Research on Strategy of HS300 Index Based on Random Forest

Xinrun Huang
Computer Science and Technology
Yanshan University
Qinhuangdao, China
huangxinrun100@163.com

***Abstract*—With the development of China's economy and the continuous improvement of the stock market, stock investment has gradually become one of the ways for people in financial management. It also attracts non-mathematicians to participate in the prediction of stock prices. Models such as machine learning in artificial intelligence are beginning to be widely used in stock forecasting to predict stock trends. This paper calculated the 11 technical indicators of the HS300 index, and used the Random Forest model to predict the movements. The results showed that the accuracy of the model calculation to 0.82 and the F-score is 0.84. Then, the daily income of the model was integrated. Then annualized return reached 28.1%, and the maximum drawdown was 0.42, which was significantly higher than the index return. The results showed that the Random Forest model adopted in this paper is feasible. With the introduction of more and more machine learning methods, artificial intelligence will be further developed in the field of quantitative investment.**

Keywords—HS300 index, random forest, model evaluation, prediction of return

I. INTRODUCTION

Stock is a kind of securities by which the holder obtains dividends [1-2]. Stockholders have two types of investment income. One is to distribute dividends in the company's earnings according to the amounts of shares; and the other is to gain income in the change of stock prices. Since China established the stock market in the early 1990s, after more than 20 years of market economy development and reform, the stock market has gone from the previous phase of rectification to the maturity stage. The improvement of the securities market system and the thriving stock market have attracted the attention of investors from all walks of life and stimulated their investment interest. Investors gain profits in stock price changes by selling the transferred shares at any time. Therefore, stock investment has gradually become one of the ways people invest in wealth management.

As an investor, how to correctly predict the trend of stock prices is their most concerned topic. There are many factors affecting stocks, such as companies' operating conditions, national policies, revenue rates, etc [3]. An efficient and accurate stock forecast is one of the hottest research areas today. Traditional stock forecasting mainly includes basic information based on macro-micro economy, company finance, and establishing mathematical models to analyze historical data to evaluate stocks and determine the future direction of stocks. The traditional stock forecasting method has great limitations, and the forecasting accuracy will be reduced due to the inequality, delay of the company's information, national interest rate factors and policy factors. With the continuous development of artificial intelligence technology, the field of artificial intelligence technology involves medical treatment, education and other aspects [4]. And artificial intelligence can help people invest in and manage assets in the financial sector. The artificial intelligence-based stock forecasting system uses machine learning methods to construct algorithms and models and predicts stock market price trends. The system is highly self-tuning and self-learning, with good fault tolerance and better prediction than traditional prediction methods.

The stock forecasting based on machine learning has a good forecasting effect. The research results of stock forecasting models include the following. In 1986, Rumelhart and McClelland proposed the BP neural network concept, which was a multilayer feedforward neural network trained according to the error back propagation algorithm [5]. BP neural network has strong nonlinear mapping ability and flexible network structure [6]. Kimoto et al. used the stock data of the Tokyo Stock Exchange to construct a neural network-based TOPIX stock forecasting system for prediction. The prediction results showed that the neural network model has a higher rate of return [7]. Guresen et al. used multi-layer contrast-aware (MLP), dynamic artificial neural networks and hybrid neural networks to predict Nasdaq stock data, which proved that hybrid neural networks have higher precision [8]. Fu et al. constructed 244 technical and fundamental indicators and proposed artificial intelligence algorithm framework to distinguish good stocks from Chinese stock market [9]. The results demonstrated that stacking method outperformed other machine learning method, such as Logistic Regression (LR) Deep Neural Network (DNN) and so on. Leung et al. presented a machine learning approach for business intelligence applications [10]. It applied SSVMs (structural support vector machines) to perform classification on sophisticated inputs such as the nodes of graph structures. The experimental results indicated the practicality of the machine learning approach in stock prices prediction. Jones et al. pointed out that a new method named Prototype Ranking (PR) in stock selection [11]. PR applied a modified learning technique to predict the ranks of stocks and perform the learning and testing in noisy stocks samples. The results showed that PR method brings obvious profit improvement. Huang et al. proposed a methodology using support vector regression (SVR) and genetic algorithms (GAs) for effective stock selection [12]. It first used the SVR method to generate alternates for actual stock. The GA is used for optimization of model parameters on top of this model. The results showed the investment returns of this model outperform the benchmark. Kong et al. investigated the approach of adaptive input selection (AIS) in stock prediction [13]. The learning models used artificial neural network

978-1-7281-7106-7/20/$31.00 ©2020 IEEE
DOI 10.1109/CDS49703.2020.00038

(ANN) and support vector machine (SVM). The investment performance evaluation demonstrated that the AIS approach offer significantly higher returns than the DIS approach. In this paper, Random Forest will be used to predict the movements of HS300 with historical data.

II. Data Reaserch

The data of the HS300 index selected in this paper was derived from the Wind Financial Terminal. The constituents of the HS300 Index have good market representation for each constituent stock. The backtracking interval is from January 1, 2000 to January 1, 2018. On the last trading day of each day, the MACD, MA, CCI, CMO, RSI, PSY, BIAS indicators of each stock were taken as the characteristics of the sample; the movement of each stock in the next day was calculated as the label of the sample. The HS300 index reflects the operational profile of stock price movements in China's securities market and serves as an evaluation criterion for investment performance, providing a basis for continued innovation in indexed investment and index-derived [14]. The sample of the HS300 index covers about 60% of the market value of the Shanghai and Shenzhen stock markets, and has a good market representation, reflecting the overall trend of the stock market.

The calculation formula for the HS300 index is as follows:

$$\text{Index}=\frac{N}{M}\cdot 1000 \tag{1}$$

where N is the adjusted market value of the constituents of the reporting period and M is the adjusted market value of the constituents of the base. M is calculated by

$$M=\sum(a\cdot b) \tag{2}$$

where a is the market price and b is the adjusted number of shares. We can adjust the number of shares to adjust the constituent stocks by means of grading. The sample selection method for the HS300 Index sample is to rank the average daily turnover of the sample space stocks from high to low in the most recent year. Therefore, the HS300 Index includes leading enterprises in major industries, reflecting the development of the stock market and market economy. The 300 constituent stocks cover leading companies in dozens of major industries such as steel, oil, coal and real estate.

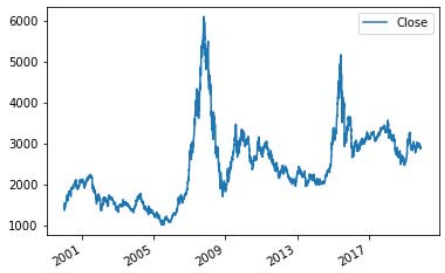

Fig. 1. Changes in the HS300 Index in the past two decades

We have made a macro analysis of the main bull markets in the stock market in the past two decades. The first was from June 2005 to October 2007, the market rose from 998.32 points to 6124.04 points, an increase of 513.5%. 6124 was the highest point in the history of China's stock market, and market sentiment was high at that time. The main reasons were that the country was in a period of rapid development, and the RMB has appreciated rapidly. The economy has recovered, and the share-trading reform has been implemented. From October 2007 to September 2008, due to the impact of the international financial crisis and domestic currency pressure, the broad market index fell by 73%. From October 2008 to August 2009, the stock market and the country's 4 trillion investment plan, the stock market has seen a big rebound. Finally, from July 2014 to June 2015, the market moved from 2081 points to 5178.19 points, an increase of 148%. The main factor was the "the Belt and Road" policy, the interest rate cut policy and a large amount of capital flowing into the stock market.

III. Models

A. Dcision Tree

Decision tree is common machine learning algorithms based on tree structure for decision making [17]. Decision trees are implemented in ID3, C4.5 and CART. The problem we have to solve is the classification problem, so we use the classification tree method. The decision tree generation process has three main parts: feature selection, decision tree generation, and decision tree pruning [18]. First, a feature is selected from the many feature features in the training set as the node classification criterion. The child node is then recursively generated from the root node according to the classification criteria until the split reaches the leaf node. To ease the over-fitting of the decision tree, the decision tree is finally pruned.

The three criteria for feature selection are information gain, information gain rate, and Gini index. In the ID3 algorithm, the information gain is evaluated according to the information theory. The C4.5 algorithm uses the information gain rate to select features. The CART algorithm uses the Gini index [19] (the feature with the smallest Gini index) as the splitting criterion. This paper selects the information gain as the evaluation criteria. The concept of entropy is introduced, and entropy represents the uncertainty of measuring random variables. Supposed the value of the random variable X is x1, x2, ..., xn, for each value xi, the entropy of the random variable X is:

$$H(X)=-\sum_{i=1}^{n}p_i log_2 p_i \tag{3}$$

Then the entropy of the sample set D is:

$$H(D)=-\sum_{i=1}^{n}\frac{|C_k|}{|D|}log_2\frac{|C_k|}{|D|} \tag{4}$$

where |Ck | is the number of samples of category k, and |D| is the total number of samples. The information gain represents the difference between the entropy before and after the data set is divided by a certain feature value. The calculation formula is:

$$g(D,A)=H(D)-H(D|A) \tag{5}$$

where H(D) is the entropy of the data set and H(D|A) is the entropy after the data set characterized by A.

There are many advantages to using the decision tree model in the classification problem. The decision tree model is simple and easy to implement for building stock forecasting, and has a high accuracy. For data on unrelated features in stock samples, it is easy to construct rules that are easy to understand and that are accurate.

B. Random Forest

Random Forest is an algorithm that integrates multiple trees through the idea of integrated learning. Its basic unit is the Decision Tree [20]. Integrated learning is to train a number of weak learners, and finally form a strong learner through a certain combination strategy. The Bagging integrated learning method is based on the self-sampling method. This method divides the training set into N new training sets, and then builds a model on each new training set, which is irrelevant. In the final prediction, we will result in the N models. Schematic diagram of Random Forest algorithm is shown in Figure 2. On the basis of building a bagging integration based on Decision Tree, the Random Forest further introduces the selection of random attributes in the training process.

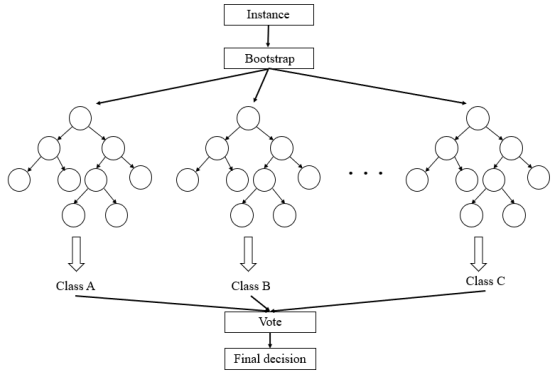

Fig. 2. Schematic diagram of Random Forest algorithm

Random Forest is a flexible and practical algorithm, which has many advantages: In all models for stock forecasting, the algorithm is simpler than neural network and has better accuracy. It has strong performance for dealing with a large number of eigenvalues existing in the stock forecasting system, and the default problem of the feature can also ensure higher accuracy.

C. Evaluation

The evaluation of a classification model is mainly based on accuracy, precision, recall and F-score indicators. These indicators are described below. The samples are divided into real categories and classifier prediction categories according to the real category. The true positive case (TP) means that if the real category is positive, the prediction category is positive; the false positive (FP) means that if the real category is negative, the prediction category is Positive class; false negative case (FN) means that if the real category is positive, the prediction category is negative; the true negative case (TN) means that if the real category is negative, and the prediction category is negative.

The accuracy indicates the proportion of all samples that are classified correctly. The calculation formula is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

The precision measures the classification accuracy of positive samples, and the calculation formula is:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

The recall indicates the proportion of positive examples in the sample that are correctly predicted. The calculation formula is:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F-score is the harmonic average of the recall rate and accuracy. The formula is:

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + TN} \tag{9}$$

IV. Strategy

This paper used a quantitative method to buy and sell HS300 index. The specific steps are as follows: First, all features described by previous technical indicators were pre-processed. Then data set were divided into training sets and test sets by the rate of 8:2. The sample was used as a positive example if the movement of HS300 index rose. The training sets were performed in Random Forest. After the model training is completed, all the pre-processed features of the last trading day sample were taken as the input of the model, and the predicted value y of the sample feature is obtained. Finally, the results of the backtest were compared with the actual sample data for model evaluation. For the classification model, the model was evaluated by comparing the accuracy, precision, recall and F-score of the training set. Strategy diagram is shown in Figure 3.

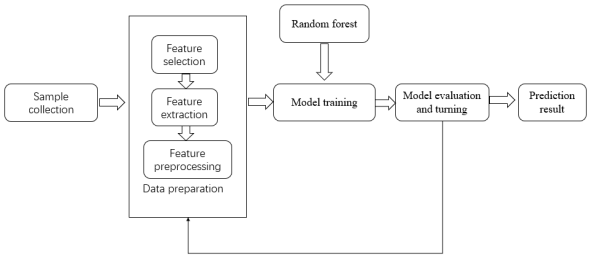

Fig. 3. Strategy diagram

V. Results and Discussion

We constructed a Random Forest model for movement prediction of HS300 index. The accuracy of the model calculation to 0.82. The recall rate is 0.85 and the F-score is 0.84. The final return on the portfolio and the yield of the HS 300 Index are shown in Figure 4. It can be seen that the return of the strategy we constructed was significantly better than

the HS300 Index. The annualized return reached 28.1% and the maximum retreat rate was 0.47. This shows that our investment strategy is basically successful.

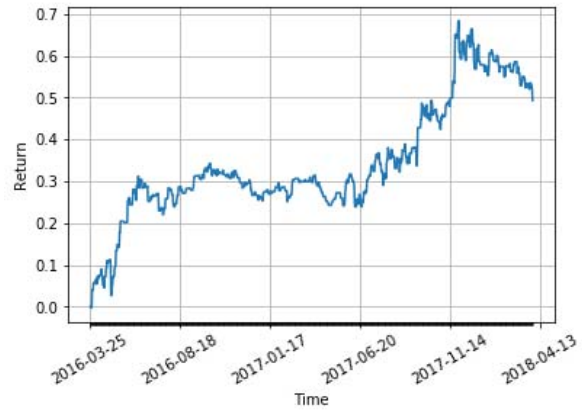

Fig. 4. Comparison of earnings and strategy results of the HS300 Index

Figure 5 presented the importance of different indicators for predicting HS300 Index. The results showed BIAS and PSY were more significant than other technical indicators. RSI and AROON had little effect on the movement prediction of HS300 index.

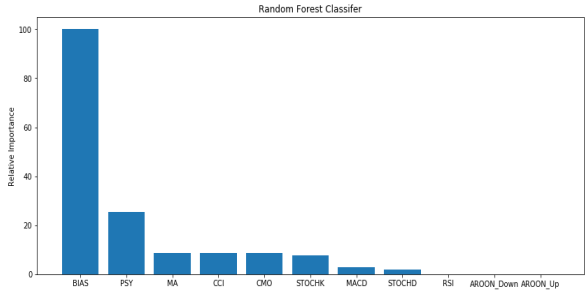

Fig. 5. Importance of different technical indicators for predicting HS300 Index

VI. Conclusion

This paper used the Random Forest model to predict the index movement through 11 indicators of China's stock market. Based on the daily income of the model training, we selected the date to buy and sell the index. Finally, the annualized return and the maximum drawdown of the portfolio to quantify the prediction of index movement measurement. The results showed that the Random Forest model adopted in this paper is feasible. We proposed a framework for using stocks to conduct stock trading. The use of Random Forest models to select stocks works well, but not good enough. The next step can be improved by the following methods: it can be extended in the diversity of indicators. By expanding the time frame and indicators, or looking for other more effective indicators, we can discover the deep changes in the stock market and improve the accuracy.

References

[1] Kang J, Liu M H, Ni S X. Contrarian and momentum strategies in the China stock market: 1993–2000[J]. Pacific-Basin Finance Journal, 2002, 10(3):243-265.

[2] Guoping. China's Stock Market : Inefficiencies and Institutional Implications[J]. China & World Economy, 2010, 16(6):81-96.

[3] Narayan P K, Zheng X. Market liquidity risk factor and financial market anomalies: Evidence from the Chinese stock market[J]. Pacific-Basin Finance Journal, 2010, 18(5):509-520.

[4] Jobling C P . Application of Artificial Intelligence in Process Control[M]// Application of Artificial Intelligence in Process Control: Lecture Notes Erasmus Intensive Course. Elsevier Science Inc. 1993.

[5] Rumelhart D E, Mcclelland J L. Parallel Distributed Processing[M]// Parallel distributed processing. 1986.

[6] White H. Economic Perdication Using Neural Networks: The Case of IBM Daily Stock Returns[J]. Earth Surface Processes & Landforms, 1988, 8(5):409-422.

[7] T. Kimoto, K. Asakawa, M. Yoda, et al. Stock Market Predietion System with Modular Neural Networks[C]// Neural Networks, 1990 IJCNN, Intemational Joint Conference on, 1990, 6(4):365-374.

[8] E.Guresen,G.Kayakutlu,,T.U.Daim.Using artificial Neural Network Models In Stock Market Index Prediction[J]. Expert Systems with Applications, 2011, 38(8): 10389-10397.

[9] Rasekhschaffe K, Jones R. Machine Learning for Stock Selection[J].

[10] Leung K S , Mackinnon R K , Wang Y . A machine learning approach for stock price prediction[C]// IDEAS 2014. 2014.

[11] Rasekhschaffe K C , Jones R C . Machine Learning for Stock Selection[J].

[12] Huang C F . A hybrid stock selection model using genetic algorithms and support vector regression[J]. Applied Soft Computing, 2012, 12(2):807-818.

[13] Kong A , Zhu H . Predicting Trend of High Frequency CSI 300 Index Using Adaptive Input Selection and Machine Learning Techniques[J]. Nephron Clinical Practice, 2018, 6(2):120-133.

[14] Wang Y , Xu Y , Liu P . The Co-integration and Causality Relationship Research of Stock Index Futures IF1006 and HS300 Stock Index[M]// Education and Management. Springer Berlin Heidelberg, 2011.

[15] Pasiphol A. Forecasting stock index direction : comparison of MACD and RSI, case study on SET 50 index[J]. Faculty of Commerce & Accountancy Thammasat University, 2009.

[16] Rosillo R , De l F D , Brugos J A L . Technical analysis and the Spanish stock exchange: testing the RSI, MACD, momentum and stochastic rules using Spanish market companies[J]. Applied Economics, 2013, 45(12):1541-1550.

[17] Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1):81-106.

[18] Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems Man & Cybernetics, 2002, 21(3):660-674.

[19] Gastwirth J L. The Estimation of the Lorenz Curve and Gini Index[J]. Review of Economics & Statistics, 1972, 54(3):306-316.

[20] Archer K J, Kimes R V. Empirical characterization of Random Forest variable importance measures[M]. 2008.