ISSN 2616-5902 Vol. 4, Issue 3: 42-47, DOI: 10.25236/AJBM.2022.040308

Research on Stock Price Prediction Based on PCA-LSTM Model

Ziheng Zhang*

School of Ande, Xi'an University of Architecture and Technology, Xi'an, 710311, China
**Corresponding author: 2724301800@qq.com*

Abstract: *When technical methods are used to establish LSTM stock prediction model, traditional methods often lead to poor generalization and poor prediction effect due to many input data variables selected, data information overlap, outliers have a great influence on training and other factors. To solve these problems, this paper proposes to use principle component analysis method to reduce the dimension of basic data, combine with stock related technical indicators KDJ and MACD as input data, and make prediction after adjusting the model according to stock characteristics. The experimental results show that the PCA-LSTM model reduces the average error of prediction, greatly reduces the running time, improves the stability of prediction, and more accurately predicts the closing price of Yaxing, which has application value.*

Keywords: *LSTM model, Stock price prediction, Neural network, Principal component analysis*

1. Introduction

With the rapid development of China's economy and people's economic consciousness improve, financial management behavior has gradually become the majority of families to achieve asset appreciation is one of the mainstream methods. Stock as a kind of free maturity has negotiable securities, with its high risk, high yield characteristics become many families' primary choice of financial investment. Therefore, it is of great significance for the country and the family to predict the trend of the stock market by reasonable analysis of its fluctuation characteristics and existing specific rules.

The forecasting of financial stocks has been a hot research topic in the field of finance, and in terms of methods, it can be broadly classified into linear forecasting models and nonlinear forecasting models. Among them, linear forecasting models mainly include moving average autoregressive model (ARIMA), generalized autoregressive conditional heteroskedasticity model (GARCH)[1], exponential generalized autoregressive conditional heteroskedasticity (EGARCH)[2] and generalized autoregressive conditional heteroskedasticity of integer product (IGARCH) [3]. With the rapid development of computer technology and the increasing sophistication of deep learning research[4], neural networks in the field of machine learning are increasingly used in stock forecasting and have achieved more efficient and accurate forecasting results than linear forecasting models. Cao[5] used BP neural networks and gray GARCH-BP models to predict significantly more accurately than GARCH models. Deng[6] proposed a DAE-BP model to perform DAE dimensionality reduction for stocks and then used BP neural network for stock price prediction, and achieved good prediction results. Unfortunately, the model structure is too homogeneous.

A good stock forecasting model needs good generalization ability. In order to analyze the problem more comprehensively and systematically, many indexes must be considered. But too many variables will undoubtedly increase the difficulty and complexity of the analysis. The correlation among the basic data of stock makes the information reflected by the data partly overlap, and can not show the underlying law well. In order to solve this problem, this paper improves the model based on the LSTM model[7-8] and the relevant technical indexes[9] which are more sensitive to the stock price. Principal Component Analysis (PCA) is introduced to extract the incoherent principal components, and less principal components are used to replace more variables as the input data of the training model while preserving the original data to the greatest extent. Reduce model complexity and improve learning rate.

2. Data processing

Through the stock data station of various financial websites and the tushare financial data interface package of Python, we got nine basic data of the selected stocks: The opening price, the closing price, the highest price, the lowest price, the pre-closing price, the rise and fall amount, and the rise and fall extent, volume and value of transactions. As shown in Table1, that will be calculated for the basic data KDJ, MACD indicators together as a model of training data. The selected stock is ST Yaxing, stock code is 600319.SH.Only 12 stock datas are shown here due to article space limitation.

Table 1: Basic data on stocks (ts_code: 600319.SH)

	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
0	20210308	5.15	5.2	5.05	5.09	5.14	-0.05	-0.9728	14443.01	7362.059
1	20210305	5.16	5.22	5.11	5.14	5.15	-0.01	-0.1942	17961.54	9236.323
2	20210304	5.19	5.23	5.15	5.15	5.2	-0.05	-0.9615	12382	6420.943
3	20210303	5.26	5.26	5.19	5.2	5.24	-0.04	-0.7634	7155.3	3725.668
4	20210302	5.24	5.29	5.17	5.24	5.26	-0.02	-0.3802	10133	5287.319
5	20210301	5.22	5.33	5.2	5.26	5.2	0.06	1.1538	11740.3	6168.485
6	20210226	5.23	5.28	5.12	5.2	5.21	-0.01	-0.1919	11766	6121.305
7	20210225	5.26	5.34	5.2	5.21	5.31	-0.1	-1.8832	16700	8758.157
8	20210224	5.34	5.44	5.2	5.31	5.3	0.01	0.1887	33575.04	17844.007
9	20210223	5.07	5.3	5.03	5.3	5.05	0.25	4.9505	28364.39	14793.624
10	20210222	5	5.09	4.9	5.05	5	0.05	1	19556.97	9814.468
11	20210219	4.99	5.06	4.86	5	5	0	0	16187	8008.387
12	20210218	5	5.08	4.97	5	5.07	-0.07	-1.3807	16815	8438.848

3. LSTM model based on principal component analysis

3.1. Obtaining training data

KDJ index is a kind of sensitive and fast technical analysis index, which mainly uses the real fluctuation of stock price to reflect the strength or weakness of the price change trend, and can send out buying and selling signals before the stock price rises or falls. The immature random value R SV of the last day of the cycle is calculated by the highest price, the lowest price, the closing price at the end of the cycle and the proportional relationship among them, then, K, D and J values are calculated according to the moving average method. Most of these parameters are averaged over several days, and for the first few days of the forecast, these averages are calculated by taking the average of the highest and lowest prices of the day to find the middle price.

The K value is the n-day moving average of r SV, and the K line connected by the K value is also called the fast line. The D value is the n-day moving average of the K value, and the D line is the slowest among the 3 lines, which is called the slow line. J value changes the fastest, as an auxiliary observation K line and D line issued by the buy and sell signal, J line is known as the ultra-fast line or confirmation line. Three lines in the same coordinates on the composition of the price fluctuations can reflect the trend of the KDJ index.

$$\begin{cases} RSV=(C_n-L_n)/(H_n-L_n)\times 100 \\ K=\frac{2}{3}K_p+\frac{1}{3}RSV \\ D=\frac{2}{3}D_p+\frac{1}{3}K \\ J=3\times \mathrm{K}-2\times D \end{cases} \tag{1}$$

In the formula (1): C_n is the closing price on the th day; L_n is the lowest price in n days. H_n is the highest price in n days; K_P and D_P are the K and D values of the previous day, or equals 50 if none exists.

The MACD is called the convergence-divergence moving average. The convergence and separation of the two moving averages represent changes in market trends and are a common technical indicator for stocks. The moving average EMA of fast and slow speed is usually selected 12 and 26 days, and the MACD is calculated by their 9-day moving average of the difference DIF and the difference Dif.

$$\begin{cases} EMA_{(n)} = \frac{n-1}{n+1} \times PEMA_{(n)} + \frac{2}{n+1} C \\ DIF = EMA_{(12)} - EMA_{(26)} \\ DEA = PDEA_n \times \frac{n-1}{n+1} + DIF \\ MACD = (DIF - DEA) \times 2 \end{cases} \tag{2}$$

In the formula (2), *n* is the moving average number of days, *C* is the closing price of the day, PEMA and PDEA are the EMA and DEA of the previous day.

3.2. Dimensionality reduction by principal component analysis

Principal Component Analysis (PCA) is a method that transforms a number of related raw data into a small number of unrelated linear combinations without changing the structure of the sample data. In order to reduce the dimension and simplify the complex multi-dimensional problems, we use less variables to replace more variables while reflecting the information of the original data to the greatest extent.

To extract the principal component, the original data must be standardized, that is, the mean of the strain is subtracted and then divided by the variance to eliminate the influence of different dimensions.

$$\begin{cases} Y_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{S_j} \\ i = 1, 2, ..., m; j = 1, 2, ..., n \end{cases} \tag{3}$$

Then the Correlation Coefficient Matrix R is calculated and the eigenvalue (i = 1,2, ... , n) is obtained by solving the characteristic equation $|\lambda E - R| = 0$.

$$\begin{cases} R = \left[\frac{1}{m-1} \sum_{k=1}^{n} Y_{ti} . Y_{ij} \right] \\ i = 1, 2, ..., n; j = 1, 2, ..., n \end{cases} \tag{4}$$

The eigenvalue λ_i is the variance of the principal components. Used to describe the information contained in the direction of the corresponding feature vector. The magnitude of the feature value directly reflects the influence of each principal component. The contribution to variance of an eigenvector is obtained by dividing the value of an eigenvalue by the sum of all eigenvalues.

$\frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k}$ is the *ith* principal component. $\frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{n} \lambda_k}$ is the cumulative contribution rate of the previous *i* principal components. Ensure that the selected principal component contains most of the information from the original data. Finally, the load value of the principal component is calculated and the score value of the principal component is obtained as the new training data.

$$\begin{cases} T_{ij} = p(z_i, x_j) = \sqrt{\lambda_i} a_{ij} \\ i = 1, 2, ..., m; j = 1, 2, ..., n \end{cases} \tag{5}$$

3.3. Prediction using the LSTM model

An artificial neural network for long term and short term memory, LSTM is a time recurrent neural network that is suitable for processing and predicting important events with relatively long intervals and delays in time series.

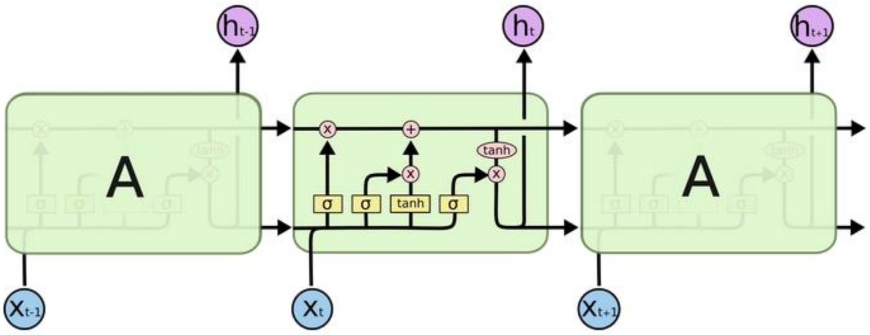

Figure 1: LSTM cellular structure

The LSTM sets up three gates in a cell: The forgetting gate, the input gate, and the output gate. Once a piece of data enters the LSTM network, it is determined according to the rules, those that conform to the algorithm rules will be left behind, and those that don't will be forgotten through the forgotten door. Only the information that conforms to the algorithm authentication will be left behind, and the information that does not conform will be forgotten through the forgetting door.

$$f_t = \sigma(W_f . \left[h_{t-1}, x_t\right] + b_f) \tag{6}$$

The input gate then updates the cell state by first determining the value to be updated by the sigmoid layer and by multiplying the candidate value vector created by the Tanh layer to get the new candidate value.

$$\begin{cases} i_t = \sigma(W_i .[h_{t-1}, x_t] + b_i) \\ C_t = \tanh(W_c .[h_{t-1}, x_t] + b_c) \end{cases} \tag{7}$$

The old cell state is then multiplied with the discarded information defined by the Forgetting Gate, plus the new candidate values to get the updated cell.

$$C_t = f_t \times C_{t-1} + i_t \times C_t \tag{8}$$

Finally, based on the current cell state, the sigmoid layer determines the output part and multiplies it with the cell state after Tanh treatment to get the output value.

$$\begin{cases} o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t = o_t \times \tanh(C_t) \end{cases} \tag{9}$$

In the LSTM model, the model can choose what to keep and what to forget so that it can analyze the data that is most relevant to the task. The LSTM model can also learn a more abstract representation of the data so that the model can learn more features of the data. These characteristics make it possible to analyze the stock trend more effectively when the LSTM model is applied in the stock market.

4. Case analysis

4.1. Sample and index data selection

The PCA-LSTM model needs to be trained several times to finally build a better model. Therefore, the training data should not only be sufficient, but also fit the training target better in order to make the model's prediction more accurate. In order to have representative data, the data related to each opening day of ST Yaxing (600319.SH) from May 2020 to March 2021 are selected as modeling data for the model, including opening and closing prices, daily high and low prices, previous closing prices, up and down amounts, up and down ranges, volume, and transaction amounts.

In the PCA-LSTM model improved for stock characteristics, the principal components of the original data were extracted based on the LSTM model combined with principal component analysis, and the correlation coefficient matrix was calculated after standardizing the original data, and then the eigenvalues and variance contribution rates were obtained by solving the characteristic equations as shown in Table 2. The variables were ranked according to the contribution of variance from largest to smallest. According to the rules of principal component selection, the first three components have the eigenvalues greater than 1 and the cumulative variance contribution rate reaches 98%, so the first three components are selected as principal components. After selecting the principal components, the data after dimensionality reduction is used as the input of LSTM.

Table 2: Eigenvalues and contribution rates

Components	**Initial Eigenvalues**			**Extraction of the sum of squares of loads**		
	Total	**Percent variance**	**Cumulative/%**	**Total**	**Percent variance**	**Cumulative/%**
1	5.164	62.560	62.560	5.164	62.560	62.560
2	1.722	20.864	83.424	1.722	20.864	83.424
3	1.213	14.702	98.125	1.213	14.702	98.125
4	0.137	1.661	99.786			
5	0.015	0.187	99.973			
6	0.001	0.016	99.989			
7	0.001	0.010	99.990			
8	0.000	0.006	99.996			
9	0.000	0.000	100.000			

4.2. Model Training and Analysis

In this paper, the relevant data for each business day between the start and end dates of the training set as input values and the closing price of the next day as output values. And 80% of the historical stock price data is used as the training set, and the remaining 20% of the historical stock price data is used as the prediction set. Data relating to the stock from May 2020 to July 2021 was used as the test set and the rest of the data was used as the training set.

The input values of the original model are the basic data such as opening and closing prices, and the neural network parameters are set as follows: input layer dimension is 12, output layer dimension is 1, initialized learning rate is 0.00006, training time step is 50. The predicted stock price curve using the model is plotted against the trend of the real stock price movement as shown in the figure below.

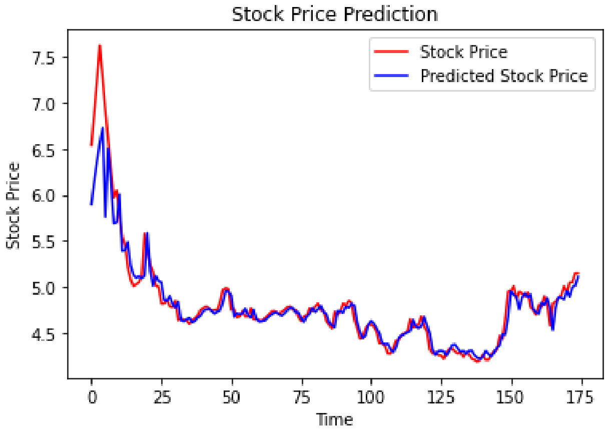

Figure 2: Model prediction results curve

The predicted curve and the true value curve basically fit in the graph, which can reflect the overall stock trend. It can be seen that the predicted value of PCA-LSTM model based on principal components has less deviation from the real value, the overall trend is the same, the curve fits better, and the results are more accurate. The data are predicted and the predictions are tested on request, here using the MAPE error tes, calculated as shown below.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{10}$$

The MAPE error test is used here, and the error MAPE obtained from the calculation of the data is less than 0.1. Therefore, it is a good prediction.

5. Conclusion

This research uses the improved LSTM model to forecast the Stock Price, introduces the Stock Correlation Technical Index, and adjusts the model according to the stock characteristic, the principal components extracted by principal component analysis and the calculated data are used as a new training sample set. The sample quality is improved, the hidden information in the data is mined deeply, the specificity of the model to the stock is increased, the dimension of the model input data is reduced, the correlation of the input features is eliminated, and the number of input layers of the LSTM neural network is reduced, at the same time of improving the input data simplification degree, it also simplifies the whole network structure.

By comparing the simulation results, it is concluded that the model based on PCA is better than the original model. It not only improves the precision of prediction, but also shortens the training time, improves the learning rate of neural network, and makes the prediction effect more stable. Because the trend of the stock market is also influenced by the external factors and its own instability, although the result and the true value will have the deviation, but the forecast general trend is consistent.

Biographical notes

Ziheng Zhang is an undergraduate student of Xi’an University of Architecture and Technology, China. His research interest focuses on Computer science and technology. Email: 2724301800@qq.com.

References

[1] Yang Q, Cao Xianbing. Stock price analysis and forecasting based on ARMA-GARCH model [J]. Practice and understanding of mathematics, 2016, 46(06): 80-86.
[2] Huang LJ, Jin TX. A study of Islamic stock market stability based on EGARCH-M model [J]. Gansu Journal of Theory, 2019, (06): 107-115+2.
[3] Fang J. An empirical study of VaR measures in Chinese stock market: A semi-parametric approach based on IGARCH [J]. Financial Theory and Teaching, 2018, (03): 15-18.
[4] Jing Nan, Wu Zhao, Wang Hefei. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction [J]. Expert Systems with Applications, 2021, 178.
[5] Cao Xiao, Sun Hongbing. Stock price prediction based on gray GARCH model and BP neural network [J]. Software, 2017, 38(11): 126-131.
[6] Deng Xuan-Kun, Wan-Liang, Huang Na-Na. Research on stock prediction based on DAE-BP neural network [J]. Computer Engineering and Applications, 2019, 55(03): 126-132.
[7] Budiharto Widodo. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM) [J]. Journal of Big Data, 2021, 8(1).
[8] Ko ChingRu, Chang HsienTsung. LSTM-based sentiment analysis for stock price forecast. [J]. PeerJ. Computer science, 2021, 7.
[9] Applied Intelligence; Researchers from University of Science and Technology Beijing Detail New Studies and Findings in the Area of Applied Intelligence (A Hybrid Stock Price Index Forecasting Model Based on Variational Mode Decomposition and Lstm Network) [J]. Journal of Robotics&amp; Machine Learning, 2020.