A PCA-LSTM Model for Stock Index Prediction
De-hua LIU¹,* and Jing-jing WANG²

¹School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
²School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China

*Corresponding author

Keywords: Financial market, LSTM model, Principle component analysis, Stock prediction model.

Abstract. This paper proposed a LSTM network model to predict stock index closing price. During the research process, we noticed the multicollinearity of the variables in the volume-price information and solved it by using PCA principal component analysis. A stock index closing price forecasting model based on historical stock index price and volume-price information was constructed. The empirical study of the CSI300 index data in the model shows that the generalization ability of LSTM network after PCA processing is better than that of normal LSTM network and normal BP network. The prediction of training set and test set is more reliable.

Introduction

The stock index price trend can reflect the current macroeconomic and financial market development. There are three main methods for stock index forecasting: one is based on historical price information and technical analysis of trading indicators to achieve stock market trend forecasting; the other is based on time series method [1], including linear model and nonlinear model; Meanwhile, the application of machine learning algorithms in financial data analysis is developing rapidly, thanks to its nonlinear fitting ability superior to traditional methods, and the role of neural network, SVM and other regression algorithms in stock index price prediction is verified by many researchers [2-3], the LSTM network evolving from the recurrent neural network performs well in time series prediction [4]. However, these algorithms often fall into local optimal solutions and oscillations during training. Some scholars have also explored such problems, and proposed BP neural network based on principal component analysis and SVM support vector machine for predictive problem research [5], to overcome the above shortcomings by reducing the input data dimension. The principle component analysis method has been verified in the existence of redundancy and duplication in the stock price information, but the research on the combination of principal component analysis and LSTM network for stock price forecasting is still relatively rare.

From the perspective of preventing over-fitting and improving generalization ability, this paper attempts to combine the LSTM (long-short-term memory) network that can make full use of historical information with the principal component analysis method to construct a stock index closing price forecasting model with high accuracy.

Research Theory

Principle Component Analysis

Principle Component Analysis is a commonly used method in statistical analysis and econometric analysis. This method is mainly used when data variables have strong correlation and information overlaps, PCA can establish a set of variables that are dimension-reduced on the basis of the original data. The variables after decomposition are as irrelevant as possible and can save the information in the original data to the maximum extent. The preprocessed variables are the principal components of the original data. The measure of multicollinearity is usually calculated by calculating the Variance...
Inflation Factor (VIF) of each variable, that is, the variable can be linearly represented by the remaining variables. VIF can be expressed as:

\[ VIF_i = \frac{1}{1-R_i^2} \]  

(1)

Where \( R_i^2 \) is the goodness of fit. Usually, the principal component is selected according to the order of the contribution rate of the feature vector, and the sum of the contribution rates of the selected principal components is recorded as the cumulative contribution rate, and the cumulative contribution rate can reflect the inclusion of the selected principal component set for the original information.

**LSTM Network**

The LSTM (Long Short Term Memory) network is improved on the recurrent neural network RNN. It replaces the neurons in the hidden layer of the traditional neural network with memory cells and computational units. This structure makes LSTM able to use historical information efficiently in time series to achieve reliable dynamic prediction. At the same time, the gradient vanishing in the circulating neural network is effectively improved on the LSTM [6]. Figure 1 shows the memory cell structure in the LSTM network. The input gate, output gate, and forget gate in the module are nonlinear summary units containing excitation functions. They control the excitation of cell cells through three nodes (forked circles in the figure). Forget gates usually use the logistic sigmoid function to ensure that the value is between 0 and 1 (indicating that the gate is closed and the gate is open). The activation functions of the input and output gates are generally tanh or logistic sigmoid functions.

\[ f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \]  

(2)

\[ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \]  

(3)

\[ c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \]  

(4)

\[ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \]  

(5)
\begin{equation}
    h_t = o_t * \tanh(c_t)
\end{equation}

where \( \sigma \) is the sigmoid activation function; \( h_{t-1} \) the output of the previous memory cell is used as the input of the latter module, thus affecting \( h_t \) the output of the latter module. The processing of the layer memory cells is then selectively returned to \( h_t \).

Figure 2 is a schematic diagram of the working principle of the entire LSTM layer after the memory cell is connected, showing the transfer process of historical information; \( A \) is the memory cell described above; \( x_t \) is the time sequence of the input variable; \( h_t \) is the output value of the memory module. The LSTM system has good nonlinear fitting ability and can fully consider the historical information of the data. Therefore, the LSTM network can be established to train through historical data to predict the closing price of the stock index.

**PCA-LSTM Prediction Model**

LSTM has a good ability to combine historical information to predict time series, but we also notice that there are many parameters in each Cell, and there is a risk of over-fitting. By combining with PCA method, the network scale can be reduced, thus improving the pan of the network. On the other hand, PCA is able to reduce feature dimensions while ensuring independence between variables and reserving the original information. By combining PCA with LSTM, the generalization ability of the model can be improved, and over-fitting is restrained. The model implementation process is as follows:

1. Data set is divided into training set, test set, and then the training data is standardized. This paper uses Z-Score transform; the transform is based on the mean \( \mu \) and standard deviation \( \sigma \) of the data.
2. The variance inflation factor of each variable is calculated to analyze whether there is multicollinearity; if the variable has multicollinearity, the PCA is performed to reconstruct the data set; for the test data, we use the same transformation formula for processing the training data.
3. Constructing a three-layer LSTM network. The first input layer is the input data after inputting the reconstructed format; the second layer is the LSTM layer of 20 memory cell; the third layer is the fully connected layer of 10 nodes. The activation function is ReLU (linear rectification function, which trains the convergence fast; the fourth layer is the output layer.
4. The adaptive learning rate adjustment algorithm Adam [7] is selected, and the MSE mean square error is used as the loss function, and the equation (7) is the mean square error.

\[
    MSE = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)}{n}
\]

**Experiments and Analysis**

This paper selects the opening price, closing price, lowest price, highest price, trading volume, turnover and rate of change of the CSI 300 Index from January 4, 2005 to October 9, 2017, and constructs a prediction model. The stock index price information of the previous day predicts the closing price of the stock index on the following day. The data from January 4, 2005 to April 22, 2015 was used as the training data set, and the data from April 23, 2015 to October 9, 2017 was used as the test data set. The selected time period is based on the following considerations: The time period contains the bull market and the bear market in several stages of the A-share, and the data samples are
representative. Firstly, the multicollinearity in data is analyzed, and the VIF (variance inflation factor) of each variable after decomposition and the CR (Contribution Rate) of each variable are calculated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Open</th>
<th>Close</th>
<th>Low</th>
<th>High</th>
<th>Turnover</th>
<th>Volume</th>
<th>%Chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>31080</td>
<td>3885</td>
<td>2304</td>
<td>3491</td>
<td>24.76</td>
<td>33.44</td>
<td>2.83</td>
</tr>
<tr>
<td>CR</td>
<td>0.6796</td>
<td>0.1772</td>
<td>0.1402</td>
<td>0.0028</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

From Table 1, we can find that the VIF of the four variables is >100, and the VIF of the three variables is >10. Therefore, it can be inferred that the seven variables have severe multicollinearity. This paper selects the first four principal components, and the cumulative contribution rate reaches: 0.9998, which basically contains all the price information. The input data of the training set after PCA transformation is: \(X_{1t}; X_{2t}; X_{3t}; X_{4t}\), which is used as the input of the LSTM network. The initial learning rate during training is set to 0.002, and the number of training times is 500. Figure 3 shows the comparison of the prediction results and real values of the training set and test set of the network, and the change of the loss function MSE in 500 epochs.

<table>
<thead>
<tr>
<th>Network</th>
<th>Normal BP network</th>
<th>LSTM network</th>
<th>PCA-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE(train)</td>
<td>36.5390</td>
<td>35.4312</td>
<td>31.5753</td>
</tr>
<tr>
<td>MSE(test)</td>
<td>41.6132</td>
<td>39.8376</td>
<td>33.1794</td>
</tr>
</tbody>
</table>

Figure 3. The training results of PCA-LSTM. Figure 4. The relative error rate in test data set of three models.

In order to verify that PCA-LSTM can improve performance, we set up two sets of comparisons: 1: The input data is not subject to Principal Component Analysis and dimension reduction LSTM network; 2: The same is not dimension reduction, the number of layers and the number of nodes are the same without LSTM Layer of normal BP network. They are trained the same Adam algorithm and the same initial learning rate and number of trainings. After training, the mean square error MSE of the training sets of the three networks is shown in Table 2. The relative error rate of the test set prediction results is shown in Figure 4.

As shown in Figure 4, we can find that the relative error rate of the normal BP network without LSTM layer and the LSTM network without PCA processing in the test set 0-200 period (April 23, 2015 to February 17, 2016) Large, LSTM is slightly better than the normal BP network, but neither can reliably predict the closing price; while PCA-LSTM can better predict the closing price during this period, the relative error rate is lower; in the test set 300-600 period (From July 12, 2016 to October 9, 2017), all three networks were able to predict well and the relative error was stable. We may analyze the above situation from the fact that A shares suddenly changed from “bull” to “bear” at the end of May 2015. The generalization of the two network models in the control group is not sufficient to predict the situation, and the PCA-LSTM network. The generalization ability is better than the control group; from July 12, 2016 to October 9, 2017, the A-share trend is generally stable, and all three networks can reliably predict. It can be seen that the PCA-LSTM network can reduce the over-fitting situation and the generalization ability is better than the LSTM and normal BP neural network without PCA processing.
Figure 5 shows the MSE square error of the loss function of the training set and the test set during the 500 training epochs of the three networks. We can find that the BP network without LSTM layer has the fastest convergence rate; the convergence speed of LSTM and PCA-LSTM is slower. This result can be understood because of the existence of the memory cell in LSTM increases the complexity of the model, but the convergence speed does not slow down too much. Meanwhile, 500 training epochs is enough for LSTM network training to converge, and the performance of PCA-LSTM is reliable. 5. Acknowledgement

Conclusion

This paper takes the stock price index forecast as the research object and uses the machine learning algorithm. Firstly, PCA is used to process the multicollinearity of the stock price information data, and combined with the LSTM network, a PCA-LSTM network stock index prediction model based on the previous day's volume price information data as the input variable and the closing price of the next day as the output variable is constructed and compared with the comparison model. The relative error rate, mean square error and convergence speed of the prediction results are analyzed. It is concluded that the LSTM network combined with principal component analysis has better generalization ability and can predict stocks more reliably. The conclusion of the price index trend. Although the prediction results are better than other models, we also notice that the LSTM network is more complex and requires more training times to converge. The next step will be to study the optimization method combined with the model to improve the network convergence speed.

References