The Establishment and Research of Random Forest Quantitative Classification Mode

Junling Sun[1], Gaoping Wang[2]
[1]*College of Computer and Artificial Intelligence, Henan Finance University, Zhengzhou, Henan, China*
[2]*Department of Computer, Guangzhou College of Applied Science and Technology, Guangzhou, Guangdong, China*

Abstract: With the advancement of China's financial deepening reforms, research on the application of quantitative investment in the A-share market has gradually increased. Exploring the application of the Random Forest algorithm in stock price classification is of great significance for advancing the development of quantitative stock selection theories. This paper proposes an ensemble learning method based on Random Forest for the problem of quantitative classification. Multiple technical indicators of stock data are selected to construct feature vectors, with a prediction horizon of 5 days, and the target vector prediction results are categorized into 6 classes. By constructing multiple decision trees and incorporating feature importance evaluation, a Random Forest model corresponding to the stocks is established. Experimental results demonstrate that this model can reduce the risk of overfitting while improving classification accuracy. Compared with traditional classification algorithms, this model outperforms them in terms of accuracy, recall rate, and other metrics.

Keywords: Random Forest Algorithm; Quantitative Classification; Price Prediction; Feature Selection; Financial Data Analysis

1. Introduction

With the advent of the big data era, the demand for classifying high-dimensional data is increasingly growing in fields such as financial quantitative investment, medical diagnosis, and industrial prediction. Traditional classification models [1] (such as logistic regression and support vector machines) have limitations when dealing with nonlinear and high-noise data. Random forest, as an ensemble learning algorithm, exhibits good performance in stock price classification and prediction due to its robustness and interpretability. Compared to traditional prediction methods, it offers better characteristics. By optimizing the random forest model [2], this study aims to enhance classification performance for high-dimensional and imbalanced data, reduce the model's sensitivity to noisy data, and assist investors in decision-making analysis through its classification results.

2. Random Forest Algorithm and Principles

Random forest employs an ensemble learning approach [3], combining multiple weak learners (such as decision trees) to construct a strong learner. Multiple models are trained in parallel. Through Bootstrap sampling, (N) samples are randomly drawn with replacement from the original dataset (repetition allowed) to generate (T) subsets (each subset trains one tree). Multiple decision trees are constructed, and their results are integrated through a voting mechanism. Features are randomly selected: when each decision tree splits, only (m) candidate features are randomly selected from the total (M) features, which reduces the correlation between trees and enhances generalization ability. Unseen data is used for validation to mitigate overfitting risk. For each subset, an unpruned decision tree is generated using the CART [4] algorithm (minimizing the Gini index). Finally, the final category is determined through a voting method (majority voting).

The mathematical expression is:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_i(x), ..., h_T(x)\} \quad (1)$$

where x is the input feature vector, $h_i(x)$ is the ith decision tree that outputs a category prediction, T is the number of decision trees, and mode represents the mode operation, i.e., the category with the most votes from all

decision tree predictions.

3 Construction and Optimization of a Random Forest Model for Stock Price Prediction

Constructing a stock price prediction model based on random forest involves integrating multiple market variables, including but not limited to historical price data, trading volume, technical indicators (such as RSI, MACD), market indices, and financial data. Given the time series nature of stock price data, special attention needs to be paid to temporal dependency. The sliding window method is adopted in implementation, where the historical data window serves as features and the price at the next time point serves as the target variable. The sampled data must maintain temporal order to avoid data leakage (ensuring that the training set precedes the test set in time). Finally, the prediction classification results are converted into buy/sell signals: when the prediction classification result is significantly higher than the current price by a certain threshold, a buy signal is generated; conversely, a sell signal is generated.

Model Evaluation should not only consider prediction accuracy but also focus on the effectiveness of risk management. Key evaluation indicators include: 1) prediction error indicators [5] (MSE, MAPE); 2) strategy returns (annualized return, Sharpe ratio); 3) risk indicators (maximum drawdown, volatility), etc.

3.1 Data Preprocessing and Modeling

This model selects the date, closing price, opening price, high price, low price, and trading volume of sz000651 and sz000659 from the A-share market for each trading day from March 15, 2022, to March 15, 2025, as samples. NaN and other missing values are deleted, and the stock data is adjusted for rights and dividends.

The closing price five days later is compared to the closing price of the same day and categorized into six classes: increased by more than 4%, increased by 2-4%, increased by 0-2%, decreased by 0-2%, decreased by 2-4%, and decreased by more than 4%. Modeling and calculations are based on Python code.

Feature Selection: Some common features are extracted from the raw data, such as moving averages, Relative Strength Index (RSI), MACD, volume ratio, OBV, etc., with NaN and other missing values deleted.

Target Variable: The target variable is defined as the percentage change over a future period N (N=5 days). The closing price five days later is compared to the closing price of the same day and categorized into six classes: increased by more than 4%, increased by 2-4%, increased by 0-2%, decreased by 0-2%, decreased by 2-4%, and decreased by more than 4%. Modeling and calculations are based on Python code.

Here, the Random Forest classification model in scikit-learn [6] is used for training. Eighty percent of the data is used for training, and 20% is used for testing. The Feature Vector 1 ['SMA_10', 'SMA_50', 'RSI', 'MACD', 'Signal_Line', 'Vol-bi'] included: 10-day SMA (Simple Moving Average), 50-day SMA, RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), MACD Signal Line, and Volume Ratio (Vol-bi). The Feature Vector 2 ['SMA_10', 'ROC', 'J', 'D','K','RSI', 'MACD', 'Signal_Line', 'Vol-bi'] comprised: 10-day SMA, ROC (Rate of Change), KDJ stochastic oscillator (J-line, D-line, K-line), RSI, MACD, MACD Signal Line, and Volume Ratio (Vol-bi). The Random Forest model was constructed using stock SZ000659 (Zhuhai Zhongfu Enterprise) and SZ000651 (Gree Electric) as the underlying asset. Model performance metrics are detailed in Tables 1 through 4, where precision is the prediction accuracy, recall is the recall rate, f1-score is a comprehensive score of precision and recall, and support is the number of supported samples.

Table 1. Random Fores Model Target Performance about Feature Vectors One-sz000659

Target	Accuracy	Recall	F1-Score	support
0-2% gain	0.40	0.27	0.32	15
2-4% gain	0.11	0.14	0.12	7
>4 % gain	0.77	0.69	0.73	29
0-2% drop	0.41	0.42	0.42	26
2-4% drop	0.33	0.41	0.37	17
>4 % drop	0.59	0.61	0.60	31

Table 2. Random Fores Model Target Performance about Feature Vectors Two-sz000659

Target	Accuracy	Recall	F1-Score	support
0-2% gain	0.50	0.20	0.29	15
2-4% gain	0.29	0.29	0.29	7
>4 % gain	0.83	0.66	0.73	29

0-2% drop	0.48	0.58	0.53	26
2-4% drop	0.29	0.41	0.34	17
>4 % drop	0.50	0.55	0.52	31

Table 3. Random Fores Model Target Performance about Feature Vectors One-sz000651

Target	Accuracy	Recall	F1-Score	support
0-2% gain	0.43	0.26	0.32	39
2-4% gain	0.33	0.22	0.27	18
>4 % gain	0.55	0.67	0.60	18
0-2% drop	0.36	0.77	0.49	22
2-4% drop	0.36	0.29	0.32	14
>4 % drop	0.50	0.36	0.42	14

Table 4. Random Fores Model Target Performance about Feature Vectors Two-sz000651

Target	Accuracy	Recall	F1-Score	support
0-2% gain	0.44	0.36	0.39	39
2-4% gain	0.50	0.28	0.36	18
>4 % gain	0.59	0.56	0.57	18
0-2% drop	0.29	0.59	0.39	22
2-4% drop	0.46	0.43	0.44	14
>4 % drop	0.50	0.29	0.36	14

3.2 Model Classification Performance Analysis

From the classification performance perspective, the models demonstrate higher accuracy and recall rates for categories with >4% price increases and >4% declines.

The four models achieved classification precision scores of 2.61, 2.89, 2.53, and 2.7 respectively - showing significant improvement over the baseline probability of 1. Model performance varies with feature selection. Feature Vector 2 demonstrates superior performance with an average 0.23-point improvement over Feature Vector 1.

Trading Signals: Buy Signal: When models classify into the >4% increase category. Sell Signal: When models classify into either >4% decline or 2-4% decline categories.

Most robust predictions occur for extreme movements (>4% thresholds). Vector 2 configurations consistently outperform Vector 1. All models exceed random chance benchmarks.

3.3 Parameter Optimization for Random Forest Models

3.3.1 Key parameters

Random Forest parameters can be categorized into two groups, with optimization priorities as follows:

Framework Parameters (control overall structure): n_estimators: Number of trees in the forest, increasing this enhances model stability but raises computational costs. max_features: Number/ratio of features considered for splitting at each node. Common values include sqrt, log2, or fractions (e.g., 0.3, 0.5). Single-Tree Parameters (control individual tree growth): max_depth: Maximum depth of a tree. Smaller values prevent overfitting; None allows full expansion. min_samples_split: Minimum samples required to split a node. Higher values regularize the model.

Key parameters (such as n_estimators, max_depth, max_features) are adjusted, and Grid Search is used to optimize these parameters. These parameters are summarized in Table 5.

Table 5. Random Fores Parameters

Parameters	Definition	Optimization Methods	Suggested Range
n_estimators	Number of Decision Trees	Incremental Testing	200-500
max_features	Number of Candidate Features for Node Splitting	Grid Search	Sqrt(n) or 0.3-0.5
max_depth	Maximum Depth of a Single Tree		5-15
min_samples_split	Minimum Number of Samples Required for Node Splitting	Adjust Based on Data Volume	2-20

Optimization Priority: First, increase n_estimators to stabilize the model. Then adjust max_features to balance bias-variance trade-off. Finally, fine-tune tree-specific parameters like max_depth and min_samples_split.

Optimization Tools: GridSearchCV [7]: Exhaustive search over predefined parameter combinations, ideal for small parameter spaces. RandomizedSearchCV: Efficient for large parameter spaces by sampling random combinations. Bayesian Optimization: Suitable for high computational cost scenarios, using probabilistic models to guide parameter selection.

3.3.2 Parameter optimization process

Data Preparation:

Features: Technical indicators (RSI, MACD), volatility, volume change rate, etc.

Label: Six classification based on 5-day price movement.
Splitting: Time-series cross-validation to avoid look-ahead bias
Parameter Grid setting: param_grid= {'n_estimators': [100, 200, 300], 'max_depth': [5, 10, None], 'max_features': ['sqrt', 0.3, 0.5], 'min_samples_split': [2, 5, 10]}
Optimization Steps:
Step 1. Use `GridSearchCV` to explore the grid and select the best parameters based on OOB (Out-of-Bag) score.
Step 2. Validate with time-series splits to ensure robustness.
3.3.3 Model evaluation
After conducting grid testing on the model, the optimized training parameters are {'n_estimators': 200, 'max_depth': 10, 'max_features': 0.5, 'min_samples_split': 5}. Assessing the feature importance ranking of the optimized model, The Feature Importance Scores ranking results are as follows:
No1. MACD (0.32)
No2. Volume Change Rate (0.28)
No3. RSI (0.25)
No4. MA10 (0.15)

4 Experiment and Result Analysis

4.1Experiment Design

4.1.1 Dataset
The dataset used in this experiment consists of A-share listed company stock data spanning from March 15, 2023, to March 15, 2025.
4.1.2 Feature vectors
The feature vectors include Moving Average Convergence Divergence (MACD), Moving Average (MA), Relative Strength Analysis (RSA), and Volume Ratio.
4.1.3 Target variable
The target variable is the stock price movement 5 days later, categorized into six types: upswing, downturn, increase by more than 5%, decrease by more than 5%, increase by more than 10%, and decrease by more than 10%.
4.1.4 Comparison models
The comparison models utilized in this experiment are Logistic Regression [8] and Support Vector Machine (SVM) [9].
4.1.5 Evaluation metrics
The evaluation metrics employed include Accuracy, Recall, F1-Score, and support (number of samples).

4.2 Result Comparison

Taking the prediction of a 4% increase and a 4% decrease in stock price 5 days later as examples, the performance metrics are summarized in Table 6.

Table 6. Model Performance

MODEL	Accuracy	Recall	F1-Score	support
Random fores	62.3%l	59%0	0.61	213
GBDT	49.7%	42%	0.45	153
SVM	45.2%	39%	0.42	161
Lunch	Rice	150		35

4.3 Feature Importance Analysis

The experiment revealed that MACD (with an importance of 0.32) and Volume Ratio (0.28) have a significant impact on the classification results, aligning with domain knowledge.

5. Discussion and Future Work

This paper addresses the stock price prediction problem by proposing a stock price movement prediction model based on Random Forest. Multiple pure technical indicators were constructed for stock prediction, and Grid Search was used to optimize the parameters of the Random Forest. Ultimately, through experiments on stock price movement prediction, it was found that under multiple pure technical indicators, the optimized Random Forest model for stock prediction exhibits greater reliability and can provide users with effective investment decision-making support.
The advantages of the Random Forest model primarily include:
The ability to perform explanatory analysis through feature importance ranking.
Robustness to missing values and noisy data.
However, its limitation lies in the linear increase in computational complexity with the number of trees.
Future research directions could involve integrating deep learning algorithms for automatic feature extraction [10] and utilizing incremental learning models to process streaming data.

References

[1] John Doe. The Role of Nutrition in Maintaining a Healthy Lifestyle."Journal of Nutrition and Health, 2022, 50(2), 123-124.
[2] American Diabetes Association. (2023). Nutrition Therapy Recommendations for the Management of Adults With Diabetes.

Diabetes Care, 2023, 46(1), 105-123.
[3] Van der Ploeg HP, Thomas EL, Bartels M, et al. Personalized Nutrition for the Management of Type 2 Diabetes: A Review of the Evidence. Nutrients, 2021, 13(6): 1951. https://doi.org/10.3390/nu13061951
[4] Haoxuan Li, Xueyan Zhang, Ziyan Li, et al. Overview of Machine Learning for Stock Selection Based on Multi-Factor Models. E3s Web of Conferences, 2020:214.
[5] Yihua Zhong, Lan Luo, Xinyi Wang, et al. Multi-factor Stock Selection Model Based on Machine Learning. Engineering Letters, 2021, 29 (1):20-26.
[6] Domitr P, Włostowski M. The use of machine learning for inverse uncertainty quantification in TRACE code based on Marviken experiment. Nuclear Engineering and Design, 2021, 384: 111498.
[7] Domitr P, Włostowski M, LASKOWSKI R, et al. Comparison of inverse uncertainty quantification methods for critical flow test. Energy, 2023, 263:125640.
[8] Carbon stock variability of Setiu Lagoon mangroves and its relation to the environmental parameters. Mohamad Saiful Imran Sahari; Nadiatul Azimah Mohd Razali; Nurul Shahida Redzuan; Amri Md Shah; Nor Aslinda Awang; Lee Hin Lee; Hafizan Juahir; Siti Mariam Muhammad Nor. Global Ecology and Conservation.2024
[9] Nijman S, Leeuwenberg A M, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol, 2022, 142: 218-229.
[10]High-resolution mapping of forest structure and carbon stock using multi-source remote sensing data in Japan. Hantao Li; Takuya Hiroshima; Xiaoxuan Li; Masato Hayashi; Tomomichi Kato. Remote Sensing of Environment.2024.

