Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/120/2025.18744

Optimization of XGBoost Bankruptcy Prediction Based on Four-Vector Optimization Algorithm

Yulin Jin[1,a,*]

[1]*Department of Economics, University of California, Berkeley, State of California, Berkeley, 94704, USA.*
a. yulinj@berkeley.edu
**corresponding author*

Abstract: In this paper, the XGBoost model is optimized based on the four-vector optimization algorithm to realize the effective prediction of enterprise bankruptcy. In the experiment, we adopted this advanced algorithm to greatly improve the predictive performance of the model, with the goal of providing a reliable bankruptcy risk assessment tool for enterprises and investors. By training the model, we observed a gradual decline in the fitness curve from 0.029 to 0.026. This trend indicates that with the increase of the number of training iterations, the performance of the model is constantly improving. The gradual slowing of fitness decline suggests that the model may have approached the optimal solution or become stable, which provides confidence for subsequent applications. In terms of specific prediction effect, the confusion matrix of the training set shows that the accuracy of the model is as high as 100%, while the accuracy of the test set is as high as 97.33%. Such high accuracy not only reflects the model's excellent ability on the seen data, but also shows that the model has good generalization ability and can maintain efficient performance on the unseen data. In addition, the performance indicator chart shows the trend of different performance indicators, including FM, J, Q, CA, DAUC, SE, SP. From the figure, we can see that the values of FM and J remain around 1.5, indicating that the model performs well in some ways. However, DAUC and SE showed less than ideal results, close to zero or negative, which means that on some performance indicators, the model still needs further optimization and improvement. To sum up, the research in this paper not only provides an effective tool for corporate bankruptcy prediction, but also provides data support for investors and managers in the decision-making process. By accurately predicting the risk of bankruptcy, relevant parties can take preventive measures to effectively reduce investment risks and improve decision-making efficiency. Therefore, improving the accuracy of corporate bankruptcy prediction can not only promote the sustainable development of enterprises, but also provide protection for the healthy operation of the economy in a larger scope.

Keywords: Four-vector optimization algorithm, XGBoost, Enterprise Bankruptcy prediction.

1. Introduction

Corporate bankruptcy is an important phenomenon in economic activities, and its research background mainly stems from the consideration of corporate management, financial risk and economic stability. With the change of the global economic environment, the external risks faced by

DOI: 10.54254/2755-2721/120/2025.18744

enterprises become more and more complex, and various factors such as market competition, technological change, policy adjustment and economic cycle may lead to the financial difficulties of enterprises [1]. By studying the behavioral patterns and related factors of corporate bankruptcy, scholars and practitioners hope to identify potential bankruptcy risks in advance and take necessary intervention measures [2].

In recent years, machine learning algorithms have played an increasingly important role in predicting corporate bankruptcies. Traditional bankruptcy prediction methods mostly rely on linear regression and financial ratio analysis. Although they can identify potential bankruptcy risks to some extent, their limitations lie in their inability to deal with complex nonlinear relationships and a large number of characteristic variables [3]. Machine learning techniques, especially algorithms such as decision trees, random forests, and support vector machines, can mine hidden patterns in data and effectively improve the accuracy and reliability of predictions.

The machine learning model also has the ability of automatic learning and self-optimization, and can update and adapt to the new market environment in time by learning historical data [4]. This enables companies to effectively assess their financial position and operational health in a dynamic economic context, develop responses in advance, and reduce the risk of insolvency. In addition, machine learning can process information from different data sources, including financial data, market conditions, social media sentiment, etc., to provide a more comprehensive assessment of bankruptcy risk. These characteristics make machine learning a powerful tool in the field of corporate bankruptcy prediction. In this paper, XGBoost is optimized based on the four-vector optimization algorithm to predict enterprise bankruptcy [5].

2. Data set source

The data set selected in this paper is an open source data set, which contains various business indicators of the enterprise and whether bankruptcy occurs, including a total of 18 business indicators, and the last item is whether the enterprise is bankrupt, with 1 for bankruptcy and 2 for non-bankruptcy. Some data sets are selected for presentation in this paper, as shown in Table 1.

Table 1: Some data sets.

X13	X14	X15	X16	X17	X18	status
191.226	163.816	201.026	1024.333	401.483	935.302	alive
160.444	125.392	204.065	874.255	361.642	809.888	alive
112.244	150.464	139.603	638.721	399.964	611.514	alive
109.59	203.575	124.106	606.337	391.633	575.592	alive
128.656	131.261	131.884	651.958	407.608	604.467	alive
2248	5864	5716	17730	17516	15482	failed
2583	6990	5948	19703	19037	17120	failed
-456	7512	4042	18963	27468	19419	failed
-1256	7240	-399	17299	29310	18555	failed
3010	6559	-1336	17440	29284	16858	failed

3. Correlation analysis

Correlation analysis is a statistical method used to assess the strength and direction of the relationship between two or more variables. The basic principle is to quantify the linear relationship between variables by calculating the correlation coefficient. Correlation analysis can reveal whether there is correlation between variables and its properties, such as positive correlation, negative correlation or no correlation [6]. When a change in one variable is accompanied by a change in another, it usually

DOI: 10.54254/2755-2721/120/2025.18744

indicates some correlation. The correlation coefficient between various business indicators of the enterprise and whether the enterprise is bankrupt is calculated and the correlation ranking is conducted, as shown in Figure 1.

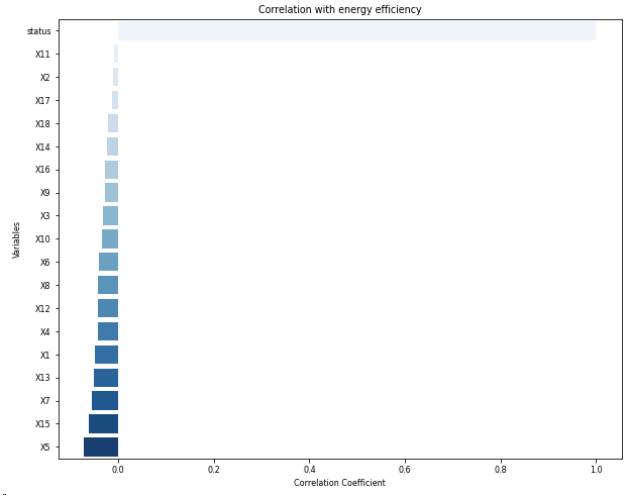

Figure 1: Correlation coefficient ranking.

4. Method

4.1. Four vector optimization algorithm

The four-vector optimization algorithm is an optimization method based on swarm intelligence, which is inspired by the cooperation and competition of different organisms in a specific environment in nature. This algorithm mainly finds the optimal solution of a problem by simulating the interaction between different individuals [7]. The core of this method is that through the combination of four key vectors, the state, goal and relative position of the individual are reflected, thus forming a dynamic search mechanism. The position of different individuals in the space is constantly updated, which simulates a process of survival struggle in the process of natural selection, making the whole group gradually move towards the optimal solution [8]. The schematic diagram of the four-vector optimization algorithm is shown in Figure 2.

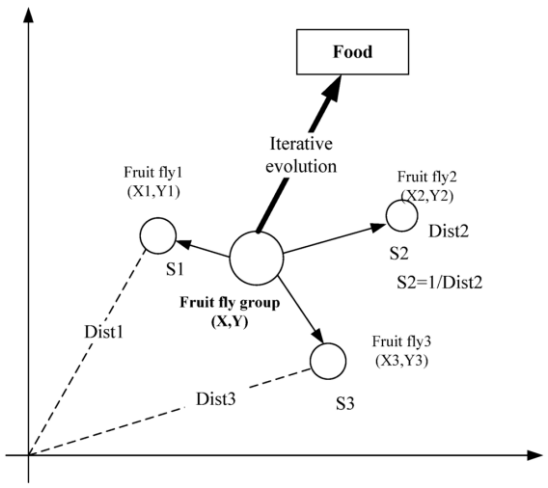

Figure 2: The schematic diagram of the four-vector optimization algorithm.

DOI: 10.54254/2755-2721/120/2025.18744

The specific operation process of the algorithm consists of several steps, each of which is realized through information sharing and self-updating among individuals. Each individual must not only consider their own state, but also pay attention to the performance of the surrounding individuals, so as to update their own position and speed. This information transfer mechanism enables individuals to conduct effective exploration in multidimensional space, avoiding the risk of falling into local optimal solutions. In addition, the four-vector optimization algorithm also increases the response to the deviation between the current state and the target through the evaluation of the individual state, and enhances the flexibility and adaptability of problem solving. Individuals in the group learn from each other and adjust continuously, thus pushing the whole towards the optimal solution.

4.2. XGBoost

XGBoost is an efficient and flexible gradient lifting decision tree (GBDT) algorithm that is widely used in classification and regression tasks. It gradually improves the predictive performance of the model by integrating multiple weak learners [9]. The key is the use of a gradient lifting method, where each new tree is built to correct the errors of the existing model on the training set, using the residual as the target of the new tree, thereby continuously improving the accuracy of the overall model.

The strength of XGBoost lies in the multiple technical improvements in its optimization process, including regularization strategies (L1 and L2 regularization) to prevent overfitting of the model. In addition, XGBoost uses parallel and distributed computing to speed up the training process, especially when the data sets are large. By automating feature selection and importance assessment, XGBoost further enhances the interpretability of the model, enabling users to better understand the impact of features on predicted results. The schematic diagram of the XGBoost algorithm is shown in Figure 3.

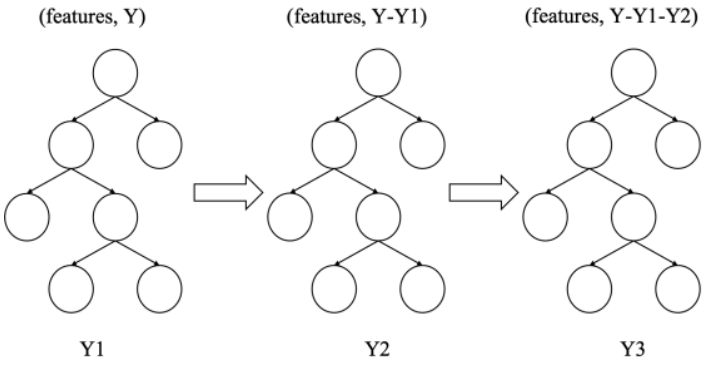

Figure 3: The schematic diagram of the XGBoost algorithm.

4.3. Optimization of XGBoost based on four-vector algorithm

Four-vector optimization algorithm is an emerging swarm intelligence optimization method, which shows unique effects and advantages in the process of optimizing machine learning models, especially XGBoost algorithm [10]. XGBoost is a powerful and flexible gradient lifting decision tree (GBDT) algorithm. Although XGBoost has high performance itself, it can provide optimization support for hyperparameter tuning, feature selection and model complexity control. By applying this optimization algorithm to XGBoost, it is possible to automatically find the best combination of hyperparameters in a wide search space, thereby improving the overall performance of the model.

In a four-vector optimization algorithm, each individual is represented as four key vectors that represent the individual's current state, target, velocity, and position change. By simulating the interaction between individuals, the four-vector optimization algorithm can effectively explore the combined space of hyperparameters in the population. In optimizing XGBoost, this approach starts

DOI: 10.54254/2755-2721/120/2025.18744

by randomly generating a group of individuals, each representing a parameter setting. Next, the population continually renews itself by evaluating the fitness of each individual, i.e. how its corresponding XGBoost model performs against cross-validation or other evaluation criteria. Based on the feedback information, the individual adjusts its position and speed, so that the overall search gradually moves closer to the parameter combination with better effect. This adaptive search capability based on swarm intelligence can avoid the local optimal problem in the traditional method and achieve the global optimal solution.

The four-vector optimization algorithm not only improves the performance of XGBoost model, but also provides an effective guarantee for the interpretability and robustness of the model. By monitoring the importance of features in the optimization process, the four-vector algorithm can automatically select the features that contribute the most to the model prediction results, reduce unnecessary noise and dimensions, and improve the stability and generalization ability of the model.

5. Result

The objective of this experiment is to optimize XGBoost based on the four-vector algorithm to predict whether an enterprise is bankrupt. In the parameter setting, the learning rate is set to 0.0003, min_child_weight is set to 3, the proportion of data used for training is set to 0.7 per tree, the proportion of features used for each tree is set to 0.7, the proportion of features used for each layer is set to 0.7, and the minimum loss reduction is set to 0.5.

Firstly, the fitness change curve of the model is output, as shown in Figure 4.

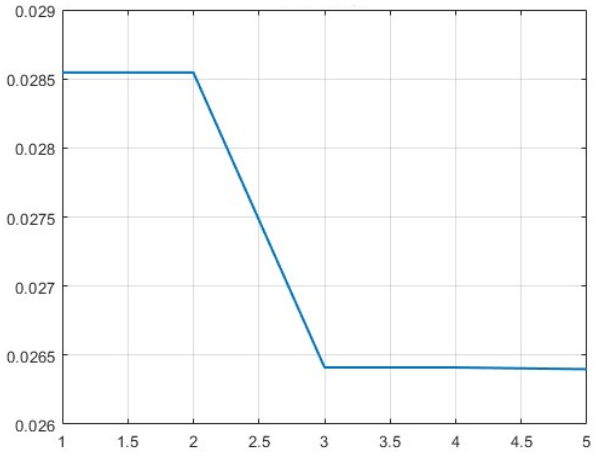

Figure 4: The fitness change curve of the model.

Figure 4 shows the trend of fitness with algebra. It can be seen from the figure that the fitness gradually decreases from 0.029 to 0.026, indicating that the performance of the model is gradually improved with the increase of the number of iterations. The decline rate of fitness gradually slows down, which means that the model is approaching the optimal solution or has become stable.

DOI: 10.54254/2755-2721/120/2025.18744

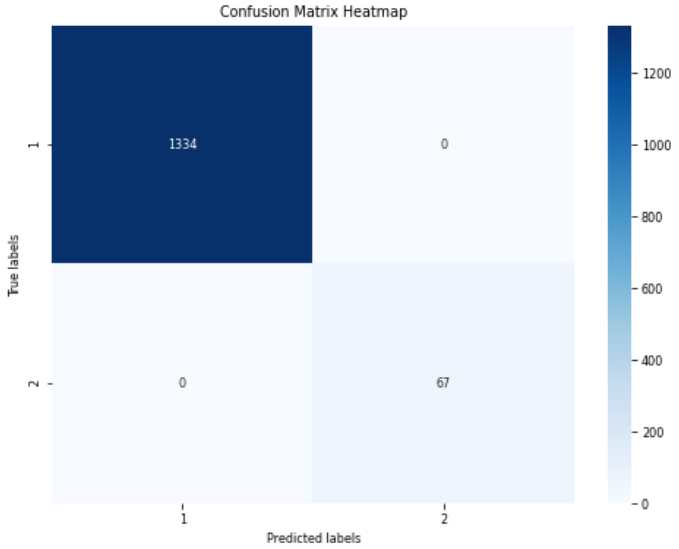

Figure 5: Tthe confusion matrix of the training set.

Figure 5 shows the confusion matrix of the training set. From the confusion matrix, it can be seen that the prediction accuracy of the model for enterprise bankruptcy and non-bankruptcy is 100%.

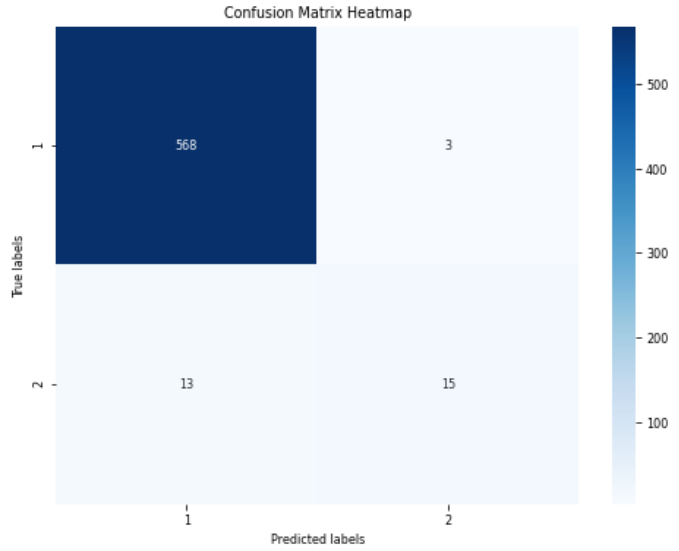

Figure 6: The confusion matrix of the test set.

Figure 6 shows the confusion matrix of the test set. According to the confusion matrix, the prediction accuracy of the model in the test set is 97.33%, which also reaches a high accuracy, indicating that the model has good generalization ability.

DOI: 10.54254/2755-2721/120/2025.18744

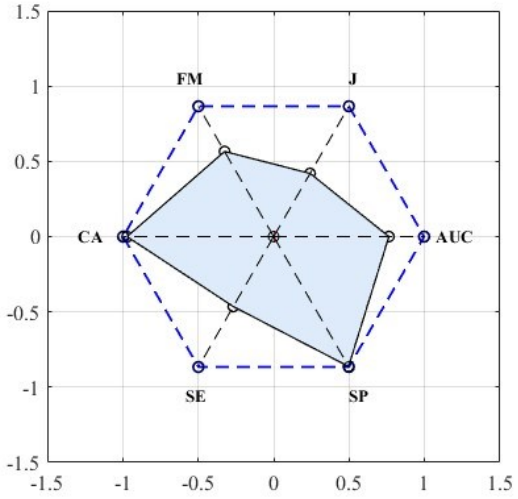

Figure 7: The performance indicator graph of the model.

Figure 7 is the performance indicator graph of the model, which shows the trend of different performance indicators (FM, J, Q, CA, DAUC, SE, SP). As can be seen from the graph: FM and J are around 1.5, while DAUC and SE are around 0 or negative. The model has different performance indicators, some indicators perform well, while others need to be further optimized, and the overall performance is good.

6. Conclusion

This paper optimizes the XGBoost model through the four-vector optimization algorithm, and successfully realizes the research goal of enterprise bankruptcy prediction. With the development of modern financial technology, accurately predicting the bankruptcy risk of enterprises is of great significance to enterprises, investors and the whole economic system.

Experimental results show that the fitness value of the model gradually decreases from 0.029 to 0.026, which indicates that the performance of the model is significantly improved with the increase of the number of iterations. The gradual slowing down of fitness decline rate also suggests that the model may be close to the optimal solution or tends to be stable, pointing to the effectiveness and scientificity of the model training process. After analyzing the confusion matrix between the training set and the test set, we found that the prediction accuracy of the model reached 100% on the training set, and also maintained a high level of 97.33 on the test set. This not only highlights the model's excellent ability to judge enterprise bankruptcy and non-bankruptcy, but also shows that the model has excellent generalization ability and can maintain good performance on unknown data.

In addition, through the analysis of performance indicators, the chart shows the change trend of different indicators (such as Fowlkes-Mallows index FM, J, Q, CA, DAUC, SE, SP) and the differences in the performance of the model on these indicators. While the values of FM and J are stable around 1.5, showing the effectiveness of the model in specific aspects, the results of DAUC and SE show certain limitations, even near zero or negative. This phenomenon suggests that the model still needs to be improved in some performance indexes to further improve its overall performance.

In summary, the research in this paper not only provides a new approach to corporate bankruptcy prediction based on modern machine learning techniques, but also provides a data-driven empirical foundation for broader economic decision-making. By accurately predicting the risk of enterprise bankruptcy, investors and business managers can take preventive measures, adjust investment

DOI: 10.54254/2755-2721/120/2025.18744

strategies, and make more objective and wise decisions in risk management, thus reducing economic losses and resource waste. The findings highlight the importance of machine learning technology in the economic field and point the way for future research to explore its application potential in other financial fields while continuously optimizing model performance, providing more scientific basis for corporate and investor decisions, thereby promoting sustainable economic development and prosperity.

References

[1] *Charalambous, Chris, Spiros H. Martzoukos, and Zenon Taoushianis. "Estimating corporate bankruptcy forecasting models by maximizing discriminatory power." Review of Quantitative Finance and Accounting 58.1 (2022): 297-328.*

[2] *Alam, Talha Mahboob, et al. "Corporate bankruptcy prediction: An approach towards better corporate world." The Computer Journal 64.11 (2021): 1731-1746.*

[3] *Kim, Hyeongjun, Hoon Cho, and Doo** Ryu. "Corporate bankruptcy predi*

[4] *Matenda, Frank Ranganai, et al. "Bankruptcy prediction for private firms in develo** economies: a sco** review and guidance for future research." Management Review Quarterly (2021): 1-40.*

[5] *du Jardin, Philippe. "Designing topological data to forecast bankruptcy using convolutional neural networks." Annals of Operations Research 325.2 (2023): 1291-1332.*

[6] *Cao, Yi, et al. "A two-stage Bayesian network model for corporate bankruptcy prediction." International Journal of Finance & Economics 27.1 (2022): 455-472.*

[7] *Chen, Tsung-Kang, et al. "Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports." Expert Systems with Applications 233 (2023): 120714.*

[8] *Jones, Stewart. "A literature survey of corporate failure prediction models." Journal of Accounting Literature 45.2 (2023): 364-405.*

[9] *Vuković, Bojana, et al. "Corporate bankruptcy prediction: evidence from wholesale companies in the Western European countries." Ekonomicky casopis 68.5 (2020): 477-498.*

[10] *Rajendran, Surendran, et al. "Automated segmentation of brain tumor MRI images using deep learning." IEEE Access 11 (2023): 64758-64768.*