Financial Risk Warning in Listed Manufacturing Enterprises Using a Huffman Tree Enhanced Support Vector Machine with Arithmetic Optimization

Li Zhao
Lu'an Vocational Technical College, Lu'an 237000, China
Email: zl619003518@163.com

Keywords: HT-SVM, manufacturing enterprises, financial risk, nonlinear mapping, warning testing

Received: June 6, 2025

During the production and operation process, manufacturing enterprises may experience financial instability due to factors such as capital flows, cost control, and market changes, which can affect their profitability and debt-paying ability. Although certain progress has been made in financial risk early warning, there are still obvious problems such as lagging early warning and incomplete indicator systems. To further optimize the early warning mechanism of enterprise financial risks and improve the response efficiency, an improved Huffman tree support vector machine algorithm is proposed. This algorithm combines arithmetic optimization algorithms and is applied to the early warning and control of financial risks in listed manufacturing enterprises. This method converts the low-dimensional space into a high-dimensional space through nonlinear mapping, thereby enhancing the computing speed and prediction accuracy. The study adopts five publicly available multi-class imbalanced datasets. The experimental results showed that the accuracy rates of the improved Huffman tree support vector machine algorithm on the training set were 80.3649%, 89.6989%, 90.3654%, 96.2453%, and 97.4658% respectively. The accuracy rates on the test set were 85.3694%, 91.3658%, 92.3654%, 94.2652%, and 96.7659% respectively. The prediction accuracy of the overall model reached 81.8%, which was higher than that of traditional methods. The results show that the optimization algorithm combining Huffman tree mechanism and support vector machine can effectively meet the needs of financial risk early warning in manufacturing enterprises, providing theoretical support and practical basis for subsequent financial risk diagnosis and control applications.

Povzetek: Predlagani HT-SVM s Huffmanovim drevesom in aritmetično optimizacijo ob nelinearnem preslikanju odpravlja neravnovesje razredov (IR ≈1), zmanjša število klasifikatorjev ter pospeši učenje. Na javnih naborih doseže dobre rezultate; v proizvodnih podjetjih izboljša pravočasno opozarjanje na finančna tveganja.

1 Introduction

Affected by the complex economic environment, listed manufacturing enterprises face multiple financial risks such as materials, markets, and supply chains. Once financial risks occur, it can affect the profitability and market reputation of the enterprise, and may also lead to the breakage of the funding chain or even bankruptcy. In this context, Shu et al. proposed a novel multi-signal integration method for anomaly detection in the financial market. The experimental results showed that the detection accuracy was improved by 15.4% and the average detection lead time was increased by 2.8 days [1]. Du and An proposed a method based on differential evolution algorithm to measure enterprise financial credit risk, achieving a time cost control within 0.4 seconds and an error rate of no more than 1% [2]. Chen et al. proposed an enterprise financial data risk prediction model based on entropy weight method for inaccurate financial risk prediction caused by improper risk indicators. The experimental results showed that the prediction accuracy rate always remained at about 98% [3]. Cao et al. proposed a combined model based on time series analysis and support vector machine to address low prediction accuracy and long prediction time of traditional methods. This model achieved a prediction accuracy rate of 95% to 100% and a prediction time of no more than 16 seconds in predicting financial data leakage [4]. Zhang et al. proposed a financial risk monitoring and warning method based on data mining to address the low accuracy of traditional methods. The accuracy of data mining reached 98.23%, and the accuracy of risk warning exceeded 95% [5].

Huffman tree encodes categories to make SVM more adaptable to imbalanced distributions between categories and reduce classification computational complexity. Wang et al. proposed a method based on Long Short-Term Memory Network (LSTM) for the stock market volatility, which effectively improved the stock price prediction accuracy [6]. Dessaint et al. proposed a theoretical model to address the impact of short-term data on the accuracy of long-term predictions, proving that short-term data

could cause predictors to shift their focus to the short-term, thereby reducing the effectiveness of long-term predictions [7]. Okeke et al. proposed a comprehensive analysis method for strategic budgeting and revenue management aimed at addressing financial stability issues in small and medium-sized enterprises, which helped improve financial forecasting and long-term sustainability [8]. In response to the low timeliness and inaccuracy of the budget in the finance department, Lv proposed a method for fiscal and tax data management and budget prediction based on a time series model, achieving good results with average errors of 7.3%, 7.4%, and 12.1% from 2020 to 2022 [9]. Jiao proposed a method that combined the minimum absolute contraction and selection operator with the gradient boosting tree algorithm in response to the concept drift problem in predicting financial difficulties of enterprises during economic recession cycles. The method achieved an accuracy of 92.47% in a dynamic environment [10]. The specific summary of the above-mentioned work is shown in Table 1.

Table 1: Literature summary table

Literature citation	Research method	Advantages	Disadvantage
Shu et al. [1]	Multi-signal integration method	The detection accuracy has been improved and the detection lead time has been increased	Computational complexity depends on signal quality
Du and An [2]	Differential evolution algorithm	Reduce time cost (<0.4 seconds) and have a low error rate (<1%)	The large demand for data may lead to a decline in the applicability of the model
Chen et al. [3]	Entropy weight method and financial data risk prediction model	The prediction accuracy rate is as high as 98%	Relying on the accuracy and applicability of risk indicators without considering data imbalance
Cao et al. [4]	Data mining technology	The accuracy rate of monitoring and early warning is high (>95%)	It is highly complex and requires a relatively long time for data preprocessing
Zhang et al. [5]	A combined model of time series analysis and support vector machine	High prediction accuracy (95%-100%), and fast response (<16 seconds)	Noise processing and data cleaning may introduce biases
Wang et al. [6]	The method based on LSTM	Effectively improve the accuracy of stock price prediction	The limitations of this model, such as the dependence on input data quality, are not explicitly mentioned
Dessaint et al. [7]	Theoretical model	It demonstrates the impact of short-term data on the effectiveness of long-term predictions and strengthens the theoretical foundation	It is mainly theoretical analysis and lacks empirical data support
Okeke et al. [8]	A comprehensive analytical approach to strategic budgeting and revenue management	It helps improve financial prediction and the long-term sustainability of small and medium-sized enterprises	The specific implementation and practical application details are relatively few, which may affect the universality
Lv [9]	Financial and tax data management and budget prediction based on time series models	Budget predictions with relatively low average errors are achieved from 2020 to 2022	Only time period data is provided, without discussing other possible influencing factors
Jiao [10]	The minimum absolute shrinkage and selection operator are combined with the gradient boosting tree algorithm	It achieves a high accuracy rate of 92.47% in a dynamic environment, adapting to the concept drift	It may be necessary to perform more complex parameter tuning, and the applicability may be limited by application scenarios

In current work, although some research has achieved certain results in financial risk early warning using methods such as deep learning and data mining, there are still some obvious limitations. For instance, some research relies on static models, which are vulnerable to data imbalance and noise, resulting in low warning accuracy. Although other research has increased complexity, they have not effectively improved classification performance when dealing with imbalanced samples of multiple classes. In addition, many traditional methods also have deficiencies in prediction speed and real-time performance. In contrast, the improved SVM

based on Huffman Tree (HT-SVM) algorithm proposed in the research enhances the classification accuracy and generalization ability of the model by combining Arithmetic Optimization Algorithm (AOA) and nonlinear mapping. It has made up for the shortcomings of previous methods in dealing with the complexity and imbalance of financial data, demonstrating higher response efficiency and practical application potential. This innovative improvement provides more effective tools and theoretical support for the early warning and control of financial risks in manufacturing enterprises.

2 Methods and materials

2.1 Function construction of improved HT-SVM algorithm

SVM, a supervised learning model, is commonly used for text classification, image recognition, and financial prediction in both classification and regression tasks. Its primary goal is to find an optimal hyperplane that maximizes the separation between different data categories [11, 12]. SVM excels in high-dimensional spaces, relying on support vectors for decision-making, providing high storage efficiency, and effectively handling nonlinear problems [13, 14]. It employs nonlinear mapping to transform non-separable samples from a low-dimensional space into a higher dimensional feature space, enhancing separability and improving data classification accuracy. In binary classification tasks, the training sample set is presented in equation (1).

$$Z=\{(a_1,b_1),(a_2,b_2),\cdots,(a_n,b_n),b_i\in\{-1,+1\}\} \quad (1)$$

In equation (1), $a_i \in z$ indicates that a_i belongs to an Z-dimensional space vector. $b_i \in$ {-1, +1} represents the class label of the sample. The classification is to build an optimal discriminative model from trained data that effectively differentiates between sample types and applies this model to new data for accurate classification predictions. This differentiation is represented by a hyperplane, described by the corresponding equation, as shown in equation (2).

$$k^L m+Z=0 \quad (2)$$

In equation (2), k refers to the weight of the variable. Z refers to the threshold. In high-dimensional space, the hyperplane position is determined by parameters k and z, and its geometric properties can be maintained unchanged through equal scaling (equivalent scaling). The plane scaling is shown in equation (3).

$$\left.\begin{matrix}k^L m_i+z\geq+1,b_i=+1\\k^L m_i+z<+1,b_i=-1\end{matrix}\right\} \quad (3)$$

In equation (3), while multiple feasible hyperplanes can classify positive and negative samples correctly, it is challenging to identify the one with the best generalization performance due to the non-uniqueness of the solution. Some solutions may overfit training data, leading to poor performance on new data. The SVM algorithm for linearly separable datasets aims to construct a separation hyperplane that maximizes the classification margin. Numerous hyperplanes can exist during this process. The optimal segmentation hyperplane is illustrated in Figure 1.

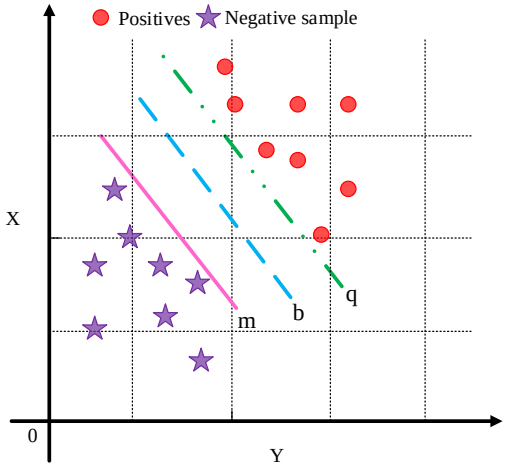


Figure 1: The optimal segmentation hyperplane graph

In Figure 1, red dots represent positive samples, and purple pentagrams denote negative samples. The hyperplanes m, b, and q can separate the samples, with the optimal one determined by maximizing the minimum distance between support vectors. Hyperplane b offers the largest margin compared to m and q, thus demonstrating the best generalization performance. Therefore, the optimal segmentation hyperplane is b. The process by which SVM transforms complex nonlinear classification problems into linearly separable mapping is shown in Figure 2.

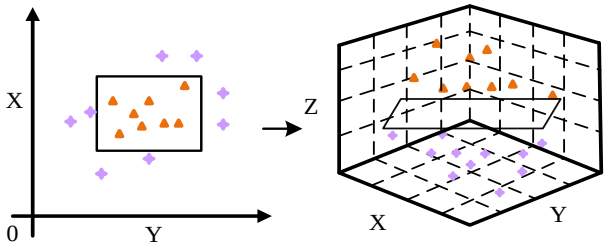


Figure 2: The process of mapping to a high-dimensional spatial graph

In Figure 2, SVM uses a kernel function to map linearly inseparable data from low-dimensional to high-dimensional space, making it separable in the latter, which enhances generalization, prevents overfitting, and improves computational efficiency [15]. AOA, a meta-heuristic optimization algorithm inspired by arithmetic operations, solves continuous optimization problems. It consists of three stages: initialization, global exploration, and local development. The hierarchical structure of each arithmetic operator is shown in Figure 3.

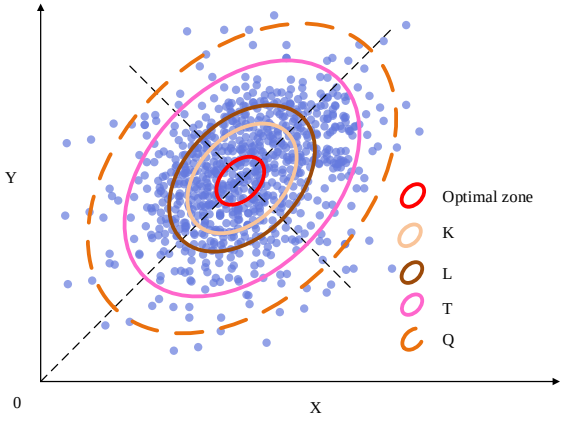


Figure 3: Hierarchical structure diagram of arithmetic operators

In Figure 3, QT denotes global exploration, while KL represents local development. The AOA optimization process starts with a randomly initialized candidate solution set. During iterations, the algorithm evaluates solution quality, designating the best candidate as the current optimal solution, which can be viewed as the global optimal solution or its high-precision approximation under convergence. Before each iteration update, AOA calculates the current search coefficient using the Math Optimizer Accelerated (MOA) function, determining whether to engage in global exploration or local development based on the coefficient threshold. The MOA function is presented in equation (4).

$$MOA(k_b) = Min + k_b * (\frac{Max - Min}{k_{\max}}) \quad (4)$$

In equation (4), Min signifies the minimum value of MOA. Max signifies the maximum value of MOA. $MOA(k_b)$ signifies the current iteration. k_b represents the current iteration count. $k_{\max}$ represents the maximum number of iterations. According to the value of $MOA(k_b)$, the search method is shown in equation (5).

$$\mathrm{SP}=\begin{Bmatrix} \mathrm{Ge}, q_1 > MOA(k_b) \\ \mathrm{Ls}, q_1 \le MOA(k_b) \end{Bmatrix} \quad (5)$$

In equation (5), SP represents the search stage. Ge represents the global exploration. Ls represents the local search. q_1 is a random number. HT-SVM solves multi-classification problems by constructing a binary tree architecture, deploying binary SVM classifiers at each decision node in the tree structure. For datasets containing n classes, this architecture only constructs n-1 binary SVMs to complete the classification task. The construction process of HT-SVM is as follows, which includes n class datasets, as shown in equation (6).

$$M(b) = \{z_1, z_2, z_3, \cdots, z_n\} \quad (6)$$

In equation (6), z_x represents the number of the x-th class. $M(b)$ is sorted in ascending order based on the size of z_x to form a new set, as shown in equation (7).

$$M(b') = \{z_1', z_2', z_3', \cdots, z_n'\} (z_1' \le z_2' \le z_3' \le \cdots \le z_n') \quad (7)$$

In equation (7), the z_1' element in $M(b')$ is used as the child node on the left side of the equation, and the z_2' element is used as the child node on the right side. The two elements are used as left and right subtrees to construct and return a new binary tree node, as shown in equation (8).

$$z_1' + z_2' = z_{12}' \quad (8)$$

In equation (8), the sum of the left and right child nodes of the element is the value of the new node. The selected elements z_1' and z_2' in $M(b')$ are removed, and a new node z_{12}' is added to $M(b')$. The added expression is shown in equation (9).

$$M(b') = \{z_1', z_2', z_3', \cdots, z_n'\} \quad (9)$$

In equation (9), equations (6), (7), and (8) are repeated until only one node element is left in set $M(b')$. The HT-SVM is constructed.

2.2 Solution of financial risk warning model based on improved HT-SVM algorithm

The main motivation for proposing an improved HT-SVM algorithm in this study is due to the significant challenges faced by manufacturing enterprises in financial risk prediction under class imbalance. To deal with this problem, the research aims to address financial risk prediction under class imbalance conditions by developing a classification algorithm that minimizes Imbalance Rate (IR) and improves generalization ability. The study adopts the K-fold cross-validation to objectively evaluate the generalization ability and stability of the SVM through multiple rounds of partitioning and testing [16, 17]. The cross-validation process is shown in Figure 4.

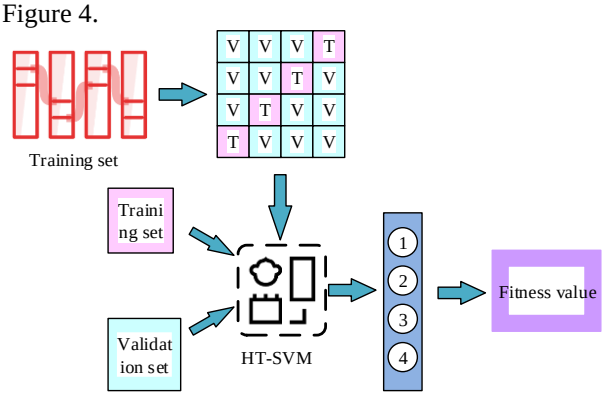


Figure 4: Cross-validation process diagram

In Figure 4, the training set samples undergo four rounds of four-fold cross-validation, where the training set is divided into four equal parts (each representing 25% of the samples). In each round, one-fold is used as the validation set, while the remaining three serve as the training set, allowing the model's performance to be tested. This process, repeated four times, ensures each fold is validated. The class average of the four validation results is then calculated and used as the fitness function value to optimize model parameters or select the best

feature combination. Each binary SVM aims to maintain similar sample sizes for both classes during classification, thereby directly addressing data imbalance without needing resampling or algorithm modification [18-20]. The data imbalance is quantified by the IR, calculated by equation (10).

$$IR = \frac{M_x}{M_y} \tag{10}$$

In equation (10), for a multi-class dataset, IR represents the data imbalance rate. M_y signifies the number of classes with the smallest sample size. M_x signifies the number of categories with the largest sample size. The HT-SVM is constructed by addressing imbalanced data IM in a multi-class dataset, as shown in Figure 5.

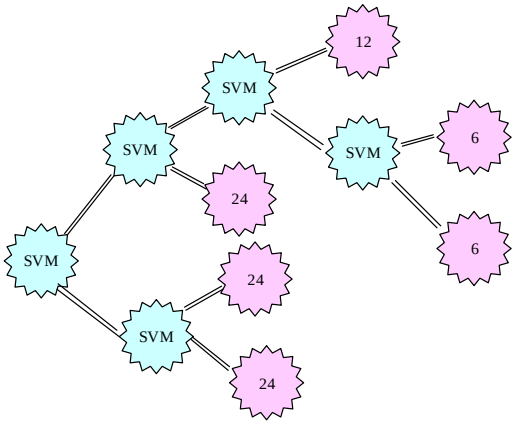


Figure 5: Construction of HT-SVM model diagram

As shown in Figure 5, the IR of the dataset before processing is 6 (24/4), requiring five SVMs to construct HT-SVM. The first SVM is (24, 24, 12, 6, 6), the second SVM is (24, 24), the third SVM is (24, 12, 6, 6), the fourth SVM is (12, 6, 6), and the fifth SVM is (6, 6). Unlike traditional methods, HT-SVM ensures that the IR value of each binary SVM is equal to 1, which means that it completely eliminates the impact of data imbalance. By setting different misclassification cost parameters for the majority and minority classes, the classification performance on imbalanced datasets is improved. The classification cost is shown in equation (11).

$$\min_{\alpha,\beta,\eta} = \frac{1}{2}\|\varphi\|^2 + B_+ \sum_{k\in L_+} \eta_k + B_- \sum_{k\in L_-} \eta_k \tag{11}$$

In equation (11), α and β are decision variables in the optimization process. η_k represents the relaxation variable of each sample k. L_+ represents the index set of the majority class samples. L_- represents the index set of the minority class samples. The majority class introduces a penalty factor B_+, and the minority class introduces a penalty factor B_-. The range of values for the majority and minority classes is shown in equation (12).

$$h.d.y_k(\varphi^V x_k + \zeta) \geq 1 - \eta_k, \eta_k \geq 0, \forall k \tag{12}$$

In equation (12), y_k represents the label of each sample k. φ^V represents the weight vector of the decision boundary. x_k represents the eigenvector of sample k. ζ represents the bias term. The maximum Lagrangian transforms the constrained problem in the equation into an unconstrained problem, as expressed in equation (13).

$$S_q = \frac{1}{2}\|\varphi\|^2 + B_+ \sum_{k\in L_+} \eta_k + B_- \sum_{k\in L_-} \eta_k - \sum_{k=1}^{C} \theta_k [y_k(\varphi^V x_k + \zeta) - 1 + \eta_k] - \sum_{k=1}^{C} \xi_k \eta_k \tag{13}$$

In equation (13), S_q represents the objective function, which minimizes the classification error rate or related classification cost by optimizing weights, penalty factors, and slack variables. θ_k represents the Lagrange multiplier. ξ_k represents the Lagrange multiplier related to η_k. C represents the penalty parameter. The parameters φ, ζ, and η_k are subjected to partial derivative calculation, and their partial derivative expressions are shown in equation (14).

$$\varphi = \sum_{k=1}^{C} \theta_i y_k (\varphi^V x_k + \zeta), \sum_{k=1}^{C} \theta_i y_k = 0 \tag{14}$$

In equation (14), the parameters φ, ζ, and η_k are solved by partial derivatives and the result is set to 0. The final function expression is shown in equation (15).

$$g(x) = sign\left(\sum_{k=1}^{C} y_k \theta_i D(x \cdot x_k) + \zeta \right) \tag{15}$$

In equation (15), $g(x)$ represents the final classification decision function. D represents some kind of inner product or feature mapping function. $sign$ represents the sign function that returns its input to determine whether the classification is positive or negative.

The improved HT-SVM algorithm can effectively solve the insufficient warning accuracy caused by class imbalance and nonlinear characteristics in enterprise financial data. The entire financial risk warning process model is shown in Figure 6.

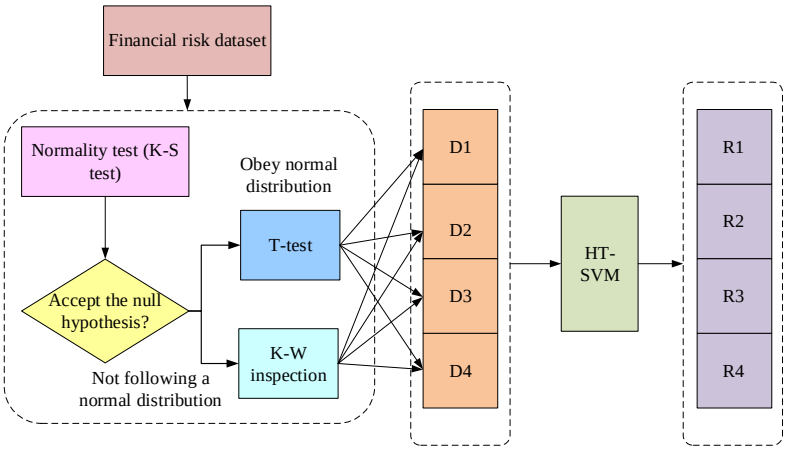


Figure 6: HT-SVM enterprise financial risk warning model diagram

In Figure 6, D1 represents operational capability, D2 denotes development capability, D3 indicates profitability, and D4 reflects debt-paying capability, with R1, R2, R3, and R4 as their respective results. The financial risk warning model process begins with preparing financial reports (income statement and balance sheet), market indicators (stock price change rate), and economic variables (GDP growth rate and interest rate). The Kolmogorov-Smirnov test (K-S test) assesses the data distribution. If normal, a T-test is used for indicator screening. If non-normal, the Kruskal-Wallis test (K-W test) selects indicators to create multidimensional sub-datasets (D1-D4) integrated into the main dataset D. Finally, the core HT-SVM algorithm generates the final result set R and outputs the results. Based on the above content analysis, the pseudo-code of the proposed method in the research is shown in Figure 7.

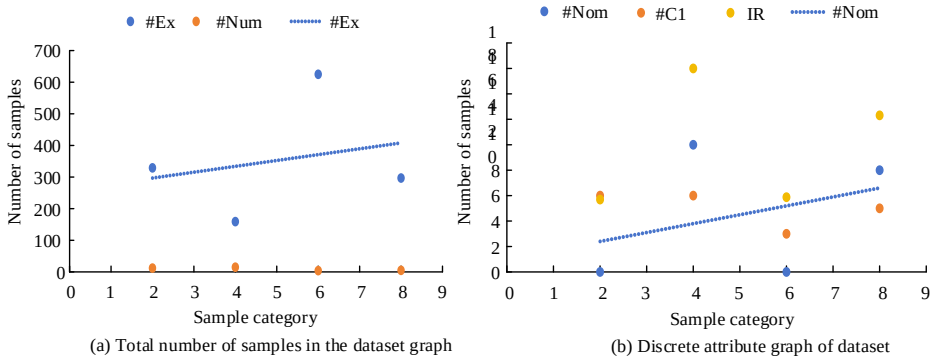
```
Algorithm HT-SVM(X_train, y_train, X_test)

Input:
    - X_train: Training feature dataset
    - y_train: Training labels
    - X_test: Testing feature dataset

Output:
    - predictions: Predicted labels for X_test

1. FUNCTION HT-SVM(X_train, y_train, X_test):

    1.1. Calculate class probabilities from y_train
    1.2. Build Huffman Tree using calculated probabilities
    1.3. Encode y_train using Huffman Tree
    1.4. Train SVM model on X_train with encoded labels
    1.5. Predict encoded labels for X_test using the trained SVM
    1.6. Decode predictions back to original categories using Huffman Tree

    RETURN predictions

2. FUNCTION BUILD_HUFFMAN_TREE(probabilities):
    - Initialize nodes for each class
    - Combine nodes until one tree remains
    RETURN Huffman tree root

3. FUNCTION ENCODE_CATEGORIES(y, tree):
    - Map each label in y to its code in the Huffman Tree
    RETURN encoded labels

4. FUNCTION DECODE_CATEGORIES(encoded_predictions, tree):
    - Map each encoded prediction back to its original label
    RETURN decoded predictions

END
```

Figure 7: Pseudo-code for improved HT-SVM algorithm

3 Results

3.1 Performance testing of improved HT-SVM algorithm

The improved HT-SVM algorithm improves the classification accuracy and generalization ability of financial risk warning models through nonlinear mapping, AOA, and cross-validation methods. To validate the effectiveness and practicality, five publicly available multi-class imbalanced datasets are selected for the study, with data sourced from two publicly available databases. The description of each dataset is shown in Figure 8.



(a) Total number of samples in the dataset graph (b) Discrete attribute graph of dataset

Figure 8: Dataset description diagram

In Figure 8, dataset abbreviations include Acc, Aut, Bal, and Cle. #Ex indicates total samples, #Nom indicates discrete attributes, #C1 signifies classes, #Num represents numerical attributes, and IR shows the imbalance rate. According to Figure 8 (a), the Acc dataset had 329 samples with 12 numerical attributes, the Aut dataset had 159 samples with 15 numerical attributes, the Bal dataset contained 625 samples with 4 numerical attributes, and the Cle dataset included 297 samples with 5 numerical attributes. In Figure 8 (b), the Acc dataset had 6 categories and an IR of 5.69, with no discrete attributes. The Aut dataset had 10 discrete attributes, 6 categories, and an IR of 16.00. The Bal dataset had 3 categories and an IR of 5.88, with no discrete attributes. The Cle dataset contained 8 discrete attributes, 5 categories, and an IR of 12.31. To address multi-class imbalance, the HT-SVM algorithm is compared with GA-SVM, PSO-SVM, and OVO-SVM to identify the optimal method. The number of classifiers and IR results for each method are provided in Table 2.

Table 2: Number of classifiers and IR results for each method dataset

/	Dataset	Acc	Aut	Bal	Cle
Number of classifiers	HT-SVM	6	5	3	3
	GA-SVM	8	7	4	6
	PSO-SVM	11	10	5	4
	OVO-SVM	14	13	4	11
IR	HT-SVM	1.00	1.00	1.00	1.00
	GA-SVM	31.82	16.37	18.14	15.944
	PSO-SVM	26.55	9.46	20.36	9.77
	OVO-SVM	6.13	15.58	8.31	13.65

According to Table 2, HT-SVM required six classifiers for the Acc dataset, five for Aut, and three for both Bal and Cle, outperforming the other algorithms. It achieved an IR of 1.66 for Acc, 3.96 for Aut, 6.02 for Bal, and 2.78 for Cle, which were the best results among the four algorithms. The HT-SVM significantly reduced the number of classifiers while maintaining classification performance and decreasing the IR of most datasets to near equilibrium levels. To confirm the effectiveness of the improved HT-SVM in addressing multi-class imbalance issues, the fitness change curves during the optimization process are analyzed, highlighting convergence characteristics that indicate optimal solution attainment. The fitness change curve is presented in Figure 9.

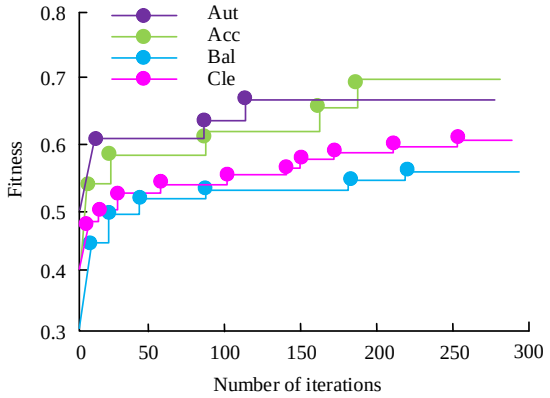


Figure 9: Adaptability change curve chart

Figure 9 demonstrates the curve variations of HT-SVM across four datasets: Acc, Aut, Bal, and Cle. The algorithm exhibited stable convergence across all datasets, significantly achieving convergence within approximately 90 iterations. The excellent convergence characteristics and global optimization ability of HT-SVM highlight its effectiveness in handling multi-class datasets. To evaluate the influence of specific parameter optimization on the performance of the HT-SVM model, ablation experiments are conducted in the study. The experiment mainly focuses on the penalty parameter C and the kernel function. The study compares the changes in model performance without optimizing these parameters to demonstrate the difference in AOA optimization and non-optimization effects. The ablation experiment is specifically shown in Table 3.

Table 3: Results of ablation experiment

Experimental group	C value	Kernel functions	Accuracy (%)	F1
Baseline model	1	Linear function	82.5	0.78
Optimized penalty parameter	10	Linear function	90.2	0.88
Optimized kernel functions	1	RBF function	89.5	0.87
Optimized penalty parameter + kernel function	10	RBF function	91.6	0.90

Table 3 presents the results of the ablation experiment, evaluating the changes in the performance of the HT-SVM model under different parameter configurations. When the penalty parameter C was 1 and the linear kernel function was used, the accuracy of the baseline model was 82.5 and the F1 score was 0.78. After optimizing the penalty parameter C value to 10, the accuracy of the model increased to 90.2, and the F1 score also rose to 0.88. When using the RBF kernel function, even if the C value was 1, the model still performed well, with an accuracy of 89.5 and an F1 score of 0.87. When the penalty parameters and kernel functions were optimized simultaneously, with the C value set to 10 and the RBF kernel used, the model achieved the best performance, with an accuracy of 91.6 and an F1 score of 0.90. These results indicate that parameter optimization has improved the classification performance of the model. The study employs paired t-tests and Wilcoxon signed-rank tests for the research results to evaluate the statistical significance of the performance differences among different models after feature selection. For the paired t-test, the *p* value threshold adopted is 0.05. If the *p* value is less than 0.05, it is considered that there is a significant difference in model performance. The t-test is chosen because the data conforms to a normal distribution, while the Wilcoxon signed-rank test is used in cases where the normality assumption is not satisfied. For feature selection, the K-S and Kruskal-Wallis tests are used to analyze the influence of different features on the target variable, and the *p* value threshold is also set at 0.05 to ensure the reliability of feature selection. The specific results of the statistical test are shown in Table 4.

Table 4: Statistical significance test results

Model	Accuracy (%)	F1	AUC-ROC	Paired t-test *p* value	Wilcoxon *p* value
GA-SVM	82.5	0.78	0.85	0.045	0.038
PSO-SVM	90.2	0.88	0.91	0.006	0.004
OVO-SVM	86	0.84	0.87	0.015	0.013
HT-SVM	91.6	0.90	0.93	0.002	0.001

Table 4 shows the performance indicators of different models and their statistical significance test results. The accuracy of the GA-SVM model was 82.5, the F1 score was 0.78, the AUC-ROC value was 0.85, the *p* value of its paired t-test was 0.045, and the *p* value of the Wilcoxon test was 0.038, indicating that its performance was significantly different from the overall level. The PSO-SVM model performed the best on accuracy, F1 score, and AUC-ROC value, which were 90.2, 0.88 and 0.91, respectively. Moreover, the *p* values of its paired t-test and Wilcoxon test were both lower than 0.01, showing a significant performance improvement. The accuracy of the OVO-SVM model was 86, the F1 score was 0.84, and the AUC-ROC value was 0.87. The statistical test results also showed differences at the 0.05 significance level. The HT-SVM model achieved the highest accuracy of 91.6, an F1 score of 0.90, and an AUC-ROC value of 0.93. Both the paired t-test and the Wilcoxon test showed extremely significant *p* values, emphasizing that this model was significantly superior to other models after feature selection.

3.2 Application effect testing of improved HT-SVM algorithm in financial risk warning

To verify the effectiveness of the improved HT-SVM algorithm in financial risk warning applications, simulation experiments are conducted. The input set of HT-SVM is the classification result, and 150 samples are

selected. The dataset used contains financial data from listed companies in multiple industries. These data mainly cover multiple dimensions such as financial statement data, operating indicators, and market performance. Financial indicators mainly include revenue, net profit, total assets, shareholders' equity, debt ratio, cash flow and price-earnings ratio, etc. The research data mainly comes from financial data providers such as Bloomberg and Reuters. These data are mostly compiled based on corporate annual reports and market transaction data, and have high authority and reliability. Another part of the data comes from open financial databases such as Yahoo Finance and Google Finance. The study adopts multiple proportion configurations to allocate the training and test sets, including 90%:10%, 80%:20%, 70%:30%, 60%:40%, and 50%:50% ratios. Each configuration is implemented through random sampling. To ensure the reliability of model validation, the dataset is first divided into mutually exclusive training and testing sets. Based on the evaluation results of the testing set, the parameters are iteratively optimized. By horizontally comparing the classification accuracy of each candidate model, the optimal warning model is ultimately selected for enterprise financial risk prediction. The financial risk warning test for manufacturing enterprises is shown in Figure 10.

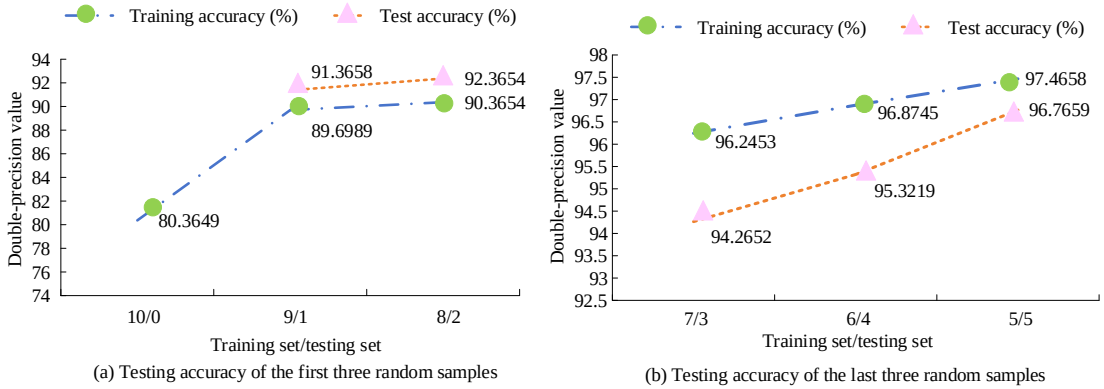


Figure 10: Financial risk warning chart for manufacturing enterprises

According to Figure 10, the accuracy of training set from multiple sessions for random samples 10/0, 9/1, and 8/2 was 80.36%, 89.70%, and 90.37%, respectively, while the accuracy of testing set was 0%, 91.37%, and 92.37%. For random samples 7/3, 6/4, and 5/5, the accuracy of training set was 96.25%, 96.87%, and 97.47%, and the corresponding accuracy of testing set was 94.27%, 95.32%, and 96.77%. The accuracy of training and testing both exceeded 80% and showed a steady growth trend, indicating that the improved HT-SVM algorithm effectively maintained high accuracy in financial risk warning for manufacturing enterprises. A financial risk warning and control model based on the improved HT-SVM algorithm is constructed using D1 (operational ability), D2 (development ability), D3 (profitability), and D4 (debt-paying ability). The predictions for each class in T-2 and T-3 years are shown in Figure 11.

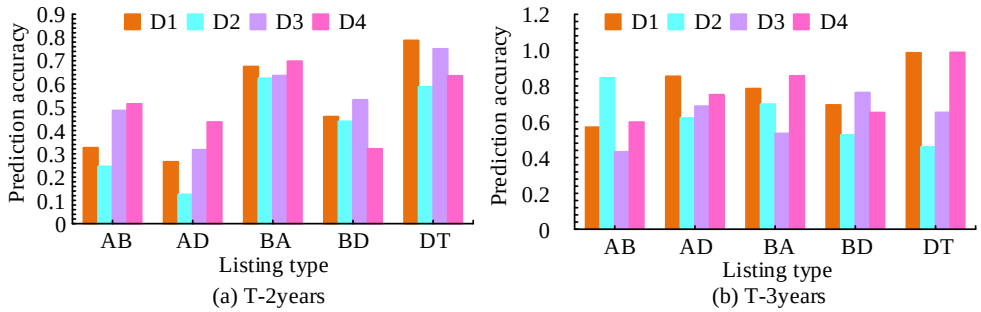


Figure 11: The predicted results of each data in each class for T-2 and T-3 years

In Figure 11, the status and transition relationship of listed companies are as follows: A (normal listing), B (ST), D (*ST), T (delisting consolidation period), and X (termination of listing). The state transition relationship is represented as: AB (normal→ST), AD (normal→*ST), BA (ST→normal), DT (*ST→delisting consolidation period), etc. According to Figure 11 (a), the average predicted values of AB, AD, BA, BD, and DT in T-2 years were 0.93%, 0.29%, 0.66%, 0.44%, and 0.69%, respectively, with a total average rate of 0.49%. According to Figure 11 (b), the average predicted values of AB, AD, BA, BD, and DT in T-3 years were 0.61%, 0.73%, 0.72%, 0.66%, and 0.77%, respectively, with a total average rate of 0.70%. From the results, the performance of T-3 increased compared to T-2, indicating good predictive ability. Five companies from the manufacturing listed companies on the main board of Shanghai and Shenzhen A-shares are taken as the financial health sample group

and six companies are taken as the financial risk sample group. In the application examples for enterprises in Shanghai and Shenzhen, the "Financial health" and "Risk" labels are determined by combining manual labeling and external audit results. First, the expert team conducts a preliminary assessment based on the enterprise's financial statements and performance indicators to identify possible financial health or risk conditions. Subsequently, after review and feedback from external auditing agencies, these labels are further verified and improved to ensure their accuracy and reliability. These two sample groups are then fed into the improved HT-SVM algorithm for financial risk warning and control of listed manufacturing companies. The predicted risk occurrence in the year T is obtained, and the risk prediction results are shown in Figure 12.

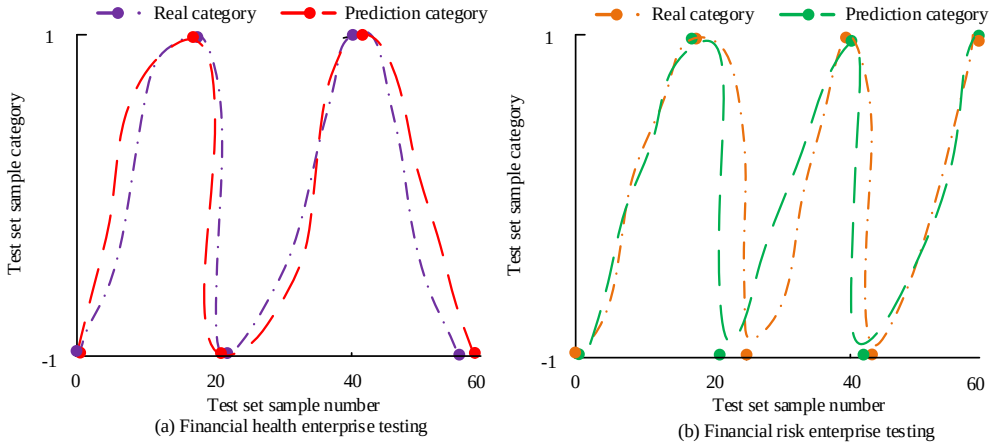


Figure 12: Financial risk warning and control model for manufacturing enterprises

From Figure 12, the error between the predicted results of the proposed method and the actual results was relatively small. In the health data samples, when the sample sizes were 10, 20, and 40, the predicted results were almost the same as the actual results. In the risk data sample, when the sample size is 20, there was a certain error between the predicted results and the true results. However, when the sample size was 40, the predicted results highly overlapped with the true results. The research has achieved good results in predicting financial risks in manufacturing enterprises through the HT-SVM model. To further verify its advancement, the proposed method in the research is compared with those in references [1-5]. The specific results are shown in Table 5.

Table 5: Comparative analysis of performance of different models

Method	Accuracy of the training set	Accuracy of the testing set	IR	Predicted time (s)
HT-SVM	0.963	0.943	1.00	8.4
Reference [1]	0.925	0.901	3.52	15.6
Reference [2]	0.911	0.885	4.04	0.4
Reference [3]	0.926	0.931	5.41	17.6
Reference [4]	0.951	0.934	3.22	16.7
Reference [5]	0.973	0.935	4.53	18.1

From Table 5, the HT-SVM model performed well in predicting financial risks in manufacturing enterprises. In the training and testing sets, the accuracy of HT-SVM reached 96.3% and 94.3% respectively, significantly higher than most methods in references [1] to [5]. The accuracy of the testing set in reference [1] was 90.1%, while the performance of reference [2] dropped more significantly, being only 88.5%. Although reference [3] had an accuracy of 93.1% in the testing set, the IR was 5.41, indicating its shortcomings in handling imbalanced samples. Furthermore, the IR of HT-SVM was 1.00, which was lower than that of all references, indicating that this model had a better ability to deal with class imbalance. Meanwhile, HT-SVM also performed well in prediction time, requiring 8.4 seconds. Compared with other methods, it had certain advantages. This series of outstanding performance indicators indicate that HT-SVM effectively enhances accuracy and efficiency in financial risk prediction tasks and has strong practical value.

4 Discussion

Based on the above experimental results, the HT-SVM has better classification performance compared with GA-SVM, PSO-SVM and OVO-SVM. Furthermore, the number of classifiers of the research method when dealing with multiple datasets is lower than that of the other three algorithms, which indicates that HT-SVM can effectively reduce model complexity, save computing resources, and improve real-time performance. Meanwhile, HT-SVM

also outperforms other algorithms on IR, especially when dealing with highly imbalanced sample distributions. Through reasonable cost guidance and classifier design, HT-SVM can eliminate the impact brought by class imbalance while maintaining good prediction performance. HT-SVM can achieve better results mainly due to its unique hierarchical structure design and improved mapping mechanism. Compared with the traditional SVM, HT-SVM builds a binary tree structure, allowing each node to focus on handling the boundary between the minority class and the majority class. This not only alleviates the class imbalance, but also improves the expressive ability of the model. Meanwhile, by introducing nonlinear mapping and AOA, HT-SVM can better map the original data to the high-dimensional feature space and effectively capture complex patterns in the data.

5 Conclusion

In response to the financial data distortion, difficulties in integrating multi-source data, model lag, and supply chain financial crises, an improved HT-SVM algorithm was proposed for financial risk warning and control in listed manufacturing enterprises. The algorithm optimized parameter optimization speed and utilized hierarchical threshold adaptive enhancement to improve global and local exploration capabilities. The experimental results showed that among the six random samples, the lowest accuracy of the training set obtained through multiple training was 80.3649%, and the highest accuracy was 97.4658%. The lowest accuracy of the testing set was 85.3694%, the highest accuracy was 96.7659%. The accuracy in training and testing sets was both above 80% and steadily increasing, indicating that the improved HT-SVM algorithm could maintain good accuracy in financial risk warning for manufacturing enterprises and improve the prediction accuracy. In the financial risk identification, 4 out of 5 enterprises could be identified in the 0-60 test set sample, with an accuracy of 80%. In the 80-140 testing set sample, 5 out of 6 enterprises could be identified, with an accuracy of 83.3%. Combining two types of risk samples, the overall prediction accuracy reached 81.8%, which could timely analyze the current business situation and take relevant measures to avoid financial risks when the enterprise predicted future risks. Although the research method has achieved certain results in the experiment, there are still certain limitations. For example, although combining Huffman trees with SVM improves the processing ability of imbalanced datasets, the complexity of the model increases the computational burden, which may pose a challenge to the applicability of ultra large scale application scenarios. Future research could focus on optimizing the model structure, reducing computational complexity to adapt to real-time application scenarios, and exploring integration with other machine learning algorithms to provide more flexible and efficient solutions for various types of data processing tasks.

Funding

This study was supported by the High-level Specialty in Big Data and Accounting with Distinctive Features (Provincial level project Wan Jiao Mi Gao [2023] No. 56), Teaching Team for Big Data and Financial Management Program (Provincial level project Wan Jiao Mi Gao [2022] No. 68) and 2022 Outstanding Young Talents Support Program for Higher Education Institutions (Provincial level project Anhui Education Commission Letter [2022] No. 371).

References

[1] Shu M, Wang Z, Liang J. Early warning indicators for financial market anomalies: A multi-signal integration approach. Journal of Advanced Computing Systems, 2024, 4(9): 68-84. https://doi.org/10.69987/JACS.2024.40907.

[2] Du L, An X. An enterprise financial credit risk measurement method based on differential evolution algorithm. International Journal of Information Technology and Management, 2025, 24(1-2):67-77. https://doi.org/10.1504/IJITM.2025.144106.

[3] Chen W. An enterprise financial data risk prediction model based on entropy weight method. International journal of industrial and systems engineering: International Journal of Industrial and Systems Engineering, 2023, 45(1):89-100. https://doi.org/10.1504/IJISE.2023.133533.

[4] Cao Q. An enterprise financial data leakage risk prediction based on ARIMA-SVM combination model. International Journal of Applied Systemic Studies, 2023, 10(3):169-181. https://doi.org/10.1504/IJASS.2023.134358.

[5] Zhang X. Financial risk monitoring and warning method of listed enterprises based on data mining. International Journal of Business Intelligence and Data Mining, 2025, 26(1-2):133-146. https://doi.org/10.1504/IJBIDM.2025.143932.

[6] Wang J, Hong S, Dong Y, Li Z, Hu J. Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting. Journal of computer science and software applications, 2024, 4(3): 1-7. https://doi.org/index.php/jcssa/article/view/100.

[7] Dessaint O, Foucault T, Frésard L. Does alternative data improve financial forecasting? The horizon effect. The Journal of Finance, 2024, 79(3): 2237-2287. https://doi.org/10.1111/jofi.13323.

[8] Okeke N I, Bakare O A, Achumie G O. Forecasting financial stability in SMEs: A comprehensive analysis of strategic budgeting and revenue management. Open Access Research Journal of Multidisciplinary Studies, 2024, 8(1): 139-149. https://doi.org/10.53022/oarjms.2024.8.1.0055.

[9] Lv M. Integrating ARIMA model for enhanced financial and tax data management and accurate departmental budget prediction. Informatica, 2025, 49(5): 19-36. https://doi.org/10.31449/inf.v49i5.6556.

[10] Jiao Z. Dynamic financial distress prediction using combined LASSO and GBDT algorithms. Informatica, 2024, 48(17): 139-152. https://doi.org/10.31449/inf.v48i17.6493.

[11] Gupta S K, Shukla D P. Handling data imbalance in machine learning based landslide susceptibility mapping: a case study of Mandakini River Basin, North-Western Himalayas. Landslides, 2023, 20(5): 933-949. https://doi.org/10.1007/s10346-022-01998-1.

[12] Song L, Chen Y. Does a non-performing assets disposal fund help control systemic risk? evidence from an interbank financial network in China. Financial Innovation, 2025, 11(1):1-45. https://doi.org/10.1186/s40854-024-00667-7.

[13] Tribak H, Gaou M, Gaou S. QR code recognition based on HOG and multiclass SVM classifier. Multimedia Tools and Applications, 2024, 83(17): 49993-50022. https://doi.org/10.1007/s11042-023-17398-z.

[14] Gao T, Duan L, Feng L. A novel blockchain-based responsible recommendation system for service process creation and recommendation. ACM Transactions on Intelligent Systems and Technology, 2024, 15(4): 1-24. https://doi.org/10.1145/3643858.

[15] Misita M, Spasojevic Brkic V, Mihajlovic I. Selection of an algorithm for the prediction of stoppages and/or failure of excavation units using supervised machine learning. IMCSM Proceedings-International May Conference on Strategic Management–IMCSM24, May 31, 2024, Bor. Technical Faculty in Bor, 2024, 20(1): 79-91. https://doi.org/10.5937/IMCSM24008M.

[16] Pan C. Construction of risk prediction models for enterprise finance sharing operations using K-Means and C4.5 algorithms. International Journal of Computational Intelligence Systems, 2024, 17(1):1-13. https://doi.org/10.1007/s44196-024-00608-3.

[17] Ramya D, Suresha. Reinforcement learning driven trading algorithm with optimized stock portfolio management scheme to control financial risk. SN Computer Science, 2025, 6(1):1-16. https://doi.org/10.1007/s42979-024-03555-0.

[18] Li X, Wang J, Yang C. Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy. Neural Computing and Applications, 2023, 35(3):2045-2058. https://doi.org/10.1007/s00521-022-07377-0.

[19] Chen Z S, Zhou J, Zhu C Y. Prioritizing real estate enterprises based on credit risk assessment: an integrated multi-criteria group decision support framework. Financial Innovation, 2023, 9(1):2939-2991. https://doi.org/10.1186/s40854-023-00517-y.

[20] Luo N, Yu H, You Z, Li Y, Zhou T, Han N. Fuzzy logic and neural network-based risk assessment model for import and export enterprises: A review. 2023, 1(1):2-11. https://doi.org/10.47852/bonviewJDSIS32021078.