International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n1.29
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Prediction Model for Credit Risk Measurement of Small and Micro Enterprises Based On Particle Swarm Optimization random forest algorithm

Feng Chai

School of Anhui University of Finance and Economics, Bengbu, 233000, China

ABSTRACT

In the context of rapid economic development, the credit risk assessment of small and micro enterprises has become the focus of attention in the financial field. The random forest algorithm is widely used in credit risk assessment due to its high accuracy and robustness, but it has some problems, such as difficulty in parameter selection and performance degradation in processing unbalanced datasets. In this study, we propose a credit risk prediction model for small and micro enterprises based on Particle Swarm Optimization Random Forest Algorithm (PSO-RF). The particle swarm optimization effectively solves the problem of random forest parameter selection and improves the prediction performance of the model through the balance of global search and local search. Experimental results show that the PSO-RF algorithm shows significant performance advantages in credit risk prediction, performs well in dealing with unbalanced datasets, and has certain advantages in feature selection. This study provides new ideas and methods for the credit risk assessment of small and micro enterprises, and is of great significance for improving the risk management capabilities of financial institutions and supporting the healthy development of small and micro enterprises.

KEYWORDS

Particle swarm optimization; Random forest algorithm; Credit risk; Predictive models

1. INTRODUCTION

With the development of the economy, the credit risk assessment of small and micro enterprises has attracted the attention of the financial community. Artificial intelligence technology, especially random forest algorithms, is widely used in credit risk assessment due to its high accuracy and robustness. However, random forests have limitations such as parameter selection, unbalanced data processing, and model interpretability. In this study, a random forest algorithm (PSO-RF) based on particle swarm optimization is proposed to address these challenges. The particle swarm optimization algorithm effectively optimizes the random forest parameters and improves the prediction performance by balancing global and local searches. The purpose of this study is to construct a more efficient and accurate credit risk prediction model for small and micro enterprises through PSO-RF, and to explore its potential in dealing with unbalanced data and improving the explanatory nature of the model, so as to provide a new method for credit risk assessment.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

2. PARTICLE SWARM OPTIMIZATION RANDOM FOREST ALGORITHM (PSO-RF)

2.1. Random Forest Algorithm

The Random Forest (RF) algorithm is an ensemble learning algorithm proposed by Leo Breiman et al. in 2001. It improves the performance of the model by building multiple decision trees for classification or regression predictions and merging the results in a voting or averaging fashion. The core of the random forest algorithm lies in its two concepts of "random" and "forest". "Random" refers to the random selection of decision trees in the process of building them, while "forest" refers to the collection of multiple decision trees.

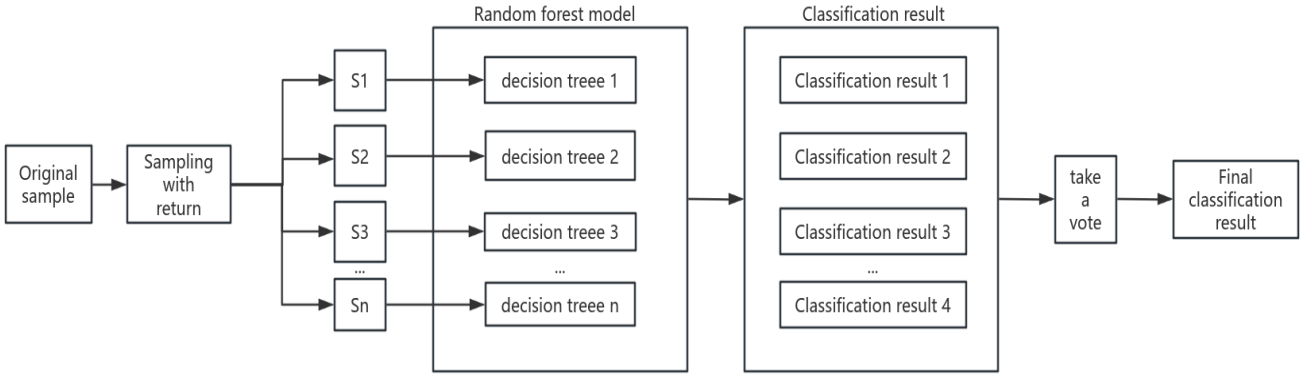

Figure 1. Schematic diagram of the random forest model

The basic principle of the random forest algorithm is "brainstorming", which means improving the accuracy and robustness of the overall model by combining the prediction results of multiple decision trees. When constructing each decision tree, the random forest adopts two randomness strategies: first, at each decision point in the decision tree, not all features are considered, but a portion of features are randomly selected as candidates for splitting nodes; The second approach is to use bootstrap sampling on the training data for generating decision trees, which involves extracting samples from the original training data with dropout, forming multiple sets of different training sets. Each tree is trained using different training sets.

2.2. Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization algorithm, PSO is an optimization algorithm based on swarm intelligence, proposed by Eberhart and Kennedy in 1995 [1-2]. The design inspiration of the PSO algorithm comes from the simulation of bird foraging behavior, which seeks the optimal solution by simulating the collective behavior of organisms such as bird or fish schools. In the PSO algorithm, each solution is abstracted as a particle in the search space, representing a potential solution in the solution space. Each particle has two properties in the solution space: position and velocity. The position of the particle represents the current state of the solution, while the velocity determines the direction and step size of the particle's movement in the solution space. Particles update their position and velocity by tracking two extremes: individual extremum (pbest) and global extremum (gbest). Individual extremum is the optimal solution found by the particle itself, while global extremum is the optimal solution for all particles in the entire particle swarm.

Particle velocity update formula:

$$v_{i,d+1} = w \cdot v_{i,d} + c_1 \cdot r_{1,i} \cdot \left(pbest_{i,d} - x_{i,d}\right) + c_2 \cdot r_{2,i} \cdot \left(gbest_d - x_{i,d}\right)$$

Among them:$v_{i,d+1}$is the velocity of the i-th particle in the d+1st iteration. W is the inertia weight used to balance the previous velocity of particles and their tendency to move towards the optimal

position.c_1 and c_2 are cognitive and social factors, respectively, typically set as constants to adjust the intensity of random searches. $r_{1,i}$ and $r_{2,i}$are independent and identically distributed random numbers within the interval [0, 1], introducing randomness into the algorithm. $pbest_{i,d}$ is the individual optimal position of a particle in the d-th iteration. $gbest_d$is the global optimal position of all particles in the d-th iteration. It is the position of the particle in the second iteration.

Particle position update formula:

$$x_{i,d+1} = x_{i,d} + v_{i,d+1}$$

Among them, $x_{i,d+1}$ is the new position of the particle in the d+1st iteration.

In the field of credit risk assessment, The PSO algorithm can be used to optimize the parameters of credit scoring models and improve their predictive accuracy. By adjusting model parameters, The PSO algorithm can help financial institutions more accurately assess the credit risk of borrowers, thereby making more reasonable credit decisions. In addition, The advantages of PSO algorithm in dealing with high-dimensional optimization problems make it widely used in financial engineering fields such as financial derivative pricing and portfolio optimization [3].

2.3. PSO-RF Algorithm

The PSO-RF algorithm is an ensemble learning framework that combines the global search ability of Particle Swarm Optimization (PSO) with the classification or regression ability of Random Forest (RF). This algorithm utilizes PSO to optimize the hyperparameters of RF, such as the number of trees, depth, and number of feature selections, to improve the predictive performance of the model. Each particle represents a combination of hyperparameters, and the performance of the constructed RF model is evaluated through cross validation [4].

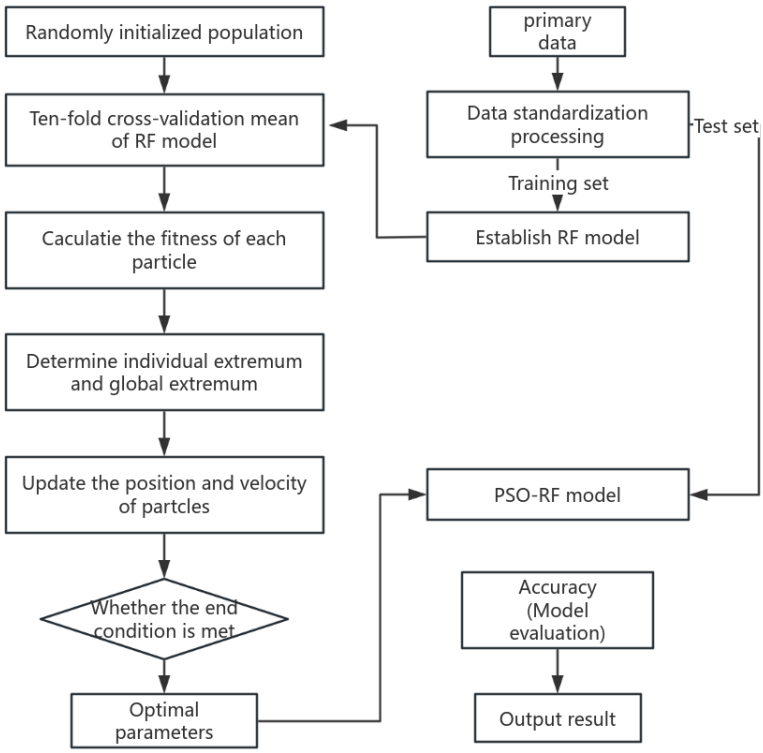

Figure 2. The prediction process of PSO-RF model

Step 1: Initialize and generate an initial particle swarm, with each particle's hyperparameter combination randomly generated.

Step 2: Fitness evaluation. For each particle, a random forest model is constructed using training data and its fitness is calculated. The fitness is usually 1 minus the model's error rate or accuracy.

Step 3: Update individual extremum. If the fitness of the current particle is better than its historical optimal fitness, update the individual extremum.

Step 4: Global extremum update. If the fitness of the current particle is better than the global optimal fitness among all particles, update the global extremum.

Step 5: Update velocity and position, adjust the hyperparameter combination of particles based on individual and global extremum using the velocity and position update formula of PSO.

Step 6: Iteration, repeat steps 2-5 until the maximum number of iterations is met or other stopping conditions are reached.

The PSO-RF algorithm combines Particle Swarm Optimization (PSO) and Random Forest (RF) to optimize model performance by iteratively adjusting the hyperparameters of RF. The position of particles represents the combination of hyperparameters, and the velocity represents the amount of adjustment. The key to the algorithm lies in selecting the appropriate inertia weight w, cognitive factor c1, and social factor c2, as well as determining the particle swarm size and maximum number of iterations. Build RF models for each iteration and evaluate performance using fitness functions based on prediction errors, such as cross entropy loss, accuracy, mean square error, or root mean square error. The global search capability of PSO enables the algorithm to effectively explore hyperparameter space, improve the model's generalization ability and prediction accuracy, especially in handling complex and high-dimensional data [5].

3. ALGORITHM EXPERIMENTS

3.1. Data Collection and Processing

The data is sourced from the financial information service platform API, including financial indicators, reputation, innovation effectiveness, and default information of small and micro enterprises. Data preprocessing is crucial, involving handling missing values, outliers, and consistency, using interpolation and mode to fill in missing data, box plots, standard deviations, or Z-scores to handle outliers. The dataset is divided into a training set and a testing set, with categorical variables converted through unique hot encoding. The problem of sample imbalance is solved through sampling, undersampling, or synthesizing samples. The preprocessed data is used to train the PSO-RF model and evaluate its performance in credit risk prediction for small and micro enterprises.

3.2. Experimental Design

Describe the experimental setup, including parameter selection and experimental steps.

3.3. Analysis of Experimental Results

Based on the above data and parameter selection, experiments were conducted using MatlabR2014a, and the experimental results are shown in Figures 3, 4, 5, and Table 1. It can be seen that the PSO-rf algorithm converges after about 40 rounds of evolution, and the mean square error tends to stabilize; According to the experimental results in Table 3, it can be seen that regardless of the MAE From the perspectives of MSE, RMSE, or MAPE, The PSO-rf algorithm has better errors and prediction accuracy than traditional random forest algorithms. According to the comparison between the two algorithms in Figure 4 and the actual values, it can be seen that the predicted values of the PSO-rf algorithm are close to the actual values and more stable compared to the random forest algorithm.

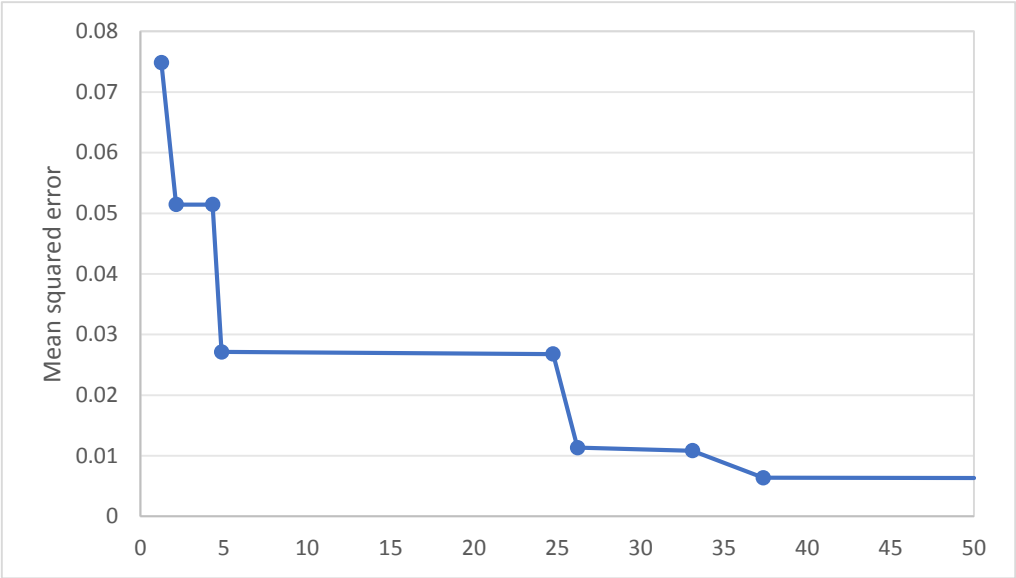

Figure 3. Convergence Curve

Table 1. Comparison of errors between rf and PSO-rf

Algorithm model	Average absolute error (MAE)	Mean squared error (MSE)	Root mean square error (RMSE)	Mean absolute percentage error (MAPE)
Traditional rf aigorithm	0.3836	0.2566	0.4921	10.36%
PSO-rf aigorithm	0.1225	0.0482	0.2133	3.20%

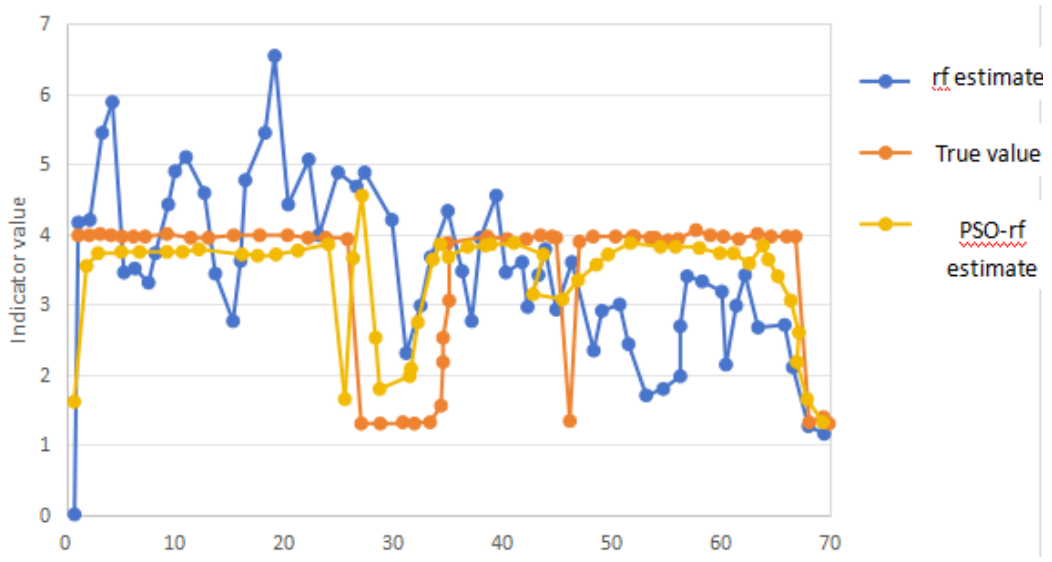

Figure 4. Comparison of True Values of Predictions Before and After Optimization

Figure 5 shows the comparison of the error between the predicted values and the true values before and after the optimization of the PSO-rf algorithm. It can be seen that the PSO-rf algorithm outperforms the RF algorithm in terms of stability and prediction error.

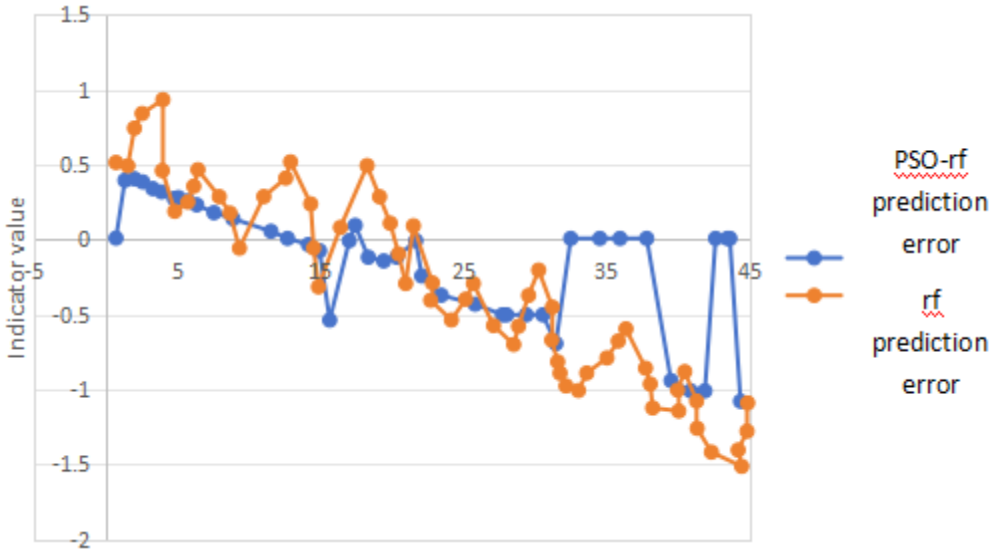

Figure 5. Comparison of errors before and after optimization

Based on the above experimental results, The PSO-RF algorithm has good performance in prediction error and stability, and has certain advantages compared to RF. Its application in credit risk measurement and prediction of small and micro enterprises has practical significance and can improve prediction accuracy.

4. RESULTS AND DISCUSSION

Algorithm performance: The PSO-RF algorithm has shown significant performance advantages in credit risk prediction tasks. Whether from MAE From the perspective of MSE, RMSE, or MAPE indicators, The PSO-RF algorithm is superior to traditional random forest algorithms. This indicates that the optimization of random forest parameters through particle swarm optimization can effectively improve the accuracy and stability of the model in credit risk prediction.

Prediction stability: Figure 4 shows the trend of prediction error variation of the PSO-RF algorithm in multiple iterations. It can be seen that as the iteration progresses, the prediction error of the model gradually decreases and tends to stabilize, reflecting the strong generalization ability and stability of the PSO-RF algorithm.

Handling imbalanced data: Experimental results show that, The PSO-RF algorithm has shown good performance in handling imbalanced datasets. The random forest model optimized by particle swarm optimization can better adapt to the imbalance of data, thereby improving the accuracy of prediction for minority categories.

Model interpretability: Although random forests themselves have certain limitations in model interpretability, the PSO-RF algorithm can improve the interpretability of the model to a certain extent by optimizing parameters. In the experiment, it was observed that the PSO-RF algorithm has more advantages in feature selection, which helps to identify key features that have a greater impact on credit risk.

Practical application potential: The PSO-RF algorithm has high potential in practical applications. For financial institutions, this algorithm can improve the efficiency and accuracy of credit risk assessment, providing more reliable decision support for loan approval, credit policy formulation, and risk control.

In summary, the credit risk prediction model for small and micro enterprises based on particle swarm optimization and random forest algorithm proposed in this study is of great significance in both theory and practice. Future research can further explore the performance of the PSO-RF algorithm on

different types and scales of datasets, as well as how to further improve the interpretability and adaptability of the model. In addition, applying this algorithm to practical financial scenarios to provide intelligent support for credit decisions of small and micro enterprises is also an important direction worthy of in-depth research.

5. CONCLUSION

This study proposes a small and micro enterprise credit risk prediction model based on Particle Swarm Optimization Random Forest Algorithm (PSO-RF), and conducts experimental verification on it. Based on the analysis of experimental results, the following conclusions can be drawn:

Performance advantage: The PSO-RF algorithm has shown significant performance advantages in credit risk prediction, with better prediction accuracy and stability than traditional random forest algorithms.

Handling imbalanced data: The PSO-RF algorithm performs well in handling imbalanced datasets, which is particularly important for credit risk assessment as default samples are often small.

Model interpretability: Although the random forest algorithm has certain limitations in model interpretability, the model optimized through PSO has advantages in feature selection, which helps identify key features that have a significant impact on credit risk.

Practical application potential: The PSO-RF algorithm has high potential in practical applications, which can improve the efficiency and accuracy of credit risk assessment for financial institutions, and provide more reliable decision support for loan approval, credit policy formulation, and risk control.

Future research directions and possible approaches for algorithm improvement

Parameter optimization: Further explore the selection of inertia weights, cognitive factors, and social factors in the PSO algorithm, as well as the determination of particle swarm size and maximum iteration times, to optimize the performance of the PSO-RF algorithm.

Feature selection and dimensionality reduction: Combining feature selection and dimensionality reduction techniques, such as principal component analysis (PCA) or linear discriminant analysis (LDA), to further improve the predictive accuracy and interpretability of the model.

Algorithm extension: Attempt to apply the PSO-RF algorithm to other types of credit risk assessment scenarios, such as personal credit scoring, corporate credit rating, etc., to verify its universality and effectiveness.

Real time data integration: Combining real-time data with external information sources such as social media, macroeconomic indicators, etc., to build a more dynamic and comprehensive credit risk assessment model.

Explanatory enhancement: Explore how to further enhance the model interpretability of the PSO-RF algorithm, such as using methods such as SHAP (Shapley Additive exPlaneations), to provide more in-depth model interpretation.

Integrate other algorithms: Attempt to integrate the PSO-RF algorithm with other machine learning algorithms (such as support vector machines, neural networks, etc.) to build a more robust and efficient credit risk assessment model.

Through the above research directions and possible approaches for algorithm improvement, it is expected to further enhance the performance and practicality of PSO-RF algorithm in the field of credit risk assessment, and provide more reliable and efficient support for the healthy development of small and micro enterprises.

ACKNOWLEDGEMENTS

Funded by the Innovation and Entrepreneurship Training Program for College Students of Anhui University of Finance and Economics (202210378242)

REFERENCES

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[2] Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization. IEEE International Conference on Neural Networks, 4, 1942-1948.

[3] Zhao Jing, Zhang Wei, Li Li. (2017). Research on credit risk assessment model based on improved particle swarm optimization random forest. Computer Engineering and Applications, 53(5), 141-146.

[4] Li Qiang, Zhang Xiaohui, Hu Xiaoming. (2018). Research on credit risk assessment model based on particle swarm optimization support vector machine. Computer Engineering and Applications, 54(6), 173-177.

[5] Wang Fang, Li Li, Liu Jie. (2016). Research on credit risk assessment model based on particle swarm optimization neural network. Computer Engineering and Applications, 52(10), 174-179.