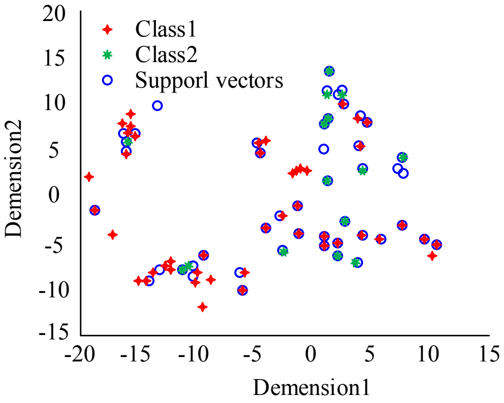
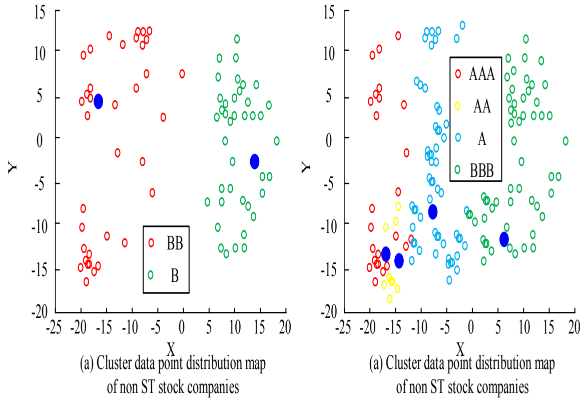


Fig. 7: SVM classification visualization

In Figure 7, the classification hyperplane can only be drawn if the financial data label is +1 or -1. Figure 7 successfully classifies the listed companies in the test dataset into two categories, in which the first category of data is clearly delineated from the second category with less overlap, and the support vectors at the edges of the hyperplane maximize the geometric spacing of the hyperplane. The results show that the classification visualization used in the study has better results. After the financial data have been downgraded and classified, the study verifies the model clustering effect. The study specifically divides the enterprise financial indexes into ten grades AAA, AA, A, A, BBB, BB, B, CCC, CC, C and D. It is not meaningful to cluster companies with continuous losses and poor credit into six clusters, so the study clusters those with the ability to repay loans into four clusters, specifically AAA, AA, A, BBB; those with the risk of default are divided into three clusters, respectively, BB, B, and CCC-D. In the face of the "ST" category the company needs to be a loss for several years, it can not be considered, and non-ST enterprises can not be considered. In the face of the "ST" category of enterprises that need to lose money in consecutive years can not be considered, the non-ST enterprises are clustered and divided into BB and B poles. The specific classification results are shown in Figure 8.



(a) Cluster data point distribution map of non ST stock companies

(a) Cluster data point distribution map of non ST stock companies

Fig. 8: Distribution of data points for the clustering K-means algorithm

Figure 8(a) shows the data point distribution map of A-share listed companies that are not losing money for two consecutive years but have the risk of repaying loans clustered into 2 clusters, and its classification effect is clear and obvious without obvious overlapping distribution. Figure 8(b) shows the data distribution map of good credit, when the distribution clusters are more similar, the data distribution points will be overlapped and crossed, but the distribution clusters are far away from each other, which has a good classification effect.

The study was conducted to further validate the model performance through a test set, while logistic regression model, random forest model, and XGBoost model were used for comparative analysis. The accuracy results of the models in fraud identification in corporate finance are shown in Figure 9.

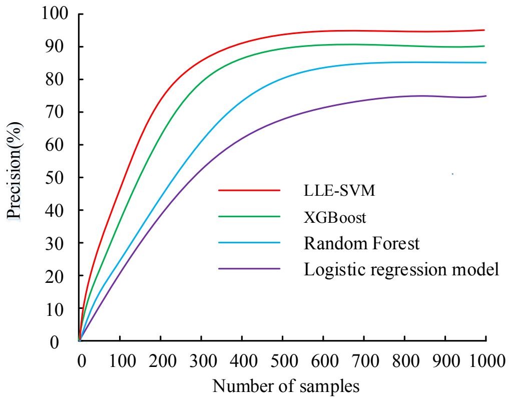


Fig. 9: Fraud identification accuracy results for each model

In Figure 9, the accuracy of the logistic regression model increases with the number of samples and stabilizes at 75.92% when the curve converges; the accuracy of the random forest has a positive correlation with the number of samples and stabilizes at 85.68% when the curve converges; the accuracy of the XGBoost model increases with the number of samples and stabilizes at 90.19% when the curve converges; and the research The accuracy of the proposed model has a positive correlation with the number of samples and is stable at 94.89% when the curve converges. The accuracy results of each model are shown in Figure 10.

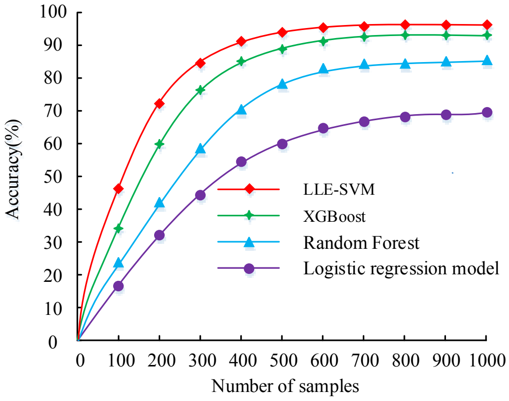


Fig. 10: Fraud identification accuracy results for each model

In Figure 10, the accuracy of the logistic regression model increases with the number of samples, and when the curve converges, the accuracy is stable at 69.66%; the accuracy of the random forest has a positive correlation with the amount of samples, and when the curve converges, the accuracy is stable at 86.82%; the accuracy of the XGBoost model increases with the number of samples, and when the curve converges, the accuracy is stable at 93.57%; and the research The precision of the proposed model has a positive correlation with the amount of samples and is stable at 96.24% when the curve converges. The recall and F1 values of each model are expressed in Table 2.

Table 2. Fraud identification recall value and F1 value results for each model

Model	Logistic regression model	Random Forest	XGBoost model	Research model
Recall value	70.05%	67.98%	82.16%	88.49%
F1 value	70.25%	73.48%	84.48%	87.08%

In Table 2, the recall value of the proposed model is 88.49%, which is 18.44%, 20.51%, and 6.33% higher compared to logistic regression, random forest, and XGBoost models, respectively; and the F1 value of the proposed model is 87.08%, which is 16.83% higher compared to logistic regression, random forest, and XGBoost models, respectively, 13.60%, and 2.6%, respectively. The experimental results validate the feasibility of the proposed model for FFI. From the above results, it was found that the proposed method outperformed other comparative algorithms in terms of accuracy, precision, recall, and F1. The possible reason for this may be that when processing financial data, research methods reduce multidimensional financial indicators to low dimensional characteristics, which can retain key information while removing noise and irrelevant information, thereby improving the recognition ability of the model. During the model training process, as the number of samples increases, the recognition ability of the model also increases. This may be because more samples provide richer information, enabling the model to better learn the distribution of data. Throughout the entire research process, there may be multiple model adjustments and optimizations, including selecting the optimal parameter settings and designing more effective loss functions.

8 Discussion

The study conducted standardized preprocessing on the original financial data, which is a prerequisite for the excellent performance of the model. Normalizing preprocessed data with z-score reduces the data differences between different dimensions, enabling all features of the data to have an equal impact on the model, avoiding the model's tendency to rely on a specific dimension due to its large data range. Secondly, the use of local linear embedding (LLE) technology can effectively reduce high-dimensional raw data to low dimensions, thereby reducing data complexity while extracting the "intrinsic" dimensions of the data, better preserving key information of the data. The study also uses SVM classifiers to construct decision boundaries in high-dimensional space, and can maximize the geometric spacing of hyperplanes by finding support vectors, greatly improving the classification accuracy of the model. In addition, the K-means clustering algorithm has also provided good validation for identifying financial risks, and its classification ability has been further improved through clear segmentation of the boundary. According to the experimental results, as the number of samples increases, the recognition performance of the model also improves. This may be because a large amount of sample data provides the model with richer information, enabling it to better capture complex patterns of data. Compared with other commonly used models, the proposed method exhibits advantages in evaluation indicators such as accuracy, precision, recall, and F1 value, which may be due to its strong robustness when dealing with complex financial data.

9 Conclusion

This paper proposes a financial fraud identification model based on temporal information, which focuses on the use of temporal indicators to detect financial fraud behavior of listed companies. The model first constructs a multi-dimensional index containing timing information, and then uses a complex algorithm called local linear embedding (LLE) to reduce the dimensionality of these data indicators, so that the high-latitude information is effectively retained in the low-latitude features. On this basis, the model further classifies and clusters the data indexes after dimensionality reduction. The study verifies the proposed model through simulation experiments. In data dimensionality reduction, the model successfully reduces the dimensionality of financial data, and the error between the intrinsic dimensions obtained from dimensionality reduction and the original dimensionality is very small; in clustering experiments, the model proposed by the study has a clear and obvious clustering effect; in the practical application and comparison experiments, the accuracy of the model developed by the study is 94.89%, the precision of the model is 96.24%, model recall is 88.49%, and model F1 value is 87.08%. The research findings denote that the FFI model grounded on time-series information has a good performance for efficiently processing high-dimensional financial data, extracting the features of financial fraud data, and learning from the data to achieve high-precision classification and clustering. Although the research results are excellent, there are some limitations and room for improvement. For the selection of non-temporal indicators for model construction, in the current study, the main focus is on continuous variables, which means that our model does not cover all types of financial information, especially discrete variables. Considering that discrete variables also play an important role in financial data, future studies should add more types of variables, including discrete variables, to further improve the processing power of the model and reveal the pattern characteristics of financial fraud more comprehensively.

References:

[1] Wei L, Peng M, Wu W, Financial Literacy and Fraud Detection-Evidence from China, *International Review of Economics and Finance*, Vol. 76, 2021, pp. 478-494.

[2] Gupta S, Kanungo RP, Financial Inclusion through Digitalisation: Economic Viability for the Bottom of the Pyramid (BOP) Segment, *Journal of Business Research*, Vol. 148, 2022, pp. 262-276.

[3] Boa M, Zimková E, Overcoming the Loan-To-Deposit Ratio by A Financial Intermediation Measure-A Perspective Instrument of Financial Stability Policy, *Journal of Policy Modeling*, Vol. 43, No. 5,

2021, pp. 1051-1069.

[4] Qiu S, Luo Y, Guo H, Multisource Evidence Theory Ased Fraud Risk Assessment of China's Listed Companies, *Journal of Forecasting*, Vol. 40, No. 8, 2021, pp. 1524-1539.

[5] Lotfi A, Salehi M, Dashtbayaz ML, The Effect of Intellectual Capital on Fraud in Financial Statements, *TQM Journal*, Vol. 34, No. 4, 2021, pp. 651-674.

[6] Gonidakis F, Koutoupis AG, Kyriakogkonas P, Lazos G, Risk Disclosures in Annual Reports: The Role of Non-Financial Companies Listed in Athens Stock Exchange, *Journal of Operational Risk*, Vol. 16, No. 3, 2021, pp. 19-45.

[7] Cao G, Zhang J, Guanxi, Overconfidence and corporate Fraud in China, *Chinese Management Studies*, Vol. 15, No. 3, 2021, pp. 501-556.

[8] Suksonghong K, Amran A, Achieving Earnings Target through Real Activities Manipulation: Lesson from Stock Exchange of Thailand, *International Journal of Monetary Economics and Finance*, Vol. 13, No. 3, 2020, pp. 260-268.

[9] Burke J, Kieffer C, Mottola G, Perez-Arce F, Can Educational Interventions Reduce Susceptibility to Financial Fraud? *Journal of Economic Behavior and Organization*, Vol. 198, 2022, pp. 250-266.

[10] Younus M, The Rising Trend of Fraud and forgery in Pakistan's Banking Industry and Precautions Taken Against, *Qualitative Research in Financial Markets*, , Vol. 13, No. 2, 2021 pp. 215-225.

[11] Nguyen AL, Mosqueda L, Nikki Windisch M, Weissberger G, Axelrod J, Han SD, Perceived Types, Causes, and Consequences of Financial Exploitation: Narratives from Older Adults, *The Journals of Gerontology: Series B*, Vol. 76, No. 5, 2021, pp. 996-1004.

[12] Cao Y, Modeling the Dependence Structure and Systemic Risk of All Listed Insurance Companies in the Chinese Insurance Market, *Risk Management and Insurance Review*, Vol. 24, No. 4, 2021, pp. 367-399.

[13] Wu S, Pang Z, Gao Y, Chen G, Xiang S, Zhao C, NEIST: A Neural-Enhanced Index for Spatio-Temporal Queries, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 4, 2021, pp. 1659-1673.

[14] De Oliveira KL, Ramos RL, Oliveira SC, Christofaro C, Water Quality Index and spatio-Temporal Perspective of A Large Brazilian Water Reservoir, *Water Science and Technology: Water supply*, Vol. 21, No. 3/4, 2021, pp. 971-982.

[15] Jones S, Luo S, Dorton HM, Angelo B, Page KA, Evidence of A Role for the Hippocampus in Food-Cue Processing and the Association with Body Weight and Dietary Added Sugar, *Obesity*, Vol. 29, No. 2, 2021, pp. 370-378.

[16] Cantor AG, Jungbauer RM, Mcdonagh M, Counseling and Behavioral Interventions for Healthy Weight and Weight Gain in Pregnancy: Evidence Report and Systematic Review for the US Preventive Services Task Force, *JAMA the Journal of the American Medical Association*, Vol. 325, No. 20, 2021, pp. 2094-2109.

[17] Jain SK, Patidar NP, Kumar Y, Real-Time Voltage Security Assessment Using Adaptive Fuzzified Decision Tree Algorithm, *International Journal of Engineering Systems Modelling and Simulation*, Vol. 13, No. 1, 2022, pp. 85-95.

[18] Guo Y, Mustafaoglu Z, Koundal D, Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms, *Journal of Computational and Cognitive Engineering*, Vol. 2, No. 1, 2022, pp. 5-9.

[19] Daho MEH, Settouti N, Bechar MEA, A New Correlation-Based Approach for Ensemble Selection in Random Forests, *International Journal of Intelligent Computing and Cybernetics*, Vol. 14, No. 2, 2021, pp. 251-268.

[20] An X, Hu C, Li Z, Lin H, Liu G, Decentralized AdaBoost Algorithm over Sensor Networks, *Neurocomputing*, Vol. 479, No. 28, 2022, pp. 37-46.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

From the proposal of the problem to the final discovery and resolution, it was solely contributed by Lili Wang.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.