Research Article

A CSI 300 Index Prediction Model Based on PSO-SVR-GRNN Hybrid Method

Jialin Chen 1 and Hanyin Yang 2

1 School of Finance, Southern University of Science and Technology, Shenzhen, Guangdong, China
2 Zhongnan University of Economics and Law, Wuhan, China

Correspondence should be addressed to Jialin Chen; 11912031@mail.sustech.edu.cn

Received 8 June 2022; Revised 27 June 2022; Accepted 5 July 2022; Published 4 August 2022

Academic Editor: Chia-Huei Wu

Copyright © 2022 Jialin Chen and Hanyin Yang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this article, a PSO-SVR-GRNN nonparametric hybrid model is proposed for the CSI 300 stock index to forecast the problem. Particle Swarm Optimization (PSO) is utilized to optimize the parameters of the SVR model to enhance the prediction ability of the support vector machine’s regression model for the original CSI 300 Index time series. The optimized residual sequence prediction results of the General Regression Neural Network (GRNN) are then used to optimize the time series prediction. The outcomes indicate that the PSO-SVR-GRNN model can greatly improve the prediction accuracy of the CSI 300 Index time series compared with individual models such as PSO-SVR, GRNN, GA-SVR, LSTM, PSO-LSTM, and SVR.

1. Introduction

Stock price indices are often compiled under a certain methodology based on the movement of stock prices during the corresponding period. As an effective measure of the capital market, the volatility of the stock index is usually observed as a barometer of changes in the national, social, economic, and political environment in which the market is based. After decades of development, the overall market capitalization of China’s capital market has increased rapidly, and its volume has developed significantly. By contrast, the fact that the Chinese stock market is not well regulated and the supervision of listed companies is incomplete has led to the high volatility of the Chinese stock market. This has a considerable impact on the Chinese stock market, which retail investors dominate, so predicting stock market movements is vital to warn investors of the risks. Furthermore, the stock index serves as a reflection of the stock market trend and as an underlying for financial innovation products such as index derivatives, achieving the establishment of the capital market in China. Accordingly, no matter for investors or regulators, the judgment of the stock index trend is of tremendous significance, so it has long been a concern by the majority of researchers. Nevertheless, the stock index is a reference for the movement of market stock prices and is influenced by diverse factors such as market sentiment. The time series created by the stock index has prominent nonstationary and nonlinear characteristics, which brings resistance to the advancement of research. This is why predicting stock index movements is still a popular research topic in the industry and academia. Currently, the research utilized to forecast stock index movements can be separated into four main categories: (1) general statistical models, (2) machine learning models, (3) combinatorial model algorithms, and (4) parameter optimization.

Researchers usually took statistical models to forecast stock price indices in the early days. Common statistical models comprise exponential smoothing models, ARIMA models (Ariyo et al. [1]), GARCH family of models, etc., (Wang et al. [2]). The autoregressive (AR) model is basically a model that operates past values and error terms of its series to forecast. Moreover, it is broadly applied for forecasting in economics for its robustness and better interpretability. However, stock markets are treacherous, and stock indices have powerful nonstationary and nonlinear features that are
not easy to be caught by ARIMA. As a result, statistical methods have a limited position in stock index forecasting.

Later, with the development of machine learning and deep learning techniques, numerous researchers experimented with artificial intelligence techniques to predict stock index movements. Support Vector Machine (SVM) is a widespread algorithm in machine learning. It has an extensive number of kernel functions available, is adjustable in managing nonlinear regression problems with high characteristics and has proven its predictive effectiveness in the financial field (Cao and Tay [3], Huang et al. [4]). Artificial Neural Networks (ANNs) are famous models in deep learning. As an influential adaptive method, ANNs are able to continually decrease the error of the objective function to approximate the actual value effectively. In addition, it is able to seize the non-linear part of the time series, which is highly attractive to researchers (Khashi and Bijari [5]). Regardless, a single SVM model has the disadvantage of being challenging to execute large-scale sample training and handle multiclassification issues. Besides, ANNs models are sensitive to parameter settings so they may be slightly less effective in generalization in the financial domain. Consequently, while adopting the above techniques, developing the advantages and avoiding the disadvantages is required to enhance the CSI 300 Index prediction effectively.

To address the drawbacks of single model forecasting, researchers have started to experiment with hybrid models and have reached some outcomes in the field of financial time series forecasting [6–8]. Pai and Lin [9] took full advantage of the special advantages of ARIMA and SVM models in forecasting stock price problems to create a hybrid ARIMA-SVM model and acquired good consequences in predicting stock prices. Zhang [10] and Khashi et al. [11] formed a hybrid ARIMA-ANN model and proved that the neural network approach is well suited to capture non-linear features such as the prediction of residual sequences filled with noise factors. Some literature suggests that generalized regression neural network (GRNN) models have more substantial predictive power than other ANNs models. Yan [12] combined several GRNN models with various properties and obtained the best outcomes in prediction with 60 different models. Hence, this article tries to fuse Support Vector Regressor (SVR) and Generalized Regression Neural Network (GRNN) in order to be able to maximize the distinct advantages of both models to fulfill the goal of enhancing the prediction accuracy.

Based on this, the application of integrating hybrid models with parameter optimization on time series forecasting is also more common. Khandelwal et al. [13] introduced separate linear and nonlinear components of time series by DWT (Discrete Wavelet Transform) and identified and forecasted new models applying ARIMA and ANN models, separately. The consequences revealed that the performance of the combined model incorporating the DWT technique was enhanced compared to both the single model and the ARIMA-ANN model. Oliveira and Ludermir [14] made an ETS-ARIMA-AR-SVR hybrid evolutionary system for time series volatility so that the prediction outcomes were remarkably advanced compared to the traditional models. In recent years, scholars have paid more and more attention to the superior performance of the PSO algorithm in parameter optimization. Qin et al. [15] formed a PSO-ARIMA-DBN model for red tide biomass time series data, which can nicely seize the temporal correlation, spatial heterogeneity, and other nonlinear features between red tide biomass and environmental factors. In a study of traffic flow prediction, Hu et al. [16] observed that the SVR model optimized with PSO parameters could get more accurate predictions than other models. Hasanipanah et al. [17] applied the PSO-SVR model to the AO prediction field and acquired the same conclusion. Therefore, this report will adopt the particle swarm algorithm (PSO) to optimize the model parameters to reach more accurate model prediction results.

As seen previously, previous literature has proved the effectiveness of “artificial intelligence techniques,” “hybrid models,” and “parametric optimization” in the time series forecasting and has acknowledged the potential of PSO algorithms for parametric optimization and the excellent performance of SVR and GRNN models in seizing financial time series features. Therefore, this article forms a PSO-SVR-GRNN hybrid model to forecast the CSI 300 stock index. First, the support vector regression (SVR) parameters are optimized with PSO parameters could get more accurate predictions than other models. Hasanipanah et al. [17] applied the PSO-SVR-GRNN hybrid model to forecast the CSI 300 stock index time series using the PSO-SVR model. Then, a generalized regression neural network (GRNN) model is constructed based on the residual series. Ultimately, the prediction values of the PSO-SVR-GRNN hybrid model for the CSI 300 stock index are reached. The main contributions of this paper are as follows:

1. Motivated by the potential of the PSO algorithm in parameter optimization and the incredible performance of SVR and GRNN models in seizing financial time series features, this article makes the first PSO-SVR-GRNN hybrid model and uses it in the field of stock index forecasting

2. Experiments reveal that the PSO-SVR-GRNN model proposed in this report greatly optimizes the single model and the base model in forecasting the CSI 300 stock index

2. Establishment of the PSO-SVR-GRNN Model

The establishment of the combined PSO-SVR-GRNN model is separated into the following processes.

2.1. Description of PSO-SVR-GRNN Process

1. Acquiring and inputting the calendar year daily HSS 300 data.

2. Apply MSE (mean square error) as the fitness evaluation index and split the time series into a training set and test set, respectively, utilizing the five-fold cross-check method. Subsequently, the particle swarm algorithm is adopted to find the optimal RBF kernel function parameters, insensitive
loss parameters, and penalty parameters C of the time series corresponding to the SVR model.

(3) Predicting the test set by the PSO-SVR model under the optimal parameters calculated in the previous step.

(4) Outputting the residual sequences after the training of the PSO-SVR model and taking the GRNN model to correct the residual sequences.

(5) Set up a combined PSO-SVR-GRNN model, train, and output the CSI 300 Index prediction outcomes.

The flow of the PSO-SVR-GRNN model building is shown in Figure 1.

2.2. Particle Swarm Algorithm. Eberhart and Kennedy first presented the particle swarm algorithm [18] in 1995. The concept of this algorithm is derived from the study of birds foraging behavior, by mimicking the foraging behavior of a flock of birds to discover the solution that makes the population optimal,

\[ v_{id}^{k+1} = v_{id}^{k} + c_1 \text{rand}(p_{best}^{k}_i - x_{id}^{k}) + c_2 \text{rand}(g_{best}^{k} - x_{id}^{k}) \]

\[ x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \cdot \]

In this algorithm, the parameters of each particle are its position \( X_i(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{id})^T \) velocity \( V_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{id})^T \), and fitness evaluation index. The original state is a flock of particles in a random position and with a random velocity (initial solution). Every particle updates the optimal value at each iteration by tracking the extreme points \( p_{best}^{k}_i \) of its trajectory and the extreme points \( g_{best}^{k} \) of the particles in the swarm, where \( v_{id}^{k} \) is the velocity of the particle at the kth latitude in the dth iteration, and \( x_{id}^{k} \) is the position of the particle \( i \) at the kth latitude in the dth iteration. Furthermore, \( C1 \) and \( C2 \) are the maximum step length of the particle \( i \) toward its historical optimal value point and the optimal value point in the population in the kth iteration, individually. If \( v_{dmax} \) is too huge, it will cause the algorithm to fall into a local optimum, so it is essential to manage the maximum step sizes \( C1 \) and \( C2 \) in a reasonable range. At the end of each iteration, every particle’s fitness is assessed, the flow of the particle swarm algorithm is shown in Figure 2.

2.3. SVR Model. In the 1990s, Vapnik [19] presented a support vector machine (SVM) model under statistical learning theory. After that, scholars took SVM to the field of regression prediction by proposing \( \epsilon \) insensitive loss function to build a support vector regression (SVR) model; SVR is mainly adopted for time series prediction and optimization control, etc. The principle method is as follows.

In formula (1), given a set of training samples \( X = \{ (x_i, y_i) | x_i \in \mathbb{R}^n, \ y_i \in \mathbb{R}, \ i = 1, 2, \ldots, N \} \) (where \( X_i \) is the input feature vector, \( Y_i \) is the target value, \( n \) is the number of samples, and SVR maps the low-dimensional nonlinear data \( \phi \) to the high-dimensional space by a non-linear function, followed by linear regression, to acquire \( f(x) \).

\[ f(x) = \omega^T \phi(x) + b \in \mathbb{R}^n, \quad b \in \mathbb{R}. \]  

(2)

Here, \( \omega \) and \( b \) are the regression coefficients and intercept terms, respectively. The idea of SVR is to find the optimal estimate of these two parameters so that the deviation between the function value corresponding to the training data \( f(x) \) and the actual value \( y_i \) is less than \( \epsilon \), and the corresponding optimization problem is to discover the optimal value of the function. The corresponding optimization problem is to find the minimum value of the following equation:

\[ \min: \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \]

\[ \left\{ \begin{array}{l}
 y_i - f(x) \leq \epsilon + \xi_i \\
 y_i - f(x) \geq \epsilon + \xi_i^* \\
 \xi_i, \xi_i^* \geq 0, \quad i = 1, 2, \ldots, N.
\end{array} \right. \]

(3)

Here, \( C > 0 \) is the penalty parameter, \( \xi_i, \xi_i^* \) is the positive slack variable, and \( \epsilon \) represents the insensitive loss function. The Lagrangian transformation is performed for the pairwise problem of the above problem, and the following equation is reached:

\[ L = \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) - \sum_{i=1}^{N} \lambda_i (\epsilon + \xi_i - y_i + f(x_i)) - \sum_{i=1}^{N} \nu_i \xi_i + \nu_i^* \xi_i^* \]

(4)

Here, \( \lambda_i, \lambda_i^*, \nu_i, \nu_i^* \geq 0, i = 1, 2, \ldots, N \), the extreme value of the function is managed by the pairwise principle, so that its partial derivative concerning the variable \( \omega, b, \xi_i, \xi_i^* \) is 0.
The kernel function $K(x, x_i) = \phi(x) \cdot \phi(x_i)$ is proposed to replace the inner product operation on the high-dimensional space, and the original regression function (1) is reduced to the following equation:

$$f(x) = \sum_{i=1}^{N} (\lambda_i - \lambda_i^*) K(x, x_i) + b. \quad (5)$$

Here, the usual kernel function selection contains linear kernel function, polynomial kernel function, and Gaussian radial basis kernel function (RBF), etc. From previous experience, the RBF kernel function is best applied in the absence of prior knowledge of the sample data, so in this paper, RBF is chosen as the kernel function, namely,

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\sigma^2}\right), \quad \sigma > 0. \quad (6)$$

Here, $\sigma$ denotes the bandwidth, and thus the original SVR model is explicitly an optimization problem solving for the bandwidth $\sigma$, insensitivity coefficient $\varepsilon$, and penalty parameters $C$.

### 2.4. GRNN Model

#### 2.4.1. Introduction of the Model Algorithm.

Generalized regression neural network (GRNN) is a kind of radial basis neural network presented by Sprecht. It has better nonlinear mapping ability and learning speed and is a generally utilized nonparametric regression model. A common nonparametric regression model can accomplish reasonable prediction results even with small sample data, GRNN largely includes a radial basis network layer and a linear network layer. Its MATLAB function-based network structure is shown below in Figure 3: (Abdollahi and Ebrahimi [20]):
There are no diagrams, tables, or images to translate.
population quantity was 20. The fitness curve of the PSO-based parameter optimization is illustrated in Figure 6.

According to Table 1, the penalty parameter C of the SVR model, the parameter $g$ of RBF kernel function, and the insensitive loss function $p$ are 2.8141, 0.0100, and 0.0100, respectively, with MSE less than 0.01. It can be judged that the three parameters obtained in the PSO method have strong fitness.

3.3. PSO-SVR Model Prediction. Based on the above-given three parameters optimized by PSO, an SVR model was established to predict the time series of CSI 300, taking the latter 20% of the trading days as the test set. The final true value and predicted value are shown in Figure 6.

As revealed from the evaluation indices of the PSO-SVR model fitted the test set (Table 2), the MSE is only 12528, showing a good fitting effect.

3.4. GRNN Model-Based Error Correction. First, the residual series of the PSO-SVR model were obtained (Figure 8).

Then, the residual series were input into the GRNN model. The 5-fold cross validation method was used to divide the dataset into a training set and a test set, obtaining the optimum smoothing factor of 0.6. As observed in Figure 9, the GRNN model with the best parameters was used to fit the residual series.

Compared with the PSO-SVR model, the fused model subjecting to residual correction by GRNN was significantly improved in effect. The RMSE decreased from 111.93 to 111.24 and the MSE from 12528–12374 (Table 3).

4. Comparison

4.1. Contrast Test with the GRNN Model. The original series of CSI 300 indices were treated as the dataset and divided into a training set and test set using a 5-fold cross validation method. Followed by, the optimum smoothing factor was obtained, 0.1, and input into the GRNN model parameter to fit the original time series.

As discovered in Figure 10, the GRNN-predicted results of the original sample comply with the actual data, presenting a good fitting effect. However as displayed in Table 4, the prediction effect of GRNN is inferior to the combined model.

4.2. Contrast Test with the LSTM Model. The original series of CSI 300 indices were treated as the dataset and divided into a training set and test set based on the above-given contrast test design. Then, the LSTM model parameter was used to fit the original time series.

As can be seen from Figure 11, the LSTM-predicted results of the original sample conform to the actual data, demonstrating a good fitting effect. However as displayed in Table 4, the prediction effect of LSTM is inferior to the combined model significantly.

4.3. Contrast Test with the SVR Model. Also, the original series of CSI 300 indices were treated as the dataset and further divided into a training set and a test set based on the said contrast test design. Then, the SVR model parameter was used to fit the original time series. Figure 12 displays the optimization results of the SVR parameter and Figure 13 presents the SVR’s fitting to the original series.

As exhibited in Figure 13, the SVR-predicted results of the original sample accord with the actual data, which presents a good fitting effect. However as implied in Table 4, the prediction effect of SVR is still inferior to the combined model.

4.4. Contrast Test with the PSO-LSTM Model. The original series of CSI 300 indices were divided into a training set and a test set, with the same design as the aforesaid test. Then, the
**Table 1:** The results of parameters optimized by PSO.

<table>
<thead>
<tr>
<th>Sequence of information granulation parameter</th>
<th>Bestc</th>
<th>Best_g</th>
<th>Best_p</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2.8141</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

**Figure 6:** MSE changes in the fitness curve during the PSO-based parameter optimization.

**Table 2:** PSO-SVR model effect.

<table>
<thead>
<tr>
<th>Evaluation index</th>
<th>RMSE</th>
<th>MSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index value</td>
<td>111.93</td>
<td>12528</td>
<td>2.7193</td>
<td>76.0757</td>
</tr>
</tbody>
</table>
PSO-LSTM model parameter was used to fit the original time series. The PSO-LSTM’s fitting to the original series is shown in Figure 14.

According to Figure 14, the PSO-LSTM model-predicted results of the original sample tally with the actual data, displaying a preferable fitting effect. However, the prediction effect of the PSO-LSTM model is still inferior to the combined model as clearly shown in Table 4.

4.5. Contrast Test with the GA-SVR Model. Taking the original series of CSI 300 indices as the dataset, the training set and the test set were also divided into with the same design as the aforesaid test. Furthermore, the GA-SVR model parameter was used to fit the original time series. The optimization results of the GA-SVR parameter are displayed in Figure 15 and the GA-SVR’s fitting to the original series is demonstrated in Figure 16.

As revealed in Figure 16, the GA-SVR model-predicted results of the original sample correspond to the actual data, demonstrating a sound fitting effect. However as observed in Table 4, the prediction effect of this model is also weaker than the combined model.

Table 3: Comparision between the effects of the PSO-SVR model and the PSO-SVR-GRNN model.

<table>
<thead>
<tr>
<th>Evaluation index</th>
<th>RMSE</th>
<th>MSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSO-SVR</td>
<td>111.93</td>
<td>12528.00</td>
<td>2.7193</td>
<td>76.0757</td>
</tr>
<tr>
<td>PSO-SVR-GRNN</td>
<td>111.24</td>
<td>12374.00</td>
<td>2.4387</td>
<td>75.8048</td>
</tr>
</tbody>
</table>

Table 4: Comparison of the models’ effects.

<table>
<thead>
<tr>
<th>Evaluation index</th>
<th>RMSE</th>
<th>MSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSO-SVR</td>
<td>111.93</td>
<td>12528.00</td>
<td>2.7193</td>
<td>76.0757</td>
</tr>
<tr>
<td>PSO-SVR-GRNN</td>
<td>111.24</td>
<td>12374.00</td>
<td>2.4387</td>
<td>75.8048</td>
</tr>
<tr>
<td>GRNN</td>
<td>128.33</td>
<td>16470.79</td>
<td>4.2362</td>
<td>96.9851</td>
</tr>
<tr>
<td>LSTM</td>
<td>123.14</td>
<td>15165.78</td>
<td>3.0256</td>
<td>83.1749</td>
</tr>
<tr>
<td>SVR</td>
<td>110.75</td>
<td>12266.14</td>
<td>2.6367</td>
<td>73.9106</td>
</tr>
<tr>
<td>GA-SVR</td>
<td>110.43</td>
<td>12197.00</td>
<td>2.6556</td>
<td>74.4111</td>
</tr>
<tr>
<td>PSO-LSTM</td>
<td>147.46</td>
<td>21744.65</td>
<td>3.2696</td>
<td>109.6961</td>
</tr>
</tbody>
</table>

4.5. Contrast Test with the GA-SVR Model. Taking the original series of CSI 300 indices as the dataset, the training...
5. Conclusion

Considering the inaccuracy of the CSI 300 index, relevant studies are analyzed and a combined CSI 300 index prediction model is proposed based on the PSO-SVR-GRNN.

First, this model has a short predictive period as it is established as per the historical data of the CSI 300 index. For a long-term prediction of the CSI 300 index, it is needed to further study how to improve the model by introducing economic change and other external factors. The complementarity of different prediction methods, the direct fusion of different methods, and how to improve the accuracy and reliability of the CSI 300 index prediction model are also the development trends in future studies on the prediction.
Second, accurate prediction of the CSI 300 index is a difficulty in many studies and is significant for the government to grasp the future macro-economic situation and trend. As demonstrated by contrast tests, the MSE, RMSE, MAPE, and MAE of the proposed model’s prediction results are smaller than other models. Hence, the proposed model’s prediction results are more accurate. This model combines the advantages of the PSO-SVR model and the GRNN model and was also used for actual prediction. The obtained prediction results varied in identical scope and trend to the actual data, which verifies the effectiveness and practicability of this model.

Third, the combined nonparametric model has more accurate and effective prediction results than the single model, as discovered over a comparison between the PSO-SVR model and the PSO-SVR-GRNN model. This result also indicates that the CSI 300 index series practically has complicated nonlinear structural features.

In further studies, more input data such as macro data, investor sentiment, and fundamental information will be used to increase the prediction accuracy. Moreover, PSO or GA algorithms will be used to find the best parameters of GRNN to optimize the existing model. Finally, the model will be extensively evaluated using various stock test data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


