_____

Lecturer Jujie WANG, PhD
E-mail: wjj0913@126.com
Danfeng QUE, PhD Candidate
E-mail: quintaqdf@gmail.com
School of Management Science and Engineering
Nanjing University of Information Science and Technology
Nanjing 210044, China

AN EXPERIMENTAL INVESTIGATION OF TWO HYBRID FRAMEWORKS FOR STOCK INDEX PREDICTION USING NEURAL NETWORK AND SUPPORT VECTOR REGRESSION

Abstract. Stock price index (SPI) prediction is essential to investors when they make investment decisions. However, due to the intrinsic complexity and non-stationary of SPI, accurate SPI prediction has been a challenge task for a longtime. Compared to a quantity of single algorithms, hybrid methods always have higher prediction accuracy. Based on the theories of genetic algorithm (GA), grey model (GM), back-propagation neural network (BPNN) and support vector regression (SVR), in this paper, two hybrid models named GA-SVR-GM and GA-BPNN-GM are proposed for improving the prediction accuracy of Composite Index in Shanghai Stock Exchange (SHSE) and Component Index in Shenzhen Stock Exchange (SZSE). The SVR (or BPNN) is first used to forecast SPI and then with the obtained errors GM is constructed taking into account the different patterns that the SVR (or BPNN) cannot identify, reducing with the final errors. Moreover, in order to obtain the better forecasting results of SVR (or BPNN), the model parameters are optimized by GA. Two prediction cases of SPI are used to test the performance of the proposed hybrid models. Compared with the random walk (RW), SVR, and BPNN, the simulation results show that the GA-SVR-GM and GA-BPNN-GM consistently have the minimum statistical error of the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Thus, it can be concluded that such improving methods are effective.

Keywords: Stock index prediction, Neural network, Support vector regression, Genetic algorithm, Grey model, Hybrid models.

JEL Classification: O30

1. Introduction

Stock price index (SPI) is an important reference for investors to make investment strategies in financial market (Arevalo et. al., 2017). Prediction of SPI has attracted more attention of both private and institution investors. However, due to the intrinsic complexity and non-stationary of SPI, accurate SPI prediction has

_____

still been a challenge task for both academic researchers and practitioners (Chen et. al., 2017). Although faced with complicated challenges, many experts and scholars have proposed a lot of models for SPI prediction over the past three decades. According to existing literature (Wang et. al., 2017),there are two main categories for the SPI forecasting. The first category can be referred to conventional statistical models, such as autoregressive integrated moving average models (ARIMA), exponential smoothing models (ESM), random walk, and so forth (Huang et. al., 2017). These approaches mainly assume that a linear correlation structure exists among SPI. Thus, the prediction performance of these models will be worse when the nonlinear characteristics of SPI are obvious. In other words, these statistical models maybe insufficient to capture the hidden nonlinear features in SPI.

To overcome the drawbacks of the conventional statistical models, artificial intelligence (AI) techniques with powerful nonlinear self-learning capacities, primarily including back-propagation Neural Network (BPNN) and support vector regression (SVR), have become increasingly popular for SPI forecasting. For example, Huang et al. (2005)evaluated the performance of SVR when forecasting the weekly movement of direction of NIKKEI 225 index compared with some classical models. Kara et al. (2011) employed BPNN and SVR to predict movement in the daily Istanbul Stock Exchange (ISE) National 100 Index. Experimental results showed that the performance of BPNN was found significantly better than that of SVR model. Okasha (2014) investigated ARIMA, BPNN and SVR models fitted into Al-Quds Index of the Palestinian Stock Market and found that SVR performed better than other models, which persuasively demonstrated that the formation of complex nonlinear system of SVR model could portray financial market.

However, these AI models cannot yield impressive results in all prediction issues. Many research efforts have indicated that there is no single best prediction method that can be applied to any specific situation(Ince and Trafalis, 2017; Li, 2017). As a result, hybrid methods that combine two or more individual models have been proposed to increase the chance of capturing different patterns in the data and improve the prediction performance. For instance, Hu et al. (2016) proposed a short-term traffic flow hybrid forecasting method based on particle swarm optimization (PSO) and SVR. Experimental results showed that the hybrid method can get accurate forecasting results than individual models. Cao and Parry(2009) examined the relative effectiveness of hybrid model based on BPNN and genetic algorithm (GA) in forecasting future earnings per share. Gu et al. (2011) proposed a housing price forecasting based on GA and SVR. Zhou and Hu (2008) proposed an effective hybrid approach based on grey model (GM) and autoregressive moving average model (ARMA) for forecasting gyro drift. Jia et al. (2010) proposed a characteristic forecasting of hydraulic valve based on GM and adaptive neuro-fuzzy system (ANFIS). Alvisi and Franchini (2012) proposed a hybrid model based on GM and ANN for river stage forecasting with uncertainty.

From the forecasting methodologies discussed above, it can be found that:
1) BPNN and SVR usually have better performance compared to the other
individual models. 2) Hybrid methods always have higher prediction accuracy in
forecasting issues. However, most of these hybrid methods tried to use one of the
techniques, such as parameter selection and combining several models to improve
the prediction performance, and few considered these techniques comprehensively.
This study proposes to build an integrated prediction strategy for SPI considering
all these perspectives. In this paper, two hybrid models named GA-SVR-GM and
GA-BPNN-GM are proposed for improving the prediction accuracy of Composite
Index in Shanghai Stock Exchange (SHSE) and Component Index in Shenzhen
Stock Exchange (SZSE). The SVR (or BPNN) is first used to forecast SPI and then
with the obtained errors GM is constructed taking into account the different
patterns that the SVR (or BPNN) cannot identify, reducing with the final errors.
Moreover, in order to obtain the better forecasting results of SVR (or BPNN), the
model parameters are optimized by GA. Two prediction cases of SPI are used to
test the performance of the proposed hybrid models. Compared with the random
walk (RW), SVR, and BPNN, the simulation results show that the GA-SVR-GM
and GA-BPNN-GM consistently have the minimum statistical error of the root
mean square error (RMSE), mean absolute error (MAE), and mean absolute
percentage error (MAPE). Thus, we can conclude that such improving methods are
effective.

This paper is organized as follows. In section 2, the basic theory of hybrid
models will be introduced. In section 3, the concrete datasets will be applied to the
hybrid models and the performance of these models will be presented. And section
4 will outline the conclusions.

2. Proposed approach

In this paper, two categories of hybrid models called GA-SVR-GM and
GA-BPNN-GM separately, which apply genetic algorithm (GA) into support
vector regression (SVR) or back-propagation neural network (BPNN), and then
use grey model (GM) to adjust residual error, are proposed in order to forecast the
movement of stock price index (SPI) in China. The algorithm is briefly described
as follows, and the flowchart is shown in Figure 1.

Step 1: Select explanatory input variables and input them into the SVR or BPNN.
Step 2: Train the SVR model or BPNN model with the output variables based on
the optimal parameters obtained from GA.
Step 3: Use GM to adjust the difference of forecast output value and actual value
called residual error, and add the forecast error of GM up to the output value of
SVR or BPNN model to obtain the final forecast value.
Step 4: Evaluate the forecasting performance of these two hybrid models according
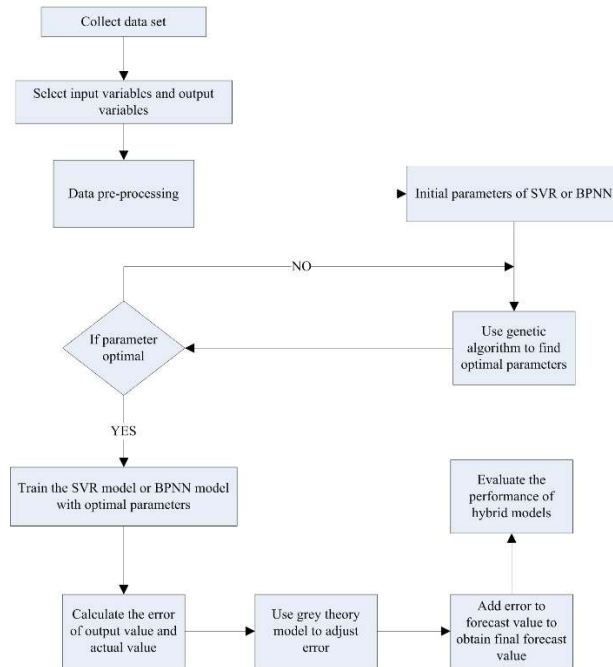to some criteria.

_____

Figure 1. GA-SVR-GM or GA-BPNN-GM flowchart

2.1 Genetic algorithm (GA)

GA is adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. The primary components of GA are the chromosome encoding, the fitness function, selection, crossover and mutation. A typical design for classical GA using standard genetic operators is below:

Step 1: Randomly generate an initial source of chromosomes called father chromosome.

Step 2: Calculate the fitness function of each chromosome in the source population.

Step 3: Create an empty successor population.

(a) Using proportional fitness selection, select two chromosomes from the source population.

(b) Apply one-point crossover with crossover rate pc to obtain a child chromosome.

(c) Or, apply uniform mutation with mutation rate pm to produce another child chromosome.

(d) Add this child chromosome to the successor population.

Step 4: Replace the source population with the successor population.

Step 5: If stopping criteria have not been met, return to Step 2.

2.2 Support vector regression (SVR)

Support vector machine (SVM) is put forward by Vapnik (1999), superior in settling the problems of limited samples, non-linearity and high-dimension. In especial, it can overcome the worrisome outcomes brought by artificial neural

196

network including dimensional disaster and over learning. When two categories of samples of training data sets $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ can be classified through linear method approximately, which means that such special circumstances where samples do not conform to limited condition $y_i((\omega * x_i) + b) \geq 1$ are permitted to exist. Consequently, to solve the problem, a slack variable $\zeta_i \geq 0 (i = 1, 2, \ldots, m)$ is introduced and the limited condition of softening is $y_i((\omega * x_i) + b) \geq 1 - \zeta_i (i = 1, 2, \ldots, m)$. To avoid a large value of ζ_i, variable ζ_i should be punished in the target function, and the quadratic programming for this problem is listed as follows:

$$
\begin{aligned}
\min \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} \zeta_i \\
s.t. \quad & y_i((\omega * x_i) + b) \geq 1 - \zeta_i \\
& \zeta_i \geq 0, i = 1, 2, \ldots, m
\end{aligned}
\tag{1}
$$

Where: C>0 and C denotes penalty cost for deviation during process, and b is a constant term.

The Lagrange function of the quadratic programming is:

$$
\begin{aligned}
L(\omega, b, \zeta, \alpha, \gamma) = \frac{1}{2}\|\omega\|^2 + \\
C\sum_{i=1}^{l} \zeta_i - \sum_{i=1}^{m} \alpha_i (y_i((\omega * x_i) + b) - 1 + \zeta_i) - \sum_{i=1}^{m} \gamma_i \zeta_i
\end{aligned}
\tag{2}
$$

Where: $\gamma_i \geq 0, \zeta_i \geq 0$. According to the principle of dual function, the dual problem of the origin problem is:

$$
\begin{aligned}
\max \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) + \sum_{i=1}^{m} \alpha_i \\
s.t. & \sum_{i=1}^{m} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, m
\end{aligned}
\tag{3}
$$

Via solving the quadratic programming problem, the global optimum solution $\alpha^* = (\alpha_1^*, \alpha_2^*, \alpha_m^*)^T$ and $\omega^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i$ can be obtained simultaneously. If

_____

choosing $0 < \alpha_j^* < C$, the optimum solution $b^* = y_i - \sum_{i=1}^{m} y_i \alpha_i^* (x_i \cdot x_j)$ can be achieved.

Grounded on the theory of linear support vector machine classification, if converting x_i to nonlinear function $\varphi(x_i)$, and converting $(x_i \cdot x_j)$ to kernel function $K(x_i, x_j)$, nonlinear support vector regression (SVR) can be obtained. The equation of kernel function is:

$$K(x_i, x_j) = (\varphi(x_i) \cdot \varphi(x_j)) \qquad (4)$$

In this paper Gaussian kernel function is selected, and its formula is:

$$K(x_i, x_j) = \exp\{-\frac{\|x_i - x_j\|^2}{\sigma^2}\} \qquad (5)$$

Therefore, the optimal regression function is:

$$f(x) = \sum_{i=1}^{m} \alpha_i^* y_i K(x_i, x_j) + b^*$$

$$b^* = y_j - \sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x_j) \qquad (6)$$

2.3 Back-propagation neural network (BPNN)

Artificial neural network (ANN) has the merits of non-linear and flexibility, thus it has a high capacity of approximation in theory. In addition, it is able to learn the patterns of data sets and can be close to any type of function, such as linear function and nonlinear function. Therefore, it is well applied in the financial market because the financial market indicates the characteristics of noise, diversity and nonlinearity. One of the most applicable ANN is Back-propagation neural network (BPNN), and it adjusts its weights based on error of output value and real value. Here the BPNN with one input layer, one hidden layer and one output layer is chosen. The concrete studying process of BPNN is as follows:

Step 1: For any set of input variable $X = (x_1, x_2, \ldots, x_d)$, initial weight of input layer $w^{(1)}$ is randomly generated to calculate the value which is put into hidden layer, and d represents for the number of cells of input layer. Therefore, the output value of input layer α_j is:

198

$$\alpha_j = \sum_{i=1}^{d} w_{ij}^{(1)} x_i \qquad (7)$$

Step 2: Logistic function as the activate function of hidden layer is used, thus the output result of hidden layer h_j is:

$$h_j = g(\alpha_j) = g(\sum_{i=1}^{d} w_{ij}^{(1)} x_i)$$

$$g(\mathrm{x}) = \frac{1}{1 + e^{-x}} \qquad (8)$$

Step 3: The linear weighted aggregative method is still applied into output layer, the weight is $w^{(2)}$, m stands for the cell number of output layer, and the final output y_k is:

$$\mathrm{y}_k = \sum_{j=1}^{m} w_{jk}^{(2)} h_j \qquad (9)$$

Step 4: Based on the output value y_k and the real value C_k, the modified error of each cell can be calculated, and the formula is:

$$d_k = (\mathrm{y}_k - C_k) * C_k * (1 - C_k) \qquad (10)$$

Step 5: According to such modified error the initial weight can be adjusted, repeating steps from 1 to 4, and defining target function. In other words, when the sum of square of difference of output value and real value is less than a pre-definitive value, the training process stops. Otherwise, when the number of learning is larger than a pre-definitive value, the training process stops as well.

2.4 Grey model (GM)

In the viewpoint of GM, the disorderly raw data can be turned to a regular series for the benefit of grey modelling, transferred to a dimensionless series to a unidirectional series for decision making. The steps for establishing GM(1,1) model for grey prediction are as follows:

Step 1: Suppose that

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n))$$

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)) \tag{11}$$

$$x^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m)$$

Where: $x^{(0)}$ denotes the raw data.

Step 2: Establish the GM(1,1) differential equation as a differential equation with one order and one variable below:

$$\frac{dx^{(1)}}{dt} + \alpha x^{(1)} = b \tag{12}$$

Where: b is a constant term. The derivative definition can be stated roughly as follows:

$$\frac{dx^{(1)}(t)}{dt} \cong \frac{x^{(1)}(k + \Delta t) - x^{(1)}(k)}{\Delta t} \tag{13}$$

Let Δt equals 1, then

$$\frac{dx^{(1)}(t)}{dt} \cong x^{(1)}(k+1) - x^{(1)}(k) = x^{(0)}(k) \tag{14}$$

In addition, set $x^{(1)} \cong z^{(1)}$ and $z^{(1)}$ can be calculated as follows:

$$z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(1)}(k+1)}{2} \tag{15}$$

Thus, based on the equation above, a new expression is obtained:

$$x^{(0)}(k) + \alpha z^{(1)} = b \tag{16}$$

Step 3: Based on the least square method to obtain the parameters a and b, therefore,

$$\theta=[a,b]^T = (B^TB)^{-1}B^T X$$

$$X = B\theta$$

$$X = [x^{(0)}(2), x^{(0)}(3),\ldots, x^{(0)}(n)]^T$$

$$B = \begin{bmatrix} -\frac{1}{2}[x^{(1)}(1)+x^{(1)}(2)] & 1 \\ -\frac{1}{2}[x^{(1)}(2)+x^{(1)}(3)] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}[x^{(1)}(n-1)+x^{(1)}(n)] & 1 \end{bmatrix} \qquad (17)$$

Step 4: Use the grey differential equation to obtain the grey AGO equation, and
the equation is as follows:

$$\hat{x}^{(0)}(k+1)=[x^{(0)}(1)-\frac{b}{a}]e^{-ak} + \frac{b}{a}, k=1,2,\ldots,n \qquad (18)$$

Next, use the IAGO for reduction to obtain the required forecasting
model:

$$\hat{x}^{(0)}(k)=[x^{(0)}(1)-\frac{b}{a}](1-e^{\alpha})e^{-a(k-1)}, k=1,2,\ldots,n \qquad (19)$$

3. Results and discussion

3.1 Datasets

The datasets investigated in this paper are Chinese Shanghai Stock
Exchange (SHSE) Composite Index and Shenzhen Stock Exchange (SZSE)
Component Index from website of NetEase, because composite index integrates all
the stocks transacted in SHSE while 40 selected stocks transacted in SZSE. The
datasets both cover the transaction time period between 04/01/2012 and
06/06/2016 amounting to 1053 transactions. To test the validity of the proposed
hybrid approaches and further evaluate the prediction accuracy, the dataset is
partitioned into a training data set (85%) and a validation data set (15%). The
training data set can be applied to train the hybrid models, and the validation data
set can be applied to evaluate the performance and effectiveness of the hybrid
models.

Before using models to train data sets, the input variables and output
variables must be determined. The input variables chosen are daily closing prices,
daily opening prices, the corresponding logarithmic absolute returns and realized
volatility at time t of the Composite Index. The output variable is daily closing

price at time t+1. The reason of choosing such explanatory input variables and output variables is that daily closing price can represent the movement of stock price index in China to a large degree and the input variables are highly relevant to the output variable according to common sense and basic tests.

Let p_t (t=1,2,….) denote the daily closing prices, thus the calculation of the logarithmic absolute returns r_t is:

$$r_t = \ln p_t - \ln p_{t-1} \qquad (20)$$

The figure of the logarithmic absolute returns of Composite Index in SHSE and Component Index in SZSE are as follows:
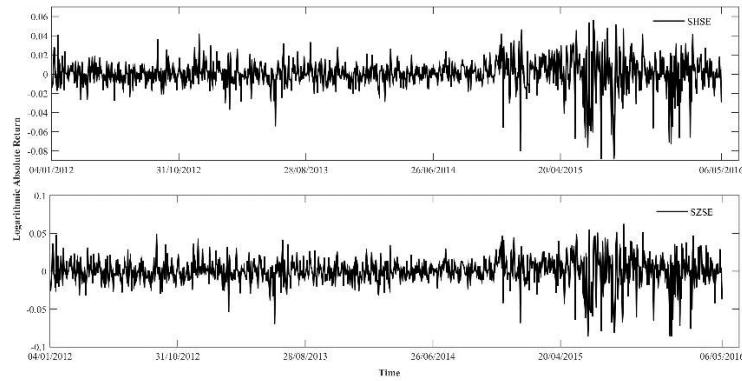
Figure 2. Logarithmic absolute returns of Composite Index in SHSE and Component Index in SZSE

Volatility is to measure the fluctuation degree, meaning risk. The equation of realized volatility is:

$$RV_t = \sqrt{n^{-1} \sum_{i=t-n+1}^{t} (r_i - \overline{r})^2}$$

$$\overline{r} = n^{-1} \sum_{i=t-n+1}^{t} r_i \qquad (21)$$

Where: r_i denotes logarithmic absolute returns, and $\overline{r}$ denotes average value of logarithmic absolute returns of n days prior to time t. Let we set that n equals 11, this is because there are 22 transaction days in a month and 11 days is approximate to half of a month. Figure3 represents realized volatility of Composite Index in SHSE and Component Index in SZSE.
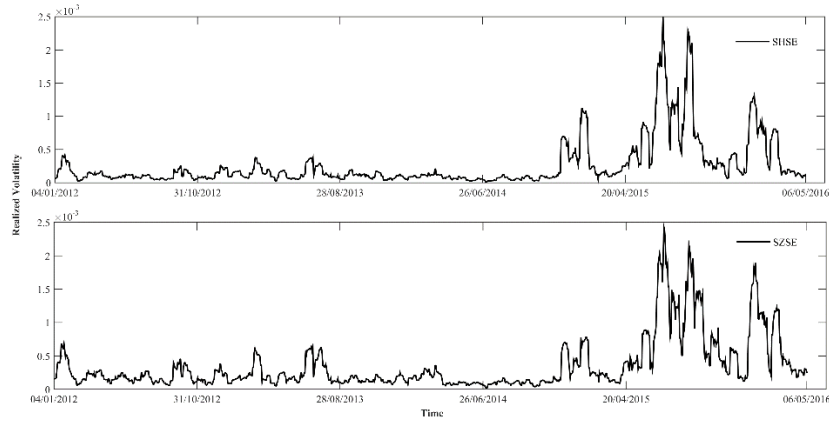
**Figure 3.Realized Volatility of Composite Index in SSE and Component Index
in SZSE**

3.2 Normalization

For sake of inconsistence of order of magnitude such the example as realized volatility is too small and daily closing prices are too large, the data should be normalized at the beginning of modeling. By way of letting the normalized data in the range of [0 , 1], the following equation is adopted:

$$X^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{22}$$

Where: X^* denotes the training data, and x denotes origin data. In order to obtain the accurate value after the forecasting, the output value can be reverted with the following formula:

$$x' = X^* \cdot (x_{max} - x_{min}) + x_{min} \tag{23}$$

Where: x' denotes the invert value of output value $X^{*'}$.

Therefore, based on such normalization method, input variables and the output variable shall be processed through hybrid models.

3.3 Forecasting results of hybrid models

When modeling with GA-SVR, the train datasets cover the period from 04/02/2012 to 10/09/2015 amounting to 894 transactions, and the prediction datasets cover the period from 11/09/2015 to 06/05/2016 amounting to 158

_____

transactions. When applying GA, the crossover rate pc equaling 0.8 is defined, and the mutation rate pm meets 0.9. To find best parameters of SVR, which means finding the optimal penalty parameter c and kernel function parameter g, the initialization of pair value (c, g) is based on grid method. In addition, the fitness function is set as to minimize least square error between output value of SVR and actual value. Specifically, the crossover operator is to get the average value of c or g proximity to the selected parameters to get the child parameters, and the mutation operator is to add a small random number to the selected parameters to get a new pair of parameters. Similarly, the principal of GA-BPNN is absolutely the same as GA-SVR, in which the biggest difference is to replace SVR with BPNN.

According to the above calculation methods, the output value of GA-SVR or GA-BPNN can be achieved, and the residual error which denotes the actual value minus the output value of daily closing price of Composite Index and Component Index can also be calculated. GM can be used to simulate the residual error prior to validation datasets but the time span equals to the period of validation sets, adding the residual error to the forecasting output value in order to obtain the final forecasting value, then this value can be reversed through reserved normalization method. Figure 4 and 5 show the forecasting performance of GA-SVR-GM model and GA-BPNN-GM model in two prediction cases.

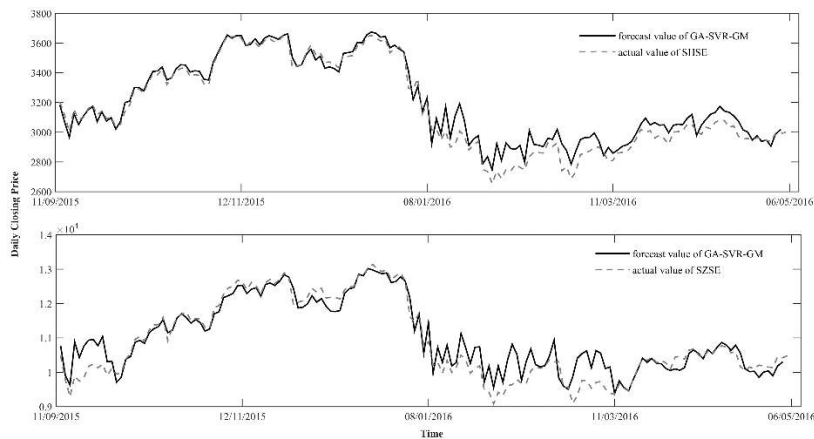### 3.4 Forecasting results of classical models

Figure 4. GA-SVR-GM Performance of Composite Index in SHSE and Component Index in SZSE

As a naïve predictor, the random walk (RW) is often used as a benchmark to evaluate other models in the field of prediction. In general, a new prediction approach is initially compared with the RW to evaluate its performance. In this paper, to assess the performance of these two hybrid models in daily closing prices

prediction, the performance of GA-SVR-GM and GA-BPNN-GM models is compared with that of RW, SVR and BPNN models.
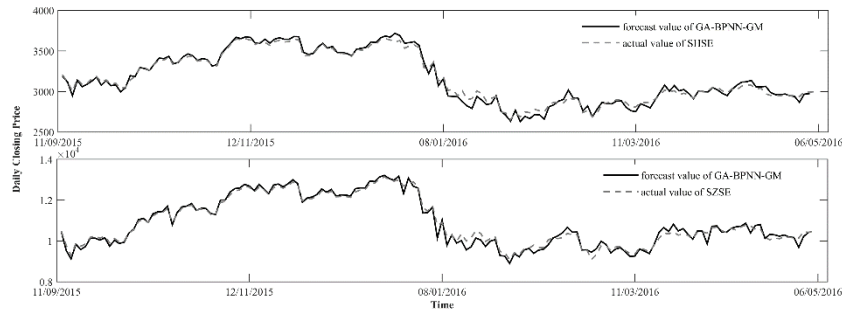

Figure 5. GA-BPNN-GM Performance of Composite Index in SHSE and Component Index in SZSE
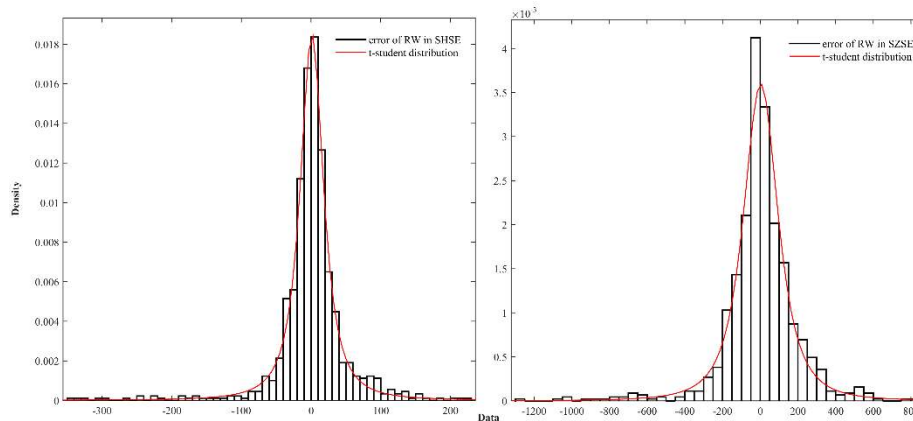
3.4.1 Random walk (RW)

Figure6. Fitted performance of t-student distribution of residue of Composite Index and Component Index

The random walk theory is that price changes have the same distribution and are independent of each other, so that the past movement or trend of a stock price or market cannot be used to predict its future movement. The formula is as follows:

$$p_t = p_{t-1} + \varepsilon_t \tag{24}$$

_____

Where p_t denotes the daily closing price on day t, and p_{t-1} denotes the daily closing price on day t-1, ε_t is a stationary random disturbance term called residue.
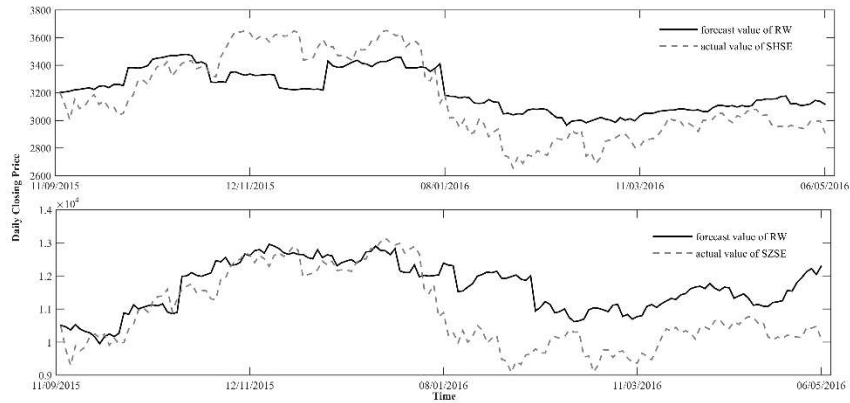

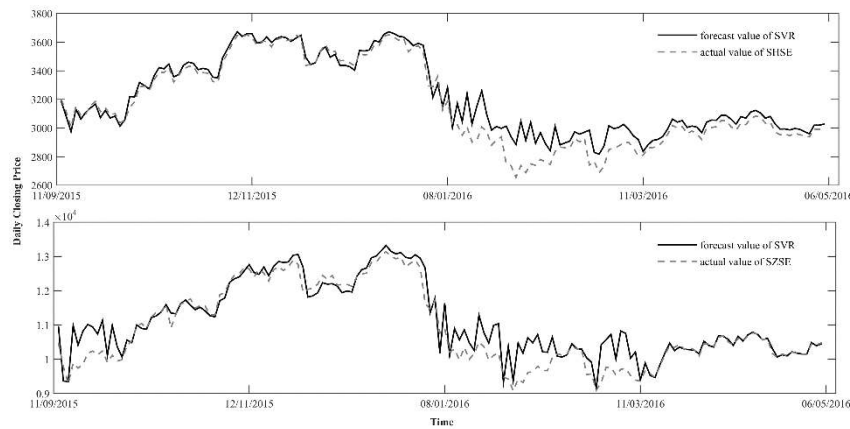**Figure 7. RW Performance of Composite Index in SSE and Component Index in SZSE**

Figure 8. SVR Performance of Composite Index in SSE and Component Index in SZSE

Before predicting daily closing prices with random walk model, which distribution the residue of daily closing price belongs to should be determined. When using t-student distribution to fit the residue of Composite Index in SHSE and Component Index in SZSE, this kind of distribution has perfect performance as Figure 6. Figure 7 shows the forecasting performance of RW model.

206

3.4.2 Classical SVR and BPNN

When training datasets with classical SVR and BPNN, the learning rate of BPNN model can be set as 0.05 and the threshold of residual function equals to 0.0001. Evaluating the performance of these two models independently with the validation data sets, Figure 8 and 9 show the results of the classical models.
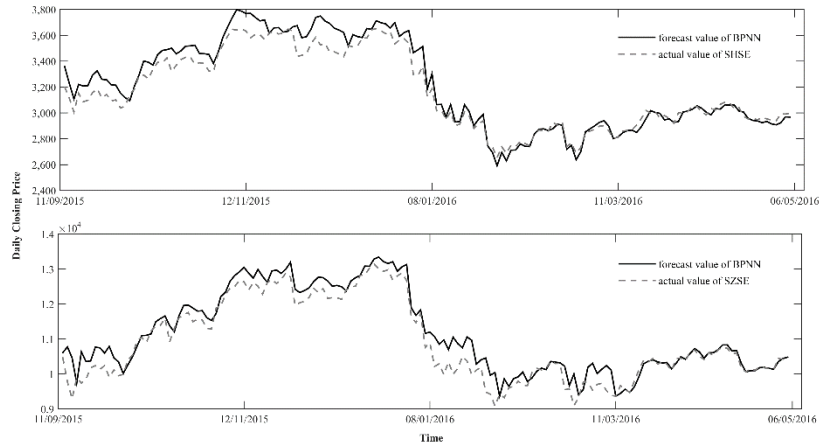

Figure 9. BPNN Performance of Composite Index in SHSE and Component Index in SZSE

3.5 Evaluate of performance of various models

Usually, three measures are used to evaluate the performance of models in forecasting daily closing prices as follows: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These measures are defined as:

$$RMSE = (\frac{1}{n}\sum_{t=1}^{n}(p_t - \hat{p}_{t-1})^2)^{1/2} \qquad (25)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|p_t - \hat{p}_t| \qquad (26)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{p_t - \hat{p}_t}{p_t}\right| \qquad (27)$$

207

_____

Where n denotes the number of forecasting samples, $\hat{p}_t$ denotes the forecasting daily closing price, and p_t denotes the real daily closing price. All the results are listed in table 1.

Table 1.Performance of Composite Index in SHSE and Component Index in SZSE

		RW	SVR	BPNN	GA-SVR-GM	GA-BPNN-GM
SHSE	RMSE	192.8566	85.1826	80.8020	65.0295	39.8736
	MAE	166.7401	57.6486	64.2551	48.1012	30.8181
	MAPE	0.0533	0.0195	0.0196	0.0161	0.01
SZSE	RMSE	1093.5	417.0634	377.3604	372.2870	171.5025
	MAE	861.7307	272.6167	268.8747	258.1719	128.0108
	MAPE	0.0843	0.0264	0.0254	0.0251	0.0121

As is the table shown, regardless of the measurement RMSE, MAE or MAPE, the classical RW model performs the worst in the Shanghai Stock Exchange or in the Shenzhen Stock Exchange, which means that the randomness of this model cannot accord to effective market theory and financial market can be predicted by past information. In addition, classical SVR and BPNN models perform better than RW model. However, BPNN model has a little better prediction performance than SVR model. When combining GA and GM with classical SVR or BPNN model so as to establish hybrid models, forecasting the movement of composite index in SHSE and component index in SZSE, three measurements of RMSE, MAE and MAPE are greatly minimized, and such improvement of models indeed enhances the prediction effect. Take RMSE for example, the GA-SVR-GM model is 20.1531 less than SVR model, and the GA-BPNN-GM model is 40.9284 less than BPNN model for sake of forecasting composite index in Shanghai. Similarly, GA-SVR-GM model enhances performance of SVR model by reducing 44.7764, GA-BPNN-GM model reducing 205.8575 to a large degree. Therefore, the hybrid GA-BPNN-GM model is superior to the hybrid GA-SVR-GM model, meaning that nonlinear and self-learning characteristic will be more accurate.

4. Conclusion

Accurate prediction of stock pricing index can assist investors in strengthening the judgment of investment. In this paper, an improving approach based on genetic algorithm and grey model is proposed to establish two hybrid models called GA-SVR-GM and GA-BPNN-GM separately. Through the validation of datasets of composite index and component index, two hybrid models indeed improve the performance compared with the classical random walk, support vector regression and back-propagation neural network. The results of three measurements of RMSE, MAE and MAPE demonstrate it. Therefore, such improvement proposed is effective, and the hybrid GA-BPNN-GM model performs the best among all the models. Thus, it is persuasive to use this model to predict the movement of stock price index.

Acknowledgments

REFERENCES

[1] **Alvisi, S., Franchini, M. (2012), Grey** *Neural Networks for River Stage Forecasting with Uncertainty;* Physica and Chemistry of the Earth, 42-44:108-118;

[2] **Arevalo, R., Garcia, J., Guijarro, F., Peris, A. (2017), A** *Dynamic Trading Rule Based on Filtered Flag Pattern Recognition for Stock Market Price Forecasting;* Expert Systems with Applications, 81: 177-192;

[3] **Cao, Q., Parry, M.E. (2009),** *Neural Network Earnings per Share Forecasting Models: A Comparison of Backward Propagation and the Genetic Algorithm;* Decision Support Systems, 47: 32–41;

[4] **Chen, Y.J., Chen, Y.M., Lu, C.L. (2017),** *Enhancement of Stock Market Forecasting Using an Improved Fundamental Analysis-based Approach;* Soft Computing, 21(13): 3735-3757;

[5] **Gu, J.R., Zhu, M.C., Jiang, L.G.Y. (2011),** *Housing Price Forecasting Based on Genetic Algorithm and Support Vector Machine***;** Expert Systems with Applications, 38: 3383-3386;

Jujie Wang, Danfeng Que

[6] **Hu, W.B., Yan, L.P., Liu, K.Z., Wang, H. (2016),** *A Short-Term Traffic Flow Forecasting Method Based on the Hybrid PSO-SVR;* Neural Processing Letter, 43: 155-172;

[7] **Huang, J.C., Huang, W.T., Chu, P.T., Lee, W.Y., Pai, H.P., Chuang, C.C., Wu, Y.W. (2017),** *Applying a Markov Chain for the Stock Pricing of a Novel Forecasting Model;* Communications in Statistics-theory and Methods, 46(9): 4388-4402;

[8] **Huang, W., Nakamori, Y., Wang, S.Y. (2005),** *Forecasting Stock Market Movement Direction with Support Vector Machine;* Computers & Operations Research, 32: 2513-2522;

[9] **Ince, H., Trafalis, T.B. (2017),** *A Hybrid Forecasting Model for Stock Market Prediction; Economic Computation and Economic Cybernetics Studies and Research; ASE Publishing;* 51(3): 263-280;

[10] **Jia, Z.Y., Ma, J.W., Wang, F.J., Liu, W. (2010),** *Characteristics Forecasting of Hydraulic Valve Based on Grey Correlation and ANFIS;* Expert Systems with Applications, 37: 1250-1255;

[11] **Kara, Y., Boyacioglu, M.A., Baykanl, O.K. (2011),** *Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange;* Expert Systems with Applications, 38: 5311-5319;

[12] **Li, H.F. (2017),** *Price Forecasting of Stock Index Futures Based on a New Hybrid EMD-RBF Neural Network Model;* Agro Food Industry Hi-tech, 28(1):1744-1747;

[13] **Okasha, M.K. (2014),** *Using Support Vector Machines in Financial Time Series Forecasting;* International Journal of Statistics and Applications, 4(1): 28-39;

[14] **Vapnik, V.N. (1999),** *An Overview of Statistical Learning Theory;* IEEE Transactions on Neural Networks, 10: 988-999;

[15] **Wang, J.J., Wang, J.Z., Zhang, Z.G., Guo, S.P. (2012),** *Stock Index Forecasting Based on a Hybrid Model;* Omega, 40: 758–766;

[16] **Zhou, Z.J., Hu, C.H. (2008),** *An Effective Hybrid Approach Based on Grey and ARMA for Forecasting Gyro Drift;* Chaos Solitons & Fractals, 35: 525-529.