

# A New Computational Method of Input Selection for Stock Market Forecasting with Neural Networks

Wei Huang<sup>1,2</sup>, Shouyang Wang<sup>2</sup>, Lean Yu<sup>2</sup>, Yukun Bao<sup>1</sup>, and Lin Wang<sup>1</sup>

<sup>1</sup> School of Management, Huazhong University of Science and Technology,  
WuHan, 430074, China  
{yukunbao, wanglin}@mail.hust.edu.cn

<sup>2</sup> Institute of Systems Science, Academy of Mathematics and Systems Sciences,  
Chinese Academy of Sciences, Beijing, 100080, China  
{whuang, sywang, yulean}@amss.ac.cn

**Abstract.** We propose a new computational method of input selection for stock market forecasting with neural networks. The method results from synthetically considering the special feature of input variables of neural networks and the special feature of stock market time series. We conduct the experiments to compare the prediction performance of the neural networks based on the different input variables by using the different input selection methods for forecasting S&P 500 and NIKKEI 225. The experiment results show that our method performs best in selecting the appropriate input variables of neural networks.

## 1 Introduction

The time series forecasting in stock market is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships[1]. Neural networks (NN) are particularly well suited to finding accurate solutions in an environment characterized by complex, noisy, irrelevant or partial information[2]. Some researchers have conducted work on the stock market time series forecasting by using past value or transformations of them as input variables of neural networks[3]. Neeraj et al studied the efficacy of neural networks in modeling the Bombay Stock Exchange SENSEX weekly closing values. They develop the two neural networks, which are denoted as NN1 and NN2. NN1 takes as its inputs the weekly closing value, 52-week Moving Average of the weekly closing SENSEX values, 5-week Moving Average of the same, and the 10-week Oscillator for the past 200 weeks. NN2 takes as its inputs the weekly closing value, 52-week Moving Average of the weekly closing SENSEX values, 5-week Moving Average of the same, and the 5-week volatility for the past 200 weeks[4]. Yim predicted Brazilian daily index returns. He mapped lagged returns to current returns by using the following three neural networks with backpropagation algorithm. The first neural network has nine lagged returns in the input layer (lags 1 to 9). The second one has only four neurons in the input layer (lags 1, 2, 5 and 9). The third one consists of two neurons (lags 1 and 9) in the input layer. The third neural network produced the best overall results[5]. Yu used the six input which use past prices or transformations of them. The inputs to the

neural networks are as follows: (1) the basis lagged six periods, (2) the RSI differential of the futures price and the index, (3) the MACD differential of the futures price and the index, (4) the change of the basis, (5) the RSI of the basis, (6) the MACD of the basis. His results for out of sample show that the neural network forecast performance is better than that of the ARIMA model[6]. Qi and Zhang use Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as several extensions to select input of neural networks for S&P 500 index. The results indicate that the information-based in-sample model selection criteria are not able to provide a reliable guide to out-of-sample performance and there is no apparent connection between in-sample model fit and out-of-sample forecasting performance[7]. Chaos analysis is a good method to analyze nonlinear dynamics in the time series. Nonlinear dynamics and chaos theory can provide information about the lag structures for the design of forecasting models using neural networks. Chaos analysis criterion (CAC) was applied to determine the lag structure for the input of neural networks based on the embedding dimensions of stock index[8, 9]. However, CAC neglect the special feature of input variables of neural networks that the input variables should not be much correlated.

Our contribution of the paper is to propose a new computational method of selecting input variables for stock market forecasting with neural networks. The computational method results from the synthetically considering the special feature of input variables of neural networks and the special feature of stock market time series. The remainder of this paper is organized as follows. Section 2 describes the new computational method. In Section 3, we conduct the experiments to compare the prediction performance of the neural networks based on the different input variables by using the different input selection methods. Finally, conclusions are given in Section 4.

## 2 Our Input Selection Method

In fact, neural networks for time series forecasting is a kind of nonlinear autoregressive (AR) model as follows:

$$\hat{y}_{t+n} = F ( y_{t-s_1}, y_{t-s_2}, \dots, y_{t-s_i} ) \quad (1)$$

where  $\hat{y}_{t+n}$  is the output of the neural network, namely the predicted value when we make a prediction of  $n$  periods ahead from the present period  $t$ ;  $y_{t-s_1}, y_{t-s_2}, \dots, y_{t-s_i}$  are the inputs of the neural network, namely the actual value at the corresponding period;  $s_i$  is the lag period from the present period  $t$ ;  $F(\bullet)$  is a nonlinear function determined by the neural networks. The problem is to select the appropriate input variables  $( y_{t-s_1}, y_{t-s_2}, \dots, y_{t-s_i} )$  of neural networks.

Usually, the rule of input variable selection is that the input variables should be as predictive as possible. As is well known, autocorrelation coefficient is a popular indicator to measure the correlation of time series. The autocorrelation coefficient of a series  $\{y_t\}$  at lag  $k$  is estimated in the following way:

$$r_k = \frac{\sum_{t=k+1} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1} (y_t - \bar{y})^2} \quad (2)$$

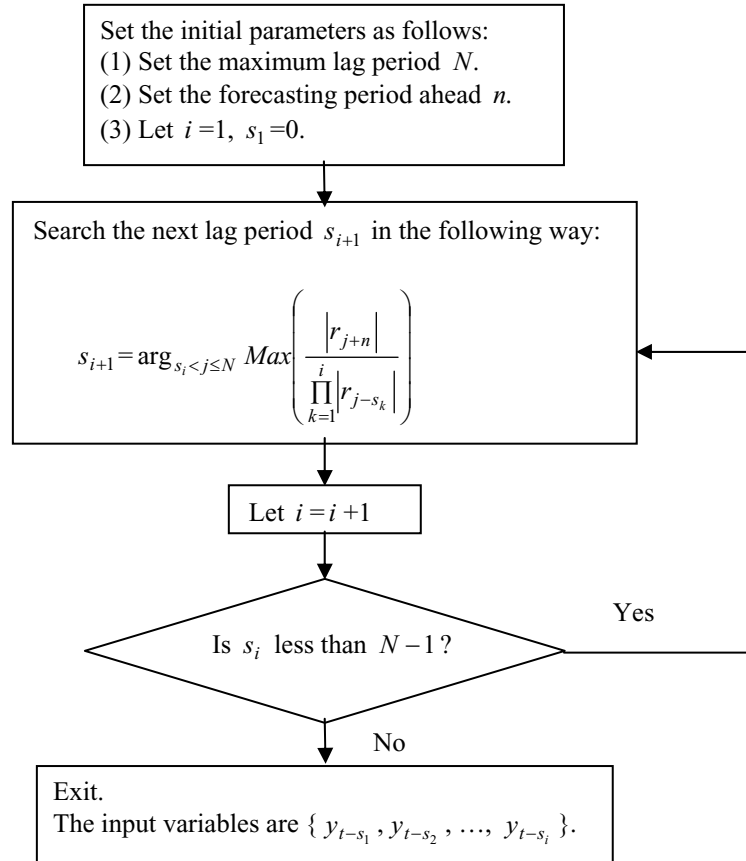
where  $\bar{y}$  is the sample mean of  $\{y_t\}$ . Table 1 shows the autocorrelation coefficients of daily observations of S&P 500 and NIKKEI 225. It indicates that the absolute value of autocorrelation coefficients of stock index prices become smaller when the lag period becomes longer.

**Table 1.** The absolute value of autocorrelation coefficients of daily observations of S&P 500 and NIKKEI 225

$k$	$ r_k $ of S&P 500	$ r_k $ of NIKKEI 225
1	0.935	0.943
2	0.877	0.872
3	0.835	0.812
4	0.790	0.764
5	0.753	0.719
6	0.702	0.660
7	0.649	0.603
8	0.598	0.535
9	0.542	0.478
10	0.487	0.433
11	0.434	0.389
12	0.368	0.345
13	0.306	0.298
14	0.229	0.265
15	0.150	0.235

The special feature of input variables of neural networks is that the input variables should not be much correlated, because the correlated input variables may degrade the prediction performance by interacting with each other as well as other elements and producing a biased effect[10]. Actually, the correlated input variables contribute the similar information for the output variable of neural networks. Therefore, the neural networks get confused and do not know to use which one. In other words, the neural networks may alternate back and forth, and over-fit.

There is a dilemma to select the input variables of neural networks. In order to let the input variables correlated to the output variable, we should choose the input variable with less lags like  $\{y_t, y_{t-1}, y_{t-2}, y_{t-3}, \dots\}$ . However, the above input variables are too correlated to each other. In order to get a trade-off of the two conflicted requirements of input variables selection, we propose a new computational method of input selection for stock market forecasting with neural networks (see Figure 1). It is a



**Fig. 1.** Our method of input selection for stock market forecasting with neural networks

process of selecting the lagged variable which is more correlated to the predicted variable and less correlated to the already selected input variables.

### 3 Experiments Analysis

In order to demonstrate our method, we conduct the experiments to compare the prediction performance of the neural networks based on the different input variables by using the different input selection methods.

#### 3.1 Neural Network Models

In order to reduce the degrees of freedom in the developed neural network models and to maintain consistency with previous research efforts, we focus on the following two popular feedforward neural network models: (1) 3-layers back-propagation networks with adaptive learning rate and momentum (BPN); (2) radial basis function networks (RBFN).

### 3.2 Naïve Prediction Hypothesis

The naïve prediction hypothesis asserts today's stock price as the best estimate of tomorrow's price. It can be expressed as follows:

$$\hat{y}_{t+1} = y_t \quad (3)$$

where  $\hat{y}_{t+1}$  is the predicted value of the next period;  $y_t$  is the actual value of current period.

### 3.3 Performance Measure

Normalized mean squared error (NMSE) is used to evaluate the prediction performance of neural networks. Given a set  $P$  comprising pairs of the actual value  $x_k$  and predicted value  $\hat{x}_k$ , the NMSE can be defined as follows:

$$\text{NMSE} = \frac{\sum_{k \in P} (x_k - \hat{x}_k)^2}{\sum_{k \in P} (x_k - \bar{x}_k)^2} \quad (4)$$

where  $\bar{x}_k$  is the mean of actual values.

### 3.4 Data Preparation

We obtain the daily observation of two stock indices, S&P500 and NIKKEI225, from the finance section of Yahoo. The entire data set covers the period from January 2001 to November 2005. The data sets are divided into two periods: the first period covers from January 2001 to December 2004 while the second period is from January 2005 to November 2005. The first period is used to estimate the model parameters. We select the appropriate size of training set by using the method in [11]. The second period is reserved for out-of-sample evaluation and comparison.

### 3.5 Results

Table 2 shows the prediction performances of the naïve prediction, which are used as benchmarks of prediction performance of S&P500 and NIKKEI225. In order to investigate the effects of the maximum lag period size on the prediction performance of neural networks, we let the maximum lag period  $N = 8, 10, 12$  respectively for one forecasting period ahead, namely  $n = 1$ . Table 3 shows the input variables of the neural networks for forecasting S&P 500 and NIKKEI 225 by using our method, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Chaos analysis criterion (CAC).

**Table 2.** The prediction performance of the naïve prediction

NMSE for S&P 500	NMSE for NIKKEI 225
0.2168	0.2743

**Table 3.** The input variables of the neural networks for forecasting S&P 500 and NIKKEI 225 by using the different input selection methods

Input selection method	Inputs variables for S&P 500	Inputs variables for NIKKEI 225
Ours ( $N = 8$ )	$\{y_t, y_{t-3}, y_{t-7}\}$	$\{y_t, y_{t-4}, y_{t-6}, y_{t-8}\}$
Ours ( $N = 10$ )	$\{y_t, y_{t-3}, y_{t-7}, y_{t-9}\}$	$\{y_t, y_{t-4}, y_{t-6}, y_{t-9}\}$
Ours ( $N = 12$ )	$\{y_t, y_{t-3}, y_{t-7}, y_{t-9}, y_{t-11}\}$	$\{y_t, y_{t-4}, y_{t-6}, y_{t-9}, y_{t-11}\}$
AIC	$\{y_t, y_{t-2}, y_{t-5}, y_{t-8}\}$	$\{y_t, y_{t-1}, y_{t-6}, y_{t-8}\}$
BIC	$\{y_t, y_{t-3}\}$	$\{y_t, y_{t-3}\}$
CAC	$\{y_t, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}\}$	$\{y_t, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}\}$

Table 4 shows the prediction performance of 3-layers back-propagation networks with adaptive learning rate and momentum(BPN) with the different input variables determined by using the different input selection methods. Table 5 shows the prediction performance of radial basis function networks(RBFN) with the different input variables determined by using the different input selection methods. The value of NMSE by using our method is the smallest among the different input selection methods, when the initial parameter maximum lag period  $N = 8, 10, 12$  respectively. It shows that our method performs best in selecting the appropriate input variable of the neural networks for forecasting S&P 500 and NIKKEI 225. Because our method balance the two conflicted need of input variables of neural networks: (1) the input variables should be more correlated to the output variable; (2) the input variables should be less correlated to each other. The chaos analysis criterion (CAC) performs worst in selecting the appropriate input variable of the neural networks for forecasting S&P 500 and NIKKEI 225. Compared with the naïve prediction, the neural networks perform better except when using the input variable determined by the chaos analysis criterion (CAC). Because CAC doesn't consider the special feature of input variable of neural networks, and the selected input variables are too correlated to each other. Our method doesn't require any assumptions, completely independent of particular class of model. The method makes full uses of information among sample observations even if the underlying relationships are unknown or hard to describe. Therefore, it is a very practical way to select the input variable of the neural networks when the financial time series is hard to model.

**Table 4.** The prediction performance of BPN for S&P 500 and NIKKEI 225 forecasting by using the different input selection methods

Input selection method	NMSE for S&P 500	NMSE for NIKKEI 225
Ours ( $N = 8$ )	0.0907	0.0915
Ours ( $N = 10$ )	0.0912	0.0923
Ours ( $N = 12$ )	0.0962	0.0983
AIC	0.1254	0.1357
BIC	0.0974	0.0992
CAC	0.3256	0.3863

**Table 5.** The prediction performance of RBFN for S&P 500 and NIKKEI 225 forecasting by using the different input selection methods

Input selection method	NMSE for S&P 500	NMSE for NIKKEI 225
Ours ( $N = 8$ )	0.0921	0.0929
Ours ( $N = 10$ )	0.0932	0.0946
Ours ( $N = 12$ )	0.0978	0.0998
AIC	0.1288	0.1396
BIC	0.1106	0.1197
CAC	0.4352	0.4879

## 4 Conclusions

In this paper, we propose a new computational method of input selection for stock market forecasting with neural networks. The method results from synthetically considering the special feature of input variables of neural networks and the special feature of stock market time series. The advantage of our method is data-driven in that there is no prior assumption about the time series under study. The experiment results show that our method outperforms the others in the prediction performance for stock market time series forecasting with neural networks.

## Acknowledgements

This work is partially supported by National Natural Science Foundation of China (NSFC No.70221001, 70401015) and the Key Research Institute of Humanities and Social Sciences in Hubei Province-Research Center of Modern Information Management.

## References

1. Hall, J. W.: Adaptive selection of US stocks with neural nets. in Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets, G. J. Deboeck, Eds. New York: Wiley; (1994) 45-65
2. Huang, W., Lai, K.K., Nakamori, Y. & Wang, S.Y.: Forecasting foreign exchange rates with artificial neural networks: a review. International Journal of Information Technology & Decision Making, 3(2004) 145-165
3. Huang, W., Nakamori, Y. & Wang, S.Y.: Forecasting Stock Market Movement Direction with Support Vector Machine. Computers & Operations Research, 32 (2005) 2513-2522
4. Neeraj, M., Pankaj, J., Kumar, L. A. & Goutam, D.: Artificial neural network models for forecasting stock price index in Bombay Stock Exchange. Working Papers with number 2005-10-01 in Indian Institute of Management Ahmedabad, Research and Publication Department, (2005)
5. Yim, J.: A comparison of neural networks with time series models for forecasting returns on a stock market index. Lecture Notes in Computer Science, Vol. 2358, Springer-Verlag Berlin Heidelberg (2002)

6. Yu, S. W.: Forecasting and arbitrage of the Nikkei stock index futures: an application of backpropagation networks. *Asia-Pacific Financial Markets*, 6(1999) 341–354
7. Qi, M. & Zhang, G. P.: An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, 132(2001) 666-680
8. Embrechts, M., Cader, M. & Deboeck, G. J.: Nolinear dimensions of foreign exchange, stock and bond markets. in *Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets*, G. J. Deboeck, Eds. New York: Wiley, (1994) 297–313
9. Oh K. J. & Kim, K.: Analyzing stock market tick data using piecewise nonlinear model. *Expert Systems with Applications*, 22(2002) 249-255
10. Zhang, G.P.: *Neural Networks in Business Forecasting*. Idea Group Inc., (2003)
11. Huang, W., Nakamori, Y., Wang, S.Y. & Zhang, H.: Select the size of training set for financial forecasting with neural networks. *Lecture Notes in Computer Science*, Vol. 3497, Springer-Verlag Berlin Heidelberg (2005) 879–884