Applied Economics and Finance
Vol. 4, No. 4; July 2017
ISSN 2332-7294 E-ISSN 2332-7308
Published by Redfame Publishing
URL: http://aef.redfame.com

Construction of Quantitative Transaction Strategy Based on LASSO and Neural Network

Xu Wang[1], Jia-Yu Zhong[2] , Zi-Yu Li[1]

[1]Finance Department of International Business School, Jinan University, Zhuhai, China

[2]Electrical Engineering and Automation of Electrical Engineering Information School, Jinan University, Zhuhai, China

Correspondence: Zi-Yu Li, Finance Department of International Business School, Jinan University, Qianshan Road 206#, Zhuhai City, Guangdong Province, Post No.519070, China.

Received: April 19, 2017 Accepted: May 9, 2017 Available online: June 21, 2017

doi:10.11114/aef.v4i4.2370 URL: https://doi.org/10.11114/aef.v4i4.2370

Abstract

Since the establishment of the securities market, there has been a continuous search for the prediction of stock price trend. Based on the forecasting characteristics of stock index futures, this paper combines the variable selection in the statistical field and the machine learning to construct an effective quantitative trading strategy. Firstly, the LASSO algorithm is used to filter a large number of technical indexes to obtain reasonable and effective technical indicators. Then, the indicators are used as input variables, and the average expected return rate is predicted by neural network. Finally, based on the forecasting results, a reasonable quantitative trading strategy is constructed. We take the CSI 300 stock index futures trading data for empirical research. The results show that the variables selected by LASSO are effective and the introduction of LASSO can improve the generalization ability of neural network. At the same time, the quantitative trading strategy based on LASSO algorithm and neural network can achieve good effect and robustness at different times.

Keywords: quantitative transaction strategy, LASSO, neural network

1. Introduction

Since the establishment of the securities market, how to accurately predict the future trend of the securities market price has been the focus of many scholars. The earliest method is the securities investment analysis method based on fundamental analysis and technical analysis. But this method can only be used for qualitative analysis of securities market. It is impossible to analyze the complex statistical characteristics of the data in the financial market and can not grasp the intrinsic relationship between these data. After that, the time series analysis method began to rise. It uses mathematical tools to establish the financial model to analyze the time series data. In 1952, Markowitz proposed the theory of portfolio. On this basis, Sharpe (1964) proposed the Capital Asset Pricing Model (CAPM), which established a linear relationship between individual asset returns and market returns. Ross (1976) extended the CAPM and proposed the arbitrage pricing theory (APT), which established the linear relationship between the asset returns and multiple factors. Although these models are successful in theory, they are not ideal in practical applications because they are based on the assumption of efficient market theory. Then, on the basis of predecessors, modern time series analysis uses probabilistic method to analyze the securities market. The ARMA model proposed by Jenkins and Box(1994) and the GARCH model proposed by Bollersllev (1986) have become the classic methods of characterization of securities market price fluctuation. In 2011, Breidt developed the All-Pass model on the basis of the ARMA model, and the model achieved good results when fitting the financial time series. However, time series analysis method has its limitations, mainly in its usual basis based on some stringent assumptions, such as the efficient market hypothesis, the assumption of normal distribution, linear hypothesis. And the time series data of the financial market usually has high noise and obvious characteristics of non-stationary, nonlinear, which are inconsistent with the hypothesis of time series analysis.

With the development of science and technology, quantitative trading, which is based on the combination of computer technology and financial engineering, has begun to rise. Kestner, L. (2003) introduced the development of quantitative trading strategies in his book and validated the rationality of the latest quantitative trading strategy using quantitative analysis. Stefano Fiorenzani (2006) systematically expounded the risk management of quantitative transactions in his article and demonstrated the risk control of quantitative transaction through advanced mathematical and statistical methods. To a certain extent, quantitative trading overcomes the weakness of human nature, and makes people more rational. More and more scholars believe that quantitative trading will become the most important way to deal with the financial market in the future.

In recent years, machine learning has achieved great success, and has been applied to the analysis and prediction of the stock market data. The most widely used is the neural network. White (1989) first predicted the use of artificial neural networks in the financial sector, but his prediction of IBM's stock returns on the use of artificial neural networks was not as good as expected. He thought it was mainly due to artificial neural network into a local minimum result. Schurmann (2000) and so on used the traditional statistical methods and artificial neural network to predict the stock market, the results found that the performance of neural networks is better than the traditional mathematical statistics forecasting method. For China's securities market, domestic scholars have established some quantitative trading models suitable for China's national conditions. According to the classification ability of BP neural network, Wei Wu (2001) tries to forecast the Shanghai Stock index. Through a large number of numerical analysis and optimization of the model, the prediction accuracy of the model reaches 70%. Jianfu Cui (2004) separately established the GARCH model and the BP neural network model to forecast the stock price. He found that the neural network model as a nonlinear system showed a strong generalization ability for the time series data of the fluctuation distance of the stock price. The accuracy is slightly better than the GARCH model.

The first step in quantifying a transaction is to acquire useful information from a large number of data to improve the accuracy and rationality of the model. Robert Tibshirani (1996) proposed a linear model estimation method, LASSO algorithm. This method uses the absolute function of the model coefficients as a penalty to compress the model coefficients so that some of the regression coefficients become smaller and overcome the shortcomings of the traditional method in selecting variables. Then, Zou (2006), Meinshansen (2006) and others have improved LASSO. And this method has become a popular variable selection model in the field of statistics.

On the basis of former research, this paper combines machine learning with statistical modeling to construct an effective quantitative trading strategy. We take the CSI 300 stock index futures as the object of study, through the LASSO algorithm to select from a number of technical indicators to get the most effective indicators. Based on the screening index, the artificial neural network is used to construct the nonlinear forecasting model to forecast the price of stock index futures. After that, we further improve the model by optimizing the trading threshold and stop loss, so as to build a quantitative trading strategy with rationality, validity and stability. The structure of this paper is as follows: The first part is the introduction of the relevant areas of the development process and research results. The second part expounds the principle of LASSO algorithm and neural network model and the construction of quantitative trading strategy. The third part uses the CSI 300 stock index futures to test the model and analyze the results. The fourth part is the conclusion.

2. Quantitative Trading Strategy

Because of the weakness of traditional forecasting methods in the stock market, this paper attempts to establish a forecasting model by combining the statistical method and the machine learning method, and then builds the quantitative trading strategy. The construction of the strategy includes the following three steps: first, to reduce the degree of redundancy data, using statistical methods for high dimensional data screening. Then, the nonlinear prediction model is established by machine learning algorithm. Finally, we introduce the trading threshold and stop loss to build an effective quantitative trading strategy.

2.1 High Dimensional Data Selection: LASSO Algorithm

In the case of forecasting the future price trend of the stock market according to the input vector, there will inevitably be multiple collinearity problems, which will affect the accuracy of the model prediction. The common methods of coping with collinearity are: subset selection, principal component analysis, ridge regression and lasso. Then we analyze and compare these four methods.

The subset selection method, which is represented by stepwise regression, needs to be repeated several times. In addition, the stepwise regression method tends to retain some unimportant variables, so the probability of selecting the correct model is low. Principal component regression can deal with collinearity. But the principal component depends on all the original variables, it is difficult to give the actual interpretation of the model. Ridge regression can reduce the regression coefficient, but it does not make any one regression coefficient reduced to zero, making the model can not be interpreted well. While this method is subjective in the choice of parameters. LASSO has the advantages of ridge regression and subset selection, which makes it better than other methods to solve the problem of high-dimensional multicollinearity. Therefore, this paper selects LASSO to screen high dimensional data.

The basic idea of LASSO is to minimize the sum of squares of residuals under the constraint that the sum of the absolute values of the regression coefficients is less than a constant, which can produce some regression coefficients that are strictly equal to 0 to achieve the purpose of screening. Suppose there is a linear regression model:

$$y_i = \alpha_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \varepsilon_i \in N\left(0, \sigma^2\right) \tag{1}$$

Existing data: $\left(X^i, y_i\right), i = 1, 2, \cdots N$, Where $X^i = \left(x_{i1}, \cdots x_{ip}\right)^T$ and y_i are the independent and dependent variables corresponding to the i-th observed values respectively. In this regression model, the LASSO estimates for α_i and β_j are defined as:

$$\left(\hat{\alpha}，\ \hat{\beta}\right) = \arg\min_{\beta}\left\{\sum_i \left(y_i - \alpha_i - \sum_j \beta_j x_{ij}\right)^2\right\}, s.t. \sum_j \left|\beta_j\right| \le t \tag{2}$$

Where t is the penalty parameter that satisfies $t \ge 0$. If $t_0 = \sum_j \left|\beta_j\right|, t \le t_0$, at this time $\hat{\beta}_j^0$ is the least squares estimator for β_j . When $t < t_0$, part of the optimal solution is equal to 0, so the variable will be removed from the model, so as to achieve the purpose of variable selection.

The estimation method of the penalty parameter t in the LASSO algorithm is mainly generalized cross validation. In this method, we rewrite the constraint condition $\sum\left|\beta_j\right| \le t$ as $\sum \beta_j{}^2 / \left|\beta_j\right| \le t$, which is equivalent to adding a penalty in the regression squared sum: $\lambda \sum \beta_j{}^2 / \left|\beta_j\right|$. We can write the sum of the squared sums with constraints:

$$\sum\left(y_i - \sum \beta_j x_{ij}\right)^2 + \lambda \sum \beta_j{}^2 / \left|\beta_j\right| = \left\|y - X\beta\right\|^2 + \lambda \beta^T \begin{pmatrix} \left|\beta_1\right| & & \\ & \ddots & \\ & & \left|\beta_p\right| \end{pmatrix}^{-1} \beta$$

$$= y^T y - 2y^T X\beta + \beta^T X^T X \beta + \lambda \beta^T W^- \beta \tag{3}$$

Where $W = diag\left(\left|\beta_1\right|, \cdots, \left|\beta_p\right|\right)$, W^- denotes the generalized inverse of W . The above formula on β to solve the derivative to zero, we can get

$$\tilde{\beta} = \left(X^T X + \lambda W^-\right)^{-1} X^T Y \tag{4}$$

So the number of valid parameters in $\tilde{\beta}$ is approximately:

$$p(t) = tr\left\{X\left(X^T X + \lambda W^-\right)^{-1} X^T\right\} \tag{5}$$

Let $rss(t)$ be the sum of squares of regression under constraint t , so we can construct generalized cross validation statistics:

$$GCV(t) = \frac{1}{t}\frac{rss(t)}{\{1 - p(t)/N\}^2} \tag{6}$$

The optimal t should make the $GCV(t)$ minimum.

2.2 Construction of Nonlinear Predictive Model: BP Neural Network

In recent years, machine learning has developed rapidly and has been applied to the analysis and prediction of the stock market data. The most widely used is the neural network. BP neural network is an artificial intelligence method, which is a multilayer feedforward network based on BP algorithm. It has strong learning ability to approximate nonlinear continuous function, and it is successfully applied in the field of economic and financial forecasting. The BP algorithm is trained by the input and output data samples. The learning process includes the forward propagation process of information and the reverse propagation process of error. These two processes are used repeatedly to continuously calculate the changes in network weights and deviations in the direction of the gradient of the relative error function, and gradually approach the target. A typical BP neural network is made up of three parts: an input layer, at least one hidden layer and an output layer, as shown in Figure 1.

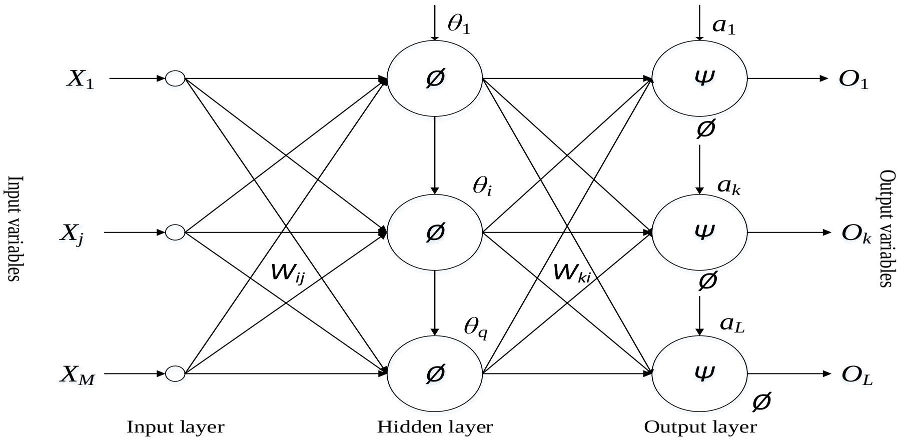

Figure 1. The Structure of BP Neural Network

In the process of forward propagation, information first enters the input layer, and then passes through the hidden layer to the output layer. If the output of the output layer does not match the expectation, the model enters the reverse propagation section, returns the error information and modifies the neuron weights between the layers. Finally make the error minimum.

2.3 The Construction of Quantitative Trading Strategy

2.3.1 Data Selection

This article selected the CSI 300 stock index futures month main contract from January 5, 2015 to December 22, 2016, a total of 482 trading days to study. Among them, from January 5, 2015 to August 19, 2016 a total of 400 sets of data for the training samples, August 22, 2016 to December 22, 2016, a total of 82 sets of data for the test samples. In addition, we also select the CSI 300 stock index futures 2007-2014 trading data to test the robustness of the model. The data source is the Cathay Pacific database.

(1) Selection of Dependent Variable

In order to effectively predict the future earnings situation, this paper does not select the future price as the dependent variable, but choose the average expected return rate in the next few days as the target value. Here we measure the expected return by the average price, the specific formula is:

$$y_m = \frac{P_{t+m} - C_t}{C_t} \tag{7}$$

Among them:

$$P_{t+m} = \frac{C_t + H_t + L_t}{3} \tag{8}$$

Where, H_t is the highest price on day t, L_t is the lowest price for day t, C_t is the closing price for day t. P_{t+m} represents the average price on day $t+m$. y_m is the price of the first day relative to the closing price of the day t.

In order to make the results more representative, we increase the total yield of the next k days to get the expected rate of return of the next k days:

$$Y_t = \sum_{m=1}^{k} y_m \tag{9}$$

If the value of Y_t is large, it means that the future rate of return is high, should be established long. If the value of Y_t is small or negative, indicating that the future rate of return is low, should be promptly closed or short.

(2) Selection of Independent Variables

According to the theory of "price, quantity, time and space" in the technical analysis theory, this paper selects different types of technical analysis indicators as input vectors to predict the future. In the stock market technical analysis process, technical analysis indicators refer to all through the mathematical formula calculated price data collection. At present, there are many technical analysis indicators in the stock market. In order to test the effectiveness of the forecasting model, this paper chooses 28 commonly used indicators from the five categories of technical analysis indicators according to the characteristics of stock index futures to carry out empirical research. In future studies, the indicators used can be expanded.

The technical indicators selected in this paper are shown in the Table 1.

Table 1. The technical indicators selected in this paper

Types	Technical analysis indicators
Trend indicator	ADXR, DEA, DIFF, MA (5), MA (10), MACD, MDI, PDI, ZIG.
Anti-trend indicator	BIAS (12), BIAS (6), CCI, KDJ.D, KDJ.J, KDJ.K, KST, QRSI, RSI1, RSI2, RSI3, WR1, WR2.
Swing indicator	ATR, TR.
Quantity price indicator	OBV.
Pressure support index	CHANNELS.HH, CHANNELS.LL.

2.3.2 Variable Selection

The information contained in these technical indicators of the input vector may be the same, so it is necessary to filter the input vectors. In this paper, LASSO algorithm is used to select variables. A penalty term is added to the computation of the RSS minimization:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| = RSS + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{10}$$

When the lambda is sufficiently large, some of the estimated coefficients can be accurately contracted to 0, so as to achieve the purpose of screening variables.

2.3.3 Training and Prediction

The neural network prediction model is constructed by using MATLAB. The input layer is the technical index variable selected by LASSO. The target variable is the expected future rate of return. Then, we will further improve the model by optimizing the trading threshold and stop loss, so as to build a quantitative trading strategy with rationality and stability.

3. Empirical Research

3.1 LASSO-ANNs

First, we use LASSO algorithm for variable selection. With above 28 technical indicators as independent variables, the average expected rate of return of K days as the dependent variable (This paper takes k=7). The Cross-validated MSE of LASSO fit is shown in the Figure 2.

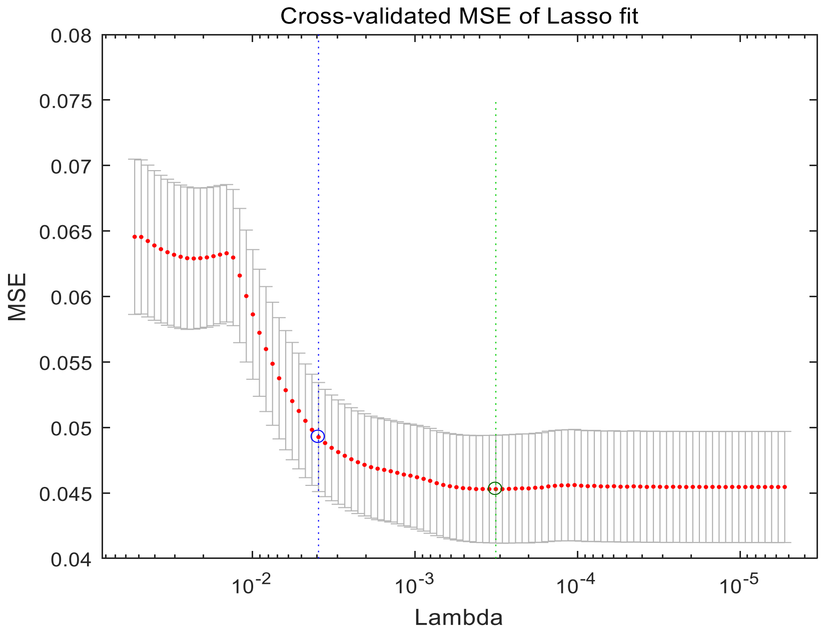

Figure 2. Cross-validated MSE of LASSO fit

Figure 2 shows the change of the number of variables in the model with the change of lambda value. As the value of λ becomes larger, the degree of model compression increases, and the function of model selection variable is enhanced. The ideal value of λ should be to maximize the degree of compression, that is, the number of variables obtained within a reasonable range as little as possible. As can be seen from the Figure 2, seven variables are sufficient to allow the MSE curve to return to gentle. LASSO finally selected seven indicators, namely RSI3, MA (10), BIAS (12), OBV, CHANNELS.LL, ZIG, and WR2.

Then the seven technical indicators will be selected as the input vector of the neural network, with the future rate of return as the target layer, training and forecasting. Figure 3 is the training results of first 400 groups of samples. It can be seen that after training, the expected rate of return is highly coincident with the true expected return. Figure 4 is the error rate histogram for sample 400 days. We can see that most of the error of the prediction results can be controlled within 7.5%, indicating that the training effect is very satisfactory. Therefore, it can be considered that the seven technical indexes selected by LASSO algorithm can well explain the average return rate in the next 7 days. The LASSO algorithm can effectively filter the variables.

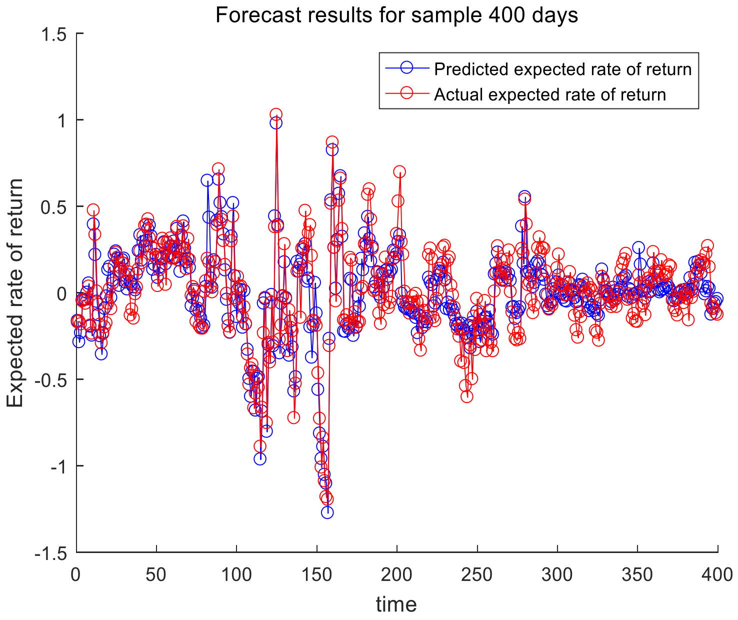

Figure 3. Forecast results for sample 400 days

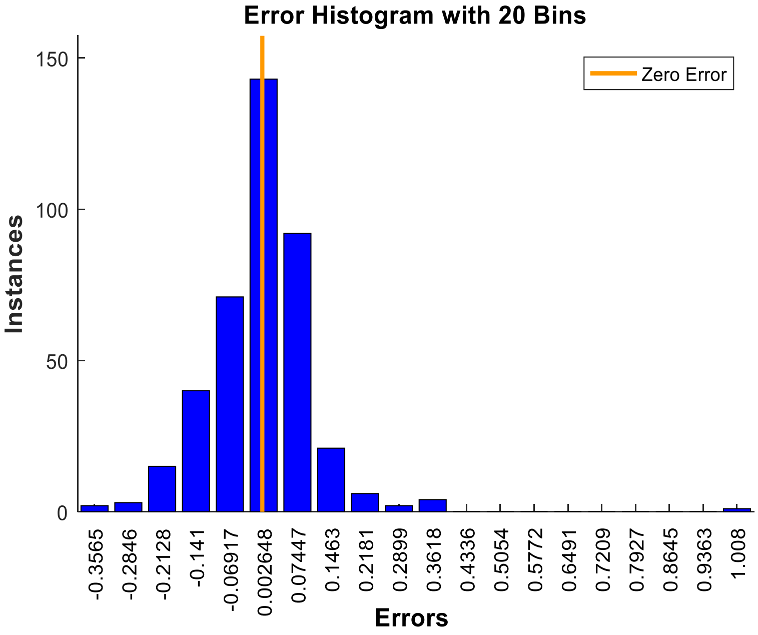

Figure 4. Error Histogram for sample 400 days

Then we use the established LASSO-ANNs model to forecast the data of the remaining 82 sets of test samples. Since the average expected rate of return for the next seven trading days is to be forecast, 75 output results are obtained from 82 sets of test samples. Figure 5 is the predictive effect of neural networks. It can be seen that although there is a slight deviation between the predicted value of the expected return on the next seven trading days and the true value, the trend is almost the same. Through the Figure 6, it can be seen that the individual errors in the test results are large, but the error rate of most predicted values can still be controlled within 7%, which indicates that the model has strong generalization ability and the prediction effect is ideal.

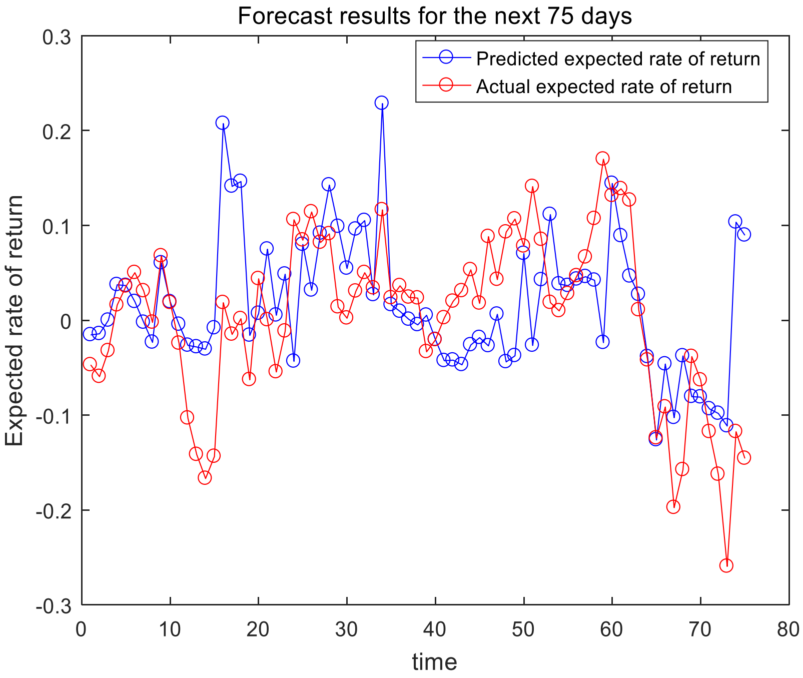

Figure 5. Forecast results for the next 75 days

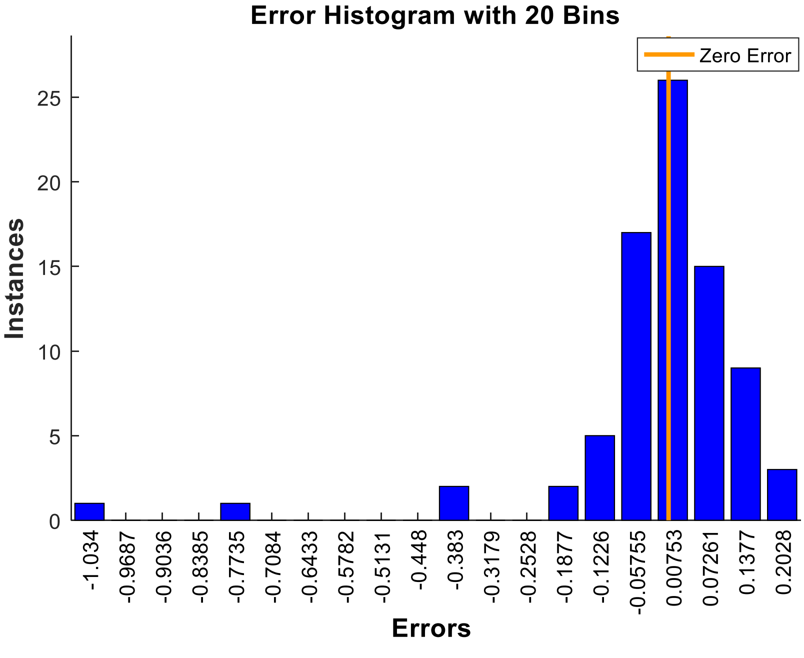

Figure 6. Error histogram for the next 75 days

Figure 7 is error histogram for the neural network training and prediction without the variable selection. By comparing with the Figure 4 and Figure 6, it can be seen that as for the neural network training without the variable selection, the accuracy of prediction accuracy of training samples is slightly higher. But the error in the test process is significantly higher than that of the neural network prediction with LASSO screening. This proves that the selection of variables by LASSO can significantly improve the generalization ability of neural networks.

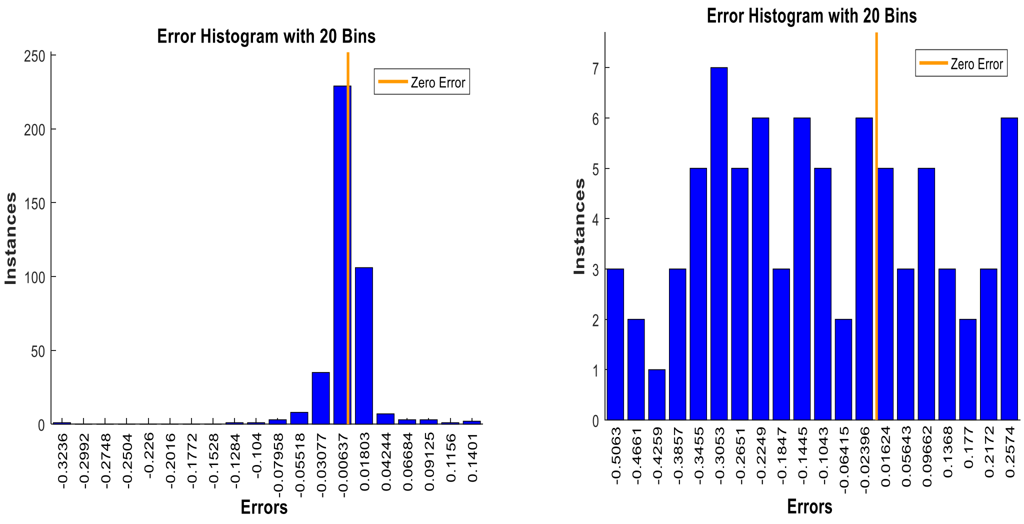

Figure 7. Error histogram for the sample and test data without using lasso

3.2 Results Analysis and Testing

According to the forecast results, we construct the stock index futures trading strategy. According to the rule of China Financial Futures Exchange, we assume that the stock index futures transaction fee is 1.5%% of the amount of turnover. We assume that only the first-hand transaction and short selling is allowed. At the same time we introduce a stop-loss strategy.

According to the LASSO algorithm and the prediction results of the neural network, the yield of the quantitative trading strategy is shown as follows.

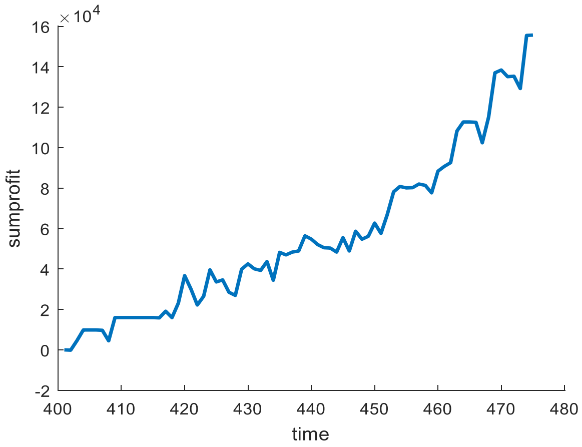

Figure 8. The fund curve of test data

The results of the retest show that the final benefit of 75 trading days is 155670, a total of 7 transactions, the maximum retracement is 23640, the maximum retracement rate is 15.19%. As can be seen from the rate of return, LASSO-ANNs model has a strong profitability. In China's securities market this model can get excess returns.

In order to further test the robustness of the trading strategy, we choose different time periods to test the trading strategy. Each time with 200 sets of data to train, with 80 sets of data to predict:

Time period 1: Training period: 2010.5.21~2011.3.21. Test period: 2011.3.22~2011.7.15.

Time period 2: Training period: 2011.7.18~2012.5.16. Test period: 2012.5.17~2012.9.6.

Time period 3: Training period: 2012.9.7~2013.7.12. Test period: 2013.7.15~2013.11.12

Time period 4: Training period: 2013.11.13~2014.9.2. Test period: 2014.9.3~2014.12.31.

The curves of the funds tested under different time periods are shown in the Figure 9.

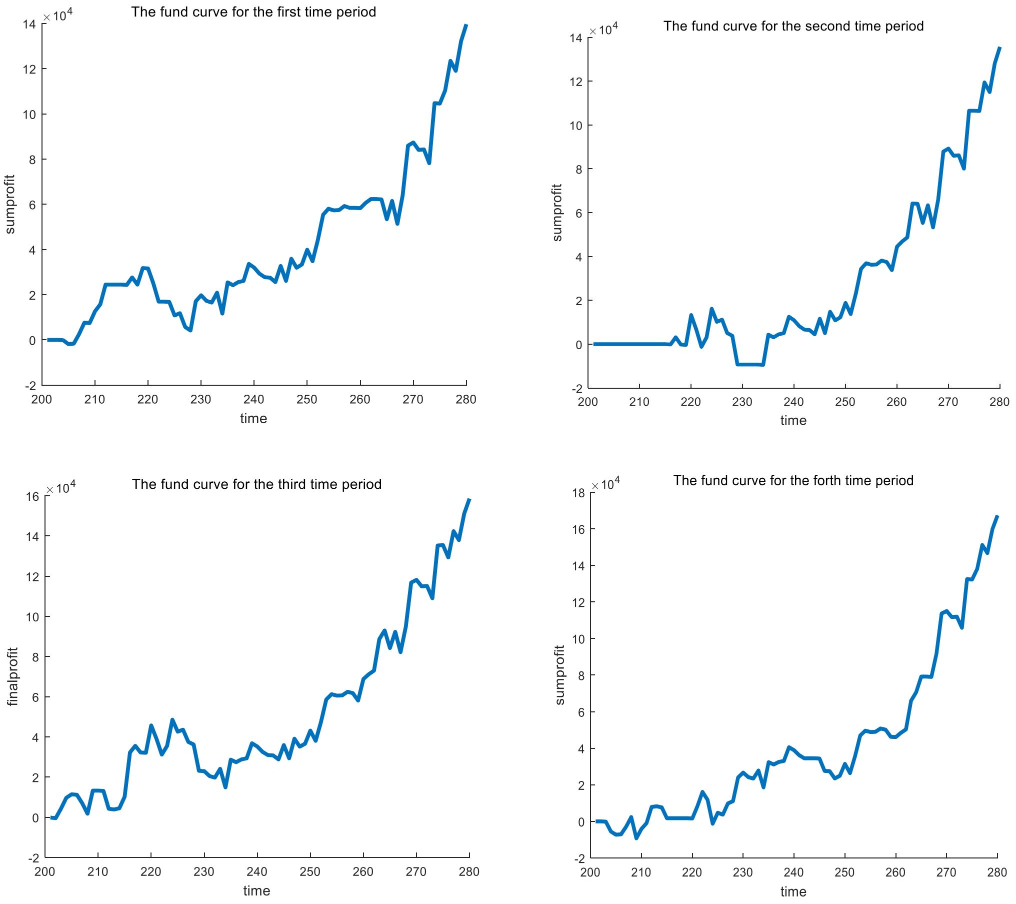

Figure 9. The fund curve of different time periods

The performances of the trading strategy in different time periods are shown in the Table 2.

Table 2. The trading strategy's performance in different time periods

time period	final profit	trade times	maximum retracement	maximum retracement rate
1	139760	8	27608	19.75%
2	135630	4	25615	18.89%
3	158610	7	33716	21.26%
4	167300	5	31076	18.58%

From the Figure 9 and Table 2, the quantitative trading strategy in the different time period all have good performance and it can get a stable income during different times. However, the maximum retracement rate is generally higher. It also can be seen from the fund curve that, there are some obvious short-term fluctuations in the income situation. This shows that the strategy has the characteristics of high risk and high return, and the risk control needs to be strengthened. The results show that the trading strategy can adapt to the stock market of the CSI 300 stock index futures in different periods. The strategy has certain robustness and can effectively carry out short-term forecast.

4. Conclusion

Under the background of the weak forecasting performance of the traditional forecasting method, this paper attempts to predict the stock market and construct the quantitative trading strategy based on the theory of statistics and machine learning. This paper takes the CSI 300 stock index futures as an example to carry on the empirical research. We use the LASSO algorithm to filter technical indexes to obtain the input vector, and take the average return rate in the next few days as the target value. Then we use the neural network to carry on the non-linear prediction and construct the quantitative transaction strategy. Through plentiful repeated tests and statistical analysis, this paper draws the following conclusions:

- LASSO algorithm can filter the variables. In the case of less information loss, the input vector latitude is reduced and the efficiency of model operation is improved.

- Compared with single neural network, the combination of LASSO algorithm and neural network can improve the prediction accuracy, and improve the generalization ability of the prediction model.
- Based on the LASSO-ANNs, the quantitative trading strategy is well performed in different periods, and the optimal returns are obtained. The risk is also within the controllable range. It can be proved that in the Chinese stock index futures market, using LASSO-ANNs to construct the quantitative trading strategy is feasible, and the results can provide some guidance for investors.

China's stock index futures market is still in the initial period of development, so the price fluctuations frequently and the overall risk is high. Compared to other mature financial markets, the forecast is more difficult. The LASSO-ANNs model constructed in this paper can achieve good results in the stock index futures market, which shows that the model can be extended to other markets. In the future study, we can further improve the risk control ability, enrich the research indicators, and try to predict the stock market in the long run.

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. (Vol.31, pp.307–327). Economics and Econometrics Research Institute (EERI), Brussels. https://doi.org/10.1016/0304-4076(86)90063-1

Box, G. E. P., & Jenkins, G. M. (1995). Time series analysis: forecasting and control. *Technometrics*, *68*(2), 303-303.

Breidt, F. J., Davis, R. A., & Trindade, A. A. (2001). Least absolute deviation estimation for all-pass time series models. , *29*(4), 919-946.

Cui, J. F., & Li, X. X.(2004). Stock price forecasting: comparison between GARCH model and BP neural network model. *Statistics and Decision,* (6), 21-22. https://doi.org/10.13546/j.cnki.tjyjc.2004.06.009

Fiorenzani, S. (2006). Quantitative methods for electricity trading and risk management. *Finance & Capital Markets*. https://doi.org/10.1057/9780230598348

Ke, Z. L. (2011). The application of Lasso and its related methods in Multiple Linear Regression Model. (Doctoral dissertation, Beijing Jiao tong University).

Kestner, L. (2003). Quantitative trading strategies. *Journal of Physics Condensed Matter, 15*(38), 6563-6579(17).

Markowitz, H. M. Portfolio selection - Markowitz, harry m. - Yale university press.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374-393. https://doi.org/10.1016/j.csda.2006.12.019

Ross, S. A. (2015). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

Schumann, M., & Lohrbach, T. (1993). Comparing artificial neural networks with statistical methods within the field of stock market prediction. *Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences*, *4*, 597-606. https://doi.org/10.1109/HICSS.1993.284239

Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance, 19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society*, *73*(3), 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

White, H. (1988). Economic prediction using neural networks: the case of IBM daily stock returns. *IEEE International Conference on Neural Networks*, *2*, 451-458. https://doi.org/10.1109/ICNN.1988.23959

Wu, W., Chen, W., & Liu, B. (2001). Prediction of ups and downs of stock market by bp neural networks. *Dalian Ligong Daxue Xuebao/journal of Dalian University of Technology, 41*(1).

Xiao, G. R. (2011). Application Research of Fund Net Value Prediction Based on Neural Network. *Computer Simulation*, *28*(3), 373-376.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418-1429. https://doi.org/10.1198/016214506000000735

Copyrights