Research on Stock Price Prediction Based on Machine Learning

Shijin Yang[1]

[1] Sun Yat-sen University, Zhuhai, Guangdong Province, 519000, China

yangshj28@mail2.sysu.edu.cn

Abstract. Stock forecasting has always been a popular research topic in the economic market. In current years, machine learning methods have been generally used in stock forecasting, which has unique advantages over traditional stock forecasting methods. This paper collects the relevant literature in recent years, first introduces the traditional methods of stock prediction, quantitative analysis, fundamental analysis and technical analysis, and then introduces modern stock analysis methods: deep learning, support vector machines and random forests, introduces various improved prediction models, as well as the practical application and research status of different models, and summarizes the application characteristics of these models. The article summarized the advantages and disadvantages of machine learning algorithms at the end, and provided a outlook on the development of machine learning algorithms based on practical situations.

Keywords: Stock Prediction, Support Vector Machine, Deep Learning, Random Forest.

1 Introduction

The stock market has a history of more than 400 years, starting in the 17th century. Stock trading activities have gradually become an important channel for people to invest and make profits, so predicting the trend of stock prices has become the focus of various researchers, hoping to reveal the trend of stocks through certain research projects. But due to the complexity, volatility, randomness, and uncertainty of stocks, predicting stock movements becomes very difficult [1].

After a long period of observation and research, various researchers have summarized many methods for predicting stocks, and applied them to stock market analysis. Early stock analysis techniques mainly include fundamental analysis, quantitative analysis, and technical analysis. Researchers need to spend a lot of time and effort on research and analysis. These methods are also affected by factors such as information asymmetry, market sentiment, researchers' own knowledge, and understanding of the stock market.

Although these traditional stock analysis methods are still applicable to a certain extent, with the development of information technology, more advanced technical

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_72

tools have emerged in modern stock analysis methods, such as deep learning, support vector machines, random forests, and other methods. These methods offer a number of advantages over traditional analytical techniques, such as the ability of modern stock analysis methods to quickly process large amounts of data and provide immediate results; Increasingly complex data sets can be analyzed and processed to provide deeper market insights, among other things, compared to manual ones. These features and advantages provide researchers and investors with more comprehensive and accurate market information and forecasting capabilities.

This paper centers on the literature on the utilization of various forecasting methods in stock prediction in current years, summarizes the stock prediction methods used in them, and classifies them by time and type of method.

2 Traditional Prediction Methods

2.1 Fundamental Analysis and Technical Analysis

In the stock market, fundamental analysis and technical analysis are two common analysis methods. Fundamental analysis, also known as basic analysis, refers to the method of predicting stock prices by analyzing a company's financial statements, management team, industry position, and profitability [2]. Investors need to use knowledge of related disciplines such as finance, financial management, economics, and securities investment, and pay attention to environmental factors such as industry trends, situation policies, etc. to predict stock price changes.

Unlike fundamental analysis, technical analysis is a method of predicting the future price of a stock based on the statistical data of the stock itself. Specific technical indicators include historical opening prices, historical closing prices, moving averages and trading volumes, the market price of the stock, changes in trading volume, and rise and fall indexes [3].

2.2 Quantitative Analysis

Quantitative analysis is a method of analyzing and predicting financial markets using tools such as mathematics, statistics, and computer programming, and can convert ambiguous or ambiguous factors into concrete data for analysis and comparison. Typically, quantitative analysis uses historical data and mathematical models for backtesting and validation to find viable investment strategies. Specifically, there are the following steps: selecting the quantization object, determining the quantization range, setting the quantization level, establishing the quantization model and implementing the quantization calculation.

3 Machine Learning Algorithm

3.1 Deep Learning

Due to the rise of big data technologies and the advancement of computer computing power, deep learning technologies represented by Convolutional Neural

Network(CNN), Recurrent Neural Network(RNN), Long Short-Term Memory Network(LSTM), etc. have begun to be used by many scholars for stock price prediction.

Guan Xueying used the Autoregressive Integrated Moving Average-Recurrent Neural Network (ARIMA-RNN) hybrid model to predict stock prices [4]. The ARIMA model, as a differential autoregressive moving average model, is particularly suitable for capturing linear trends and seasonal patterns, which flattens non-stationary series by differentially and then makes predictions using autoregressive and moving average terms. He demonstrated that the accuracy of the hybrid model was higher than that of a single recurrent neural network model and concluded that the hybrid model was more effective in short-term dynamic and static predictions.

Gao Yuan and Huang Wei used the convolutional neural network (CNN) model, the long short-term memory neural network (LSTM) model and their combined CNN-LSTM model for stock price prediction [5]. When CNN and LSTM are combined, the CNN extracts the high-level features of the data, and then the LSTM uses these features to model the time series. This combination allows the model to utilize both spatial (i.e., high-dimensional features) and temporal (i.e., sequence dynamics) information, thereby improving prediction accuracy. Through experimental demonstration, the CNN-LSTM combined model has better prediction performance than a single model, verifying the effectiveness of the combined model.

The RNN-CNN architecture model proposed by Guan Jian combines the time series modeling capabilities of RNN and the local feature detection capabilities of CNNs, which can more comprehensively capture the short-term and long-term dependence of stock data, and outperform existing models in multiple evaluation indicators [6]. Multi-source feature fusion solves the problem of incomplete feature extraction, and the model hyperparameters optimized by ISPSO further enhance the accuracy of prediction.

3.2 Support Vector Machines

Support Vector Machine (SVM) is mainly used for classification problems. Its core principle is to find an optimal hyperplane in the feature space, which can maximize the interval between data points of diverse categories, so as to achieve a good classification effect. In stock prediction, it mainly predicts the rise and fall of stock prices by analyzing historical data. SVM can process a large number of features and find the key factors that determine stock price changes.

Zhaotong Li constructed a basic support vector machine model and constructed a Support Vector Machine - Radial Basis Function(SVM-RBF) model using the Radial Basis Function (RBF) kernel, and evaluated the effects of the two models [7]. SVM implements classification by constructing a maximum margin hyperplane. A basic SVM model may behave mediocre when dealing with nonlinear relationships.The SVM_RBF model is a variant of SVM that uses RBF kernel functions to deal with nonlinear problems. The RBF core function can map the low-dimensional feature space to the high-dimensional space, making the data that is difficult to separate easily distinguish in the high-dimensional space. In the stock prediction example, the

SVM_RBF model has a prediction precision of 0.54 and a recall ratio of 0.58, showing better performance than the basic SVM model. Experiments show that compared with the basic SVM model, the SVM_RBF model has improved in dealing with nonlinear relationships, but there is still space for progress in the overall prediction ability.

The authors Zhang Xiaofang and Qian Rui used the support vector machine regression model and the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model to predict stock prices. Experiments show that both SVM and GARCH models have better performance in stock price prediction [8].

3.3 Random Forest(RF)

RF is an ensemble learning algorithm that builds multiple decision trees and integrates their results to improve the accuracy and stability of predictions. Random Forest can be used for classification and regression tasks and is one of the most popular and powerful models in machine learning.

Zou Jie and Li Lu proposed a Random Forest guided Self-Attention Bidirectional Gated Recurrent Unit(RF-SA-BiGRU) model based on random forest for stock price prediction [9]. This method is used in the fusion of Self-Attention mechanism (SA) and Bidirectional Gated Recurrent Unit (BiGRU)network, based on the Self-Attention Bidirectional Gated Recurrent Unit(SA-BiGRU) model, the dimensionality reduction processing technology random forest (RF) is introduced. Through comparative experiments, the RF-SA-BiGRU model is better than other models such as Gated Recurrent Unit(GRU), BiGRU, RF-BiGRU and CNN-SA-BiGRU in the field of prediction accuracy and stability.

Wang Huiying and Hao Yongtao proposed a Grid Search Random Forest(GS-RF) model based on technical indicators and random forest stock price trend prediction algorithm [10]. The model combines technical indicators and a random forest algorithm to predict stock price movements. The parameters of the random forest were optimized by grid search, which improved the prediction accuracy and generalization ability. Experimental results show that compared with the unoptimized random forest model, the GS-RF model performs better in terms of prediction accuracy, out-of-pocket estimation accuracy and area under the Rate of Change(ROC) curve.

Deng Jing and Li Lu established a population prediction model based on random forest, and they selected pure technical indicators such as Relative Strength Index(RSI), ROC, On-Balance Volume(OBV), Moving Average Convergence Divergence(MACD), and Williams Percent Range(Williams% R) as the characteristics of stock forecasting [11]. The parameters of the random forest were optimized by the grid search method, and a random forest stock prediction model based on these technical indicators and optimization parameters was constructed. Experimental results show that the accuracy and AUC of the optimized random forest prediction model are higher than those of other models.

4 Discussion

Deep learning models (e.g., LSTMs, RNNs) are capable of capturing complex nonlinear relationships and time series features. But it requires a lot of historical data and high computing resources, which is suitable for large data sets. It also requires high computing resources for training.

The advantage of random forests is that they are robust to noise and outliers, so they are suitable for small and medium-sized datasets, especially when there are many data features and noise. It is also relatively easy to interpret model output and feature importance and is suitable for small to medium-sized datasets, especially if the data features are numerous and noisy are present.

Support vector machines perform well on small sample datasets and are not easy to overfit. However, it requires careful selection of kernel functions and parameters, and model tuning is complex. It is suitable for small-scale datasets and high-dimensional feature data, especially if there are complex relationships between data features.

Overall, deep learning is suitable for processing large-scale, complex time series data with the ability to capture long-term dependencies. Random forests are suitable for small and medium-scale datasets, and have strong robustness and interpretability. Support vector machines are suitable for small-scale, high-dimensional feature data, and can find optimal solutions in complex relationships.

5 Conclusion

Stock price prediction is an important topic, and researchers are committed to proposing near-perfect prediction models. However, due to the randomness and uncertainty of the stock market, this goal is difficult to achieve, so the error of the forecasting model can only be reduced as much as possible. Each model has its own unique advantages and disadvantages, and scholars can summarize the profits and disprofits of each model, supplement the advantages and disadvantages, integrate the model, and improve the accuracy of the model's prediction. For example, a stacking method can be used to fuse deep learning models, random forest models, and support vector machine models to improve the accuracy of stock forecasts. The principle of the stacking method is to input the prediction results of multiple models into a meta-model as new features for final prediction. It is suitable for complex forecasting tasks and can handle both classification and regression problems. Relevant scholars can follow these ideas to fuse the model and improve the prediction accuracy.

References

1. Yan,W.X.: Application of machine learning in stock prediction [J]. Information Systems Engineering, (04):40-43(2024).
2. Hu,X.M.: On the application of fundamental analysis and technical analysis in the stock market[J]. New West (Theoretical Edition),(05):62+64(2014).
3. Li,X.J., Tang,P.: Stock price prediction based on technical analysis, fundamental analysis

and deep learning[J]. Statistics and Decision, 38(02): 146-150. DOI: 10.13546/j.cnki.tjyjc.2022.02.029(2022).

4. Guan,X.Y.: Stock price prediction based on ARIMA-RNN hybrid model[J]. Journal of Harbin University of Commerce (Natural Science Edition), 40(02): 250-256. DOI: 10.19492/j.cnki.1672-0946.2024.02.008(2024).
5. Gao,Y., Huang,W.: Stock price index prediction based on deep learning[J]. Software Engineering, 27(05): 7-13. DOI: 10.19644/j.cnki.issn2096-1472.2024.005.002(2024).
6. Guan,J.: Research on stock price prediction method based on RNN-CNN model[D]. Nanjing University of Information Science and Technology, 2024.DOI:10.27248/d.cnki.gnjqc.2023.000448(2024).
7. Li,Z.T.: Development history and application of support vector machine[J]. Information Systems Engineering, 2024(03):124-126.
8. Zhang,X.F., Qian,R.: Research on stock price prediction based on support vector machine[J]. Journal of Luoyang Normal University, 41(05): 22-26. DOI: 10.16594/j.cnki.41-1302/g4.2022.05.013(2022).
9. Zou,J., Li,L.: Research on stock price prediction based on SA-BiGRU model of random forest[J]. China Prices, (11):52-56(2023).
10. Wang,H.Y., Hao,Y.T.: Stock price trend prediction algorithm based on technical indicators and random forest[J]. Modern Computer, 27(27): 43-47+52(2021).
11. Deng,J., Li,L.: Application of parameter optimization random forest in stock prediction[J]. Software, 41(01): 178-182(2020).