Stock Price Prediction Based on Optimized LSTM Model

Wandong Zhai

E-mail: 20711032@bjtu.edu.cn

Beijing Jiaotong University (BJTU), School of Economics and Management, Beijing 100044, China

Abstract: The variation trend of stock prices is often disturbed by various factors such as investor sentiment, market sentiment, and corporate performance. The combined effect of these factors leads to stochastic fluctuations in stock price change, which results in its high-noise and unstable characteristics. In responding to the problem that results of stock price prediction of original long short-term memory neural network model (LSTM) are not accurate enough, this study combines the various analytical indicators about stock prices when applying the structure of LSTM neural network. The indicators as KDJ, BOLL, MACD, ARBR, and CR, are used as the training set data together with the basic stock price trading data to expand the training set data volume and change the model parameters for prediction. After initially improving the prediction accuracy, the study uses Pearson correlation coefficient and Principal Component Analysis (PCA) to further improve the model, and an optimized LSTM neural network model (P-I-LSTM) was proposed to improve the accuracy while reducing the amount of data in the training set and improving the training speed.

Keyworks: LSTM; Principal Component Analysis (PCA); Pearson correlation coefficient; Stock price prediction; Analysis indicators

1. Introduction

Stocks, as valuable bonds, have become an important consideration for financial investment with high-risk and high-return characteristics. Compared with traditional statistical prediction models, Deep Nerval Network (DNN) is more suitable for dealing with multifactor influenced, unstable, and complex nonlinear prediction problems like stock price by analyzing deep and complex nonlinear relationships through hierarchical feature representation. Ho-chreiter and Schmidhuber proposed the Long Short-Term Memory (LSTM) model by improving the cell structure of RNN networks [1]. However, since stock data are highly noisy, dynamic, and unstable [2], accurately predicting stock prices remains a challenging task. This study introduces a variety of technical indicators related to the analysis of stock price dynamics based on the high-noise and unstable characteristics of stock price data, using the Pearson correlation coefficient and Principal Component Analysis (PCA) to downscale and screen multiple indicators that affect stock prices [3], and proposed a novel stock price prediction model P-I-LSTM with obviously higher prediction accuracy.

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338038

2. Study Method

2.1 Each analysis indicators acquisition

2.1.1 KDJ

As an overbought and oversold indicator, the indicator KDJ reflects the sensitive fluctuations of the stock price, whose essence is reflecting stock prices' trend strength, thus reflecting the buy and sell signals before the stock price will rise or fall significantly. To obtain the KDJ indicator, we should firstly calculate the value of the immature stochastic indicator RSV by calculating the highest prices, lowest prices, and closing prices of a given period (usually 9 days). After that, we can calculate the values of K, D, and J with the following formulas:

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

C_n, L_n, and H_n respectively represent the closing price, lowest price, and highest price on day n; K_p and D_p are the value of K and D on the previous day. If K_p and D_p are not available, we can replace them with 50.

2.1.2 Bollinger Band

Bollinger Band (BOLL), invented by John Bollinger, is one of the frequently-used technical indicators in the financial markets as a price trend indicator. BOLL determines stock prices' fluctuation range and futural trend by applying statistical principles to obtain the standard deviation and trust interval of prices. It contains the middle band (MB), the upper band (UP), the lower band (DN). MA is set as the stock closing prices' moving average within a period of n days, MD is corresponding standard deviation. The calculation formulas are as follows:

$$\begin{cases} MA = \frac{\sum_i^n x_i}{n} \\ MD = \sqrt{\frac{\sum_i^n (x_i - MA)^2}{n}} \\ MB = MA \\ DN = M - k \times MD \\ UP = M + k \times MD \end{cases} \tag{2}$$

2.1.3 Moving average convergence and divergence

As a trending indicator, moving average convergence and divergence (MACD) is derived by smoothing the closing price and calculation a weighted average through the construction principle of averages. It contains two kinds of moving averages EMA, which are fast and slow average with period generally set to 12 days and 26 days respectively. To obtain the value of

MACD, we can calculate its divergence DIF and moving average DEA with the period of n days.C represents the closing price. PEMA and PDEA represent the EMA and DEA values of the previous day respectively. The calculation formulas are as follows:

$$\begin{cases} EMA_n = \dfrac{(n-1) \times PEMA_n + 2 \times C}{n+1} \\ DIF = EMA_{12} - EMA_{26} \\ DEA = \dfrac{n-1}{n+1} \times PDEA_n + DIF \\ MACD = 2 \times (DIF - DEA) \end{cases} \tag{3}$$

2.1.4 ARBR & CR

ARBR sentiment indicator, which consists of popularity indicator (AR) and willingness indicator (BR). It reflects the strength and weakness of the long and short sides in the current market situation by analyzing historical stock price trend, inferring market trading sentiment, and thus making predictions about the trends. CR intermediate willingness indicator is a medium and long-term technical analysis tool to analyze the contrast between long and short forces in the stock market and to grasp the timing of buying and selling stocks. The calculation formulas are as follows:

$$\begin{cases} AR = \dfrac{\sum_i^n (H_i - O_i)}{\sum_i^n (O_i - L_i)} \times 100 \\ BR = \dfrac{\sum_i^n (H_i - CY_i)}{\sum_i^n (CY_i - L_i)} \\ M_i = \dfrac{H_i + L_i + O_i + C_i}{4} \\ CR = \dfrac{\sum_i^n (H_i - M_i)}{\sum_i^n (M_i - L_i)} \times 100 \end{cases} \tag{4}$$

H_i, O_i, L_i, and C_i respectively represent the stock's highest price, opening price, lowest price, and closing price of the day. CY_i represents C_i of the previous day; M_i is an intermediate transformation; n is the time period parameter, which is generally set as 26 days.

2.2 Pearson correlation coefficient and Principal component analysis

As a statistical analysis indicator, Pearson correlation coefficient can reflect the degree of correlation and similarity between two variables [4], which can be used in the machine learning process to calculate the correlation between characteristic variables with the calculation formula:

$$p = \frac{\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{\sqrt{\sum_i^n x_i^2 - (\sum_i^n x_i)^2} \times \sqrt{\sum_i^n y_i^2 - (\sum_i^n y_i)^2}} \tag{5}$$

In order to obtain the correlation between each characteristic variable and the target variable, which is closing price of stock, we can set y_i as the closing price and x_i as each of the other characteristic variables of the stock. Then, the Pearson correlation coefficients are calculated separately.

As a dimensionality reduction method, Principal component analysis (PCA) can filter a few comprehensive characteristic variables out from the original multiple variables as analysis indicators based on correlation analysis. Its function is removing noise and redundant characteristic variables and making sure the variables filtered out can reflect the information of original data to the maximum extent, so as to improve the data processing's speed. Its calculation steps and formulas are as follows:

(1) Firstly, we can construct a sample matrix of size $n \times p$ using the original indicators data with n samples and p indicators. In the matrix, $x_i = (x_{i1},\ x_{i2},\ x_{i3}, \dots ,\ x_{ip})^{\mathrm{T}}$, $i = 1,2, \dots, n\ (n > p)$. We can normalize this matrix by calculating the mean value by column $\overline{x_j} = \frac{1}{n}\sum_i^n x_{ij}$, and the standard deviation $S_j = \sqrt{\frac{\sum_i^n (x_{ij}-\overline{x_j})^2}{n-1}}$.

(2) Secondly, we can obtain normalized data $X_{ij} = \frac{x_{ij}-\overline{x_j}}{s_j}$, thus obtain the normalization matrix $X_i = (X_{i1},\ X_{i2},\ X_{i3}, \dots ,\ X_{ip})^{\mathrm{T}}$. Then, we can set the normalized matrix as M. Then, determine M's correlation coefficient matrix: $R = [r_{ij}]_{p\times p} = \frac{M^T M}{n-1}$, $r_{ij} = \frac{\sum M_{kj} \times M_{kj}}{n-1}$; $i, j = 1,2, \dots, p$.

(3) Thirdly, we can calculate the characteristic equation of correlation coefficient matrix R. We can let $|R - \lambda E| = 0$, and get P eigenvalues, then determine the principal components. Each principal component's variance is eigenvalues λ_i , which illustrates the influence of that component. The i th principal component contribution rate is $\frac{\lambda_i}{\sum_k^n \lambda_k}$, while cumulative contribution rate is $\frac{\sum_k^i \lambda_k}{\sum_k^n \lambda_k}$.

2.3 Long Short-term Memory Model

Long Short-term Memory (LSTM) Neural Network is obtained by improving the recurrent neural network (RNN) [5]. LSTM Neural networks are composed of multiple isomorphic blocks, and the structure stores information over time by continuously updating the internal states. Its structure is composed of forgetting gate, input gate and output gate [6].

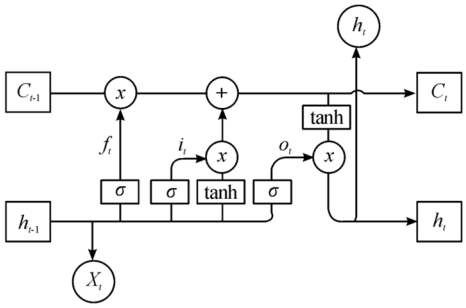

Figure 1. LSTM model's cellular structure

As depicted in Figure 1, in a single LSTM block, f、i、o denote the forgetting gate, input gate, and output gate; x denotes the input vector while h denotes the output vector; C denotes the cell state; The subscript t indicates the moment; σ、$tanh$ are the $sigmoid$、$tanh$ activation functions, respectively; W is the forgetting factor weight, b is the bias matrix;

(1) Forgetting Gate: Make data forgetting and retaining operations. The information is mapped to the interval [0,1] by the sigmoid function, if the mapped value is greater than 0.5, then it is retained; otherwise, it is forgotten. In this way, the difficult problems of gradient explosion and disappearance, and effectively deals with problems such as redundancy in historical data are solved [7]. The formulas are as follows:

$$\begin{cases} \sigma(t) = \dfrac{1}{1+e^{-t}} \\ f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right) \end{cases} \tag{6}$$

f_t denotes the forgetting factor, t is the current moment, W_f is the forgetting factor weight, h_{t-1} denotes the input of the previous loop, x_t denotes the input vector at the current moment, b_f denotes the forgetting factor bias.

(2) Input Gate: Determine which and how much information should be stored in the state cell with the following formula:

$$\begin{cases} i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \\ \quad tanh(t) = \dfrac{2}{1+e^{-2t}} - 1 \\ \tilde{C}_t = tanh\,(W_C \times [h_{t-1}, x_t] + b_C) \end{cases} \tag{7}$$

W_i denotes the forgetting weight; b_i denotes the forgetting bias; W_C denotes the memory weight; b_C denotes the memory bias. At the same time, the cell state needs to be updated at the current moment C_t; f_t denotes the input to forgotten door, C_{t-1} denotes the cell state at the previous moment:

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

(3) Output Gate: Determine which information is the output at the current moment. Its calculation formula is as follows:

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

W_o denotes the forgetting weight; b_o denotes the bias; x_t denotes the input parameter for the next neuron; h_t denotes the input state for the next moment, $\tanh\,(C_t)$ saves some of the information in the message for the output gate.

3. Empirical analysis of stock price prediction

3.1 Basic stock price trading data acquisition

Using the stock data provided by the official website of finance and the tushare finance and economics data interface package in Python, a total of 6,561 basic stock trading data of Shanghai and Shenzhen 300 index (CSI 300 index) stocks of 729 days was obtained, which contains 9 basic data indicators such as the opening price, the highest price, the lowest price, turnover, volume, rise and fall value, rise and fall percentage, turnover rate and the closing price, as shown in Table 1. Then, KDJ, BOLL, MACD, ARBR and CR analysis indicators will be further calculated as training data based on this basic data indicator.

Table 1. Basic data indicators of Stock

	OPEN	HIGH	LOW	VOLUME	AMOUNT	CHG	PCT_CHG	TURN	CLOSE
1	3894.54	3901.83	3875.97	6913053800.00	104147769298.00	-18.27	-0.47	0.25	3889.60
2	3894.17	3918.84	3833.50	9104263900.00	146981162090.00	-39.60	-1.02	0.33	3849.99
3	3850.97	3878.25	3839.42	10606628900.00	140851076272.00	28.21	0.73	0.38	3878.21
...	...	...	...	...	...	...	...	...	...
729	3769.63	3769.87	3734.00	9586696600.00	188230373665.00	-32.44	-0.85	0.33	3769.13

3.2 The selection of characteristic variables

The LSTM model requires sufficient data for multiple training to make accurate predictions. In order to make the prediction results more accurate, the training data must also be relevant to the training aim. Thus, this study proposes the preliminary improved I-LSTM model. As shown in Table 2, by further calculating basic stock trading data, we can obtain KDJ, BOLL, MACD, ARBR and CR indicators, as well as the corresponding second and third level indicators during the calculation, such as RSV, DIF, DEA, etc. Use them together as training data. Because these indicators cover the potential information of stock price fluctuations in multiple ways. I-LSTM's prediction results will be more accurate than LSTM's.

Table 2 Technical Variable Indicators Summary

First-level indicator	Second-level indicator	Third-level indicator	Variables	Explanation
Overbought & oversold Type	KDJ	KDJ.K	V1	Signals for buying and selling
		KDJ.D	V2	
		KDJ.J	V3	
Trend Type	BOLL	BOLL.UP	V4	Signals for buying and selling
		BOLL.MB	V5	
		BOLL.DN	V6	
	MACD	MACD.DIF	V7	
		MACD.DEF	V8	
		MACD.MACD	V9	
Momentum	ARBR	AR	V10	Judging transaction situation

Type				
		BR	V11	
	CR	CR.M	V12	Signals for buying and selling
		CR.CR	V13	
Basic stock data indicators	OPEN	Opening price	V14	Opening price of stock intraday
	HIGH	The highest price	V15	The highest price of stock intraday
	LOW	The lowest price	V16	The lowest price of stock intraday
	VOLUME	Trading volume	V17	The trading volume of stock intraday
	AMOUNT	Trading amount	V18	The trading amount of stock intraday
	CHG	Change	V19	The amount of change in stock price intraday
	PCT_CHG	Percentage of change	V20	The percentage for amount of change in stock price intraday
	TURN	Turnover rate	V21	The turnover rate of stock intraday

Based on the improvement of the I-LSTM model, this study proposes the P-I-LSTM model. This model combines Pearson correlation coefficient and PCA to filter characteristic variables used as training data. In this way, it can remove redundant characteristic variables, noise variables, and further improve the accuracy of prediction and training speed. By calculating the Pearson correlation coefficients between each variable characteristic and stock closing price, we can obtain the correlation coefficients of each variable. The cumulative contribution rate can be calculated by PCA, and the calculation result is taken as 97.5% contribution rate, then we can obtain 8 principal components with Pearson correlation coefficients higher than 0.9 with the stock closing price as shown in Table 3.

Table 3 The selected 8 characteristic variables

Indicator	**Variables**	**Cumulative contribution rate**	**Pearson's correlation coefficient**
OPEN	V14	0.427203	0.989096
HIGH	V15	0.622695	0.991846
LOW	V16	0.725402	0.992044
BOLL.MB	V5	0.809777	0.965524
BOLL.UP	V4	0.883606	0.949336
BOLL.DN	V6	0.930122	0.947113
MACD.DEF	V8	0.963508	0.939839
CR.M	V12	0.976925	0.998727

3.3 Determining training parameters and prediction result statistics

Using the next day's stock closing price as the output value, and the output layer dimension is 1-dimensional. For model training, the first 80% of the total characteristic variable data are used as the training set, while the final 20% as the test set. To compare LSTM, I-LSTM, and P-I-LSTM models, the Adam gradient descent algorithm is used uniformly, and the same base parameters are set, the initial learning rate is set at 0.01, the learning rate decline factor is set at 0.5, and the maximum number of iterations is set at 1200. The input values of the original LSTM model are the 8 kinds of basic stock price trading data indicators, the input layer is set as 8-dimensional. The input values of the I-LSTM improved model are the basic stock price trading data indicators and a total of 21 analytical indicators with multiple characteristics such as KDJ, BOLL and MACD, so the number of input layer increases to 21 dimensions. The input values of the P-I-LSTM model are 8 characteristic variable indicators selected by the comparison of PCA and Pearson correlation coefficients, so the input layer is 8-dimensional.

The root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and coefficient of determination R^2 were selected as evaluation indexes for quantitative evaluation of model prediction effects. The experimental results of the prediction models are shown in Table 4, and the prediction curves of the three prediction models are shown in Figures 2,3,4. n is the experimental prediction sample number; $\widehat{y_n}$ is the model prediction; y_n is the true value; $\bar{y}$ is the average of the true values:

Table 4. Error Indicator Analysis Results Summary

	RSME	R^2	MAE	MAPE
LSTM	0.024675	0.93778	0.018845	0.0004
I-LSTM	0.017286	0.96946	0.014096	0.0003
P-I-LSTM	0.011902	0.98552	0.009306	0.0002

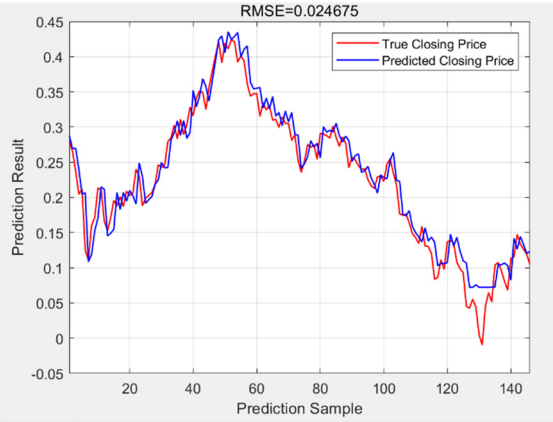

Figure 2. Prediction curve of original LSTM model

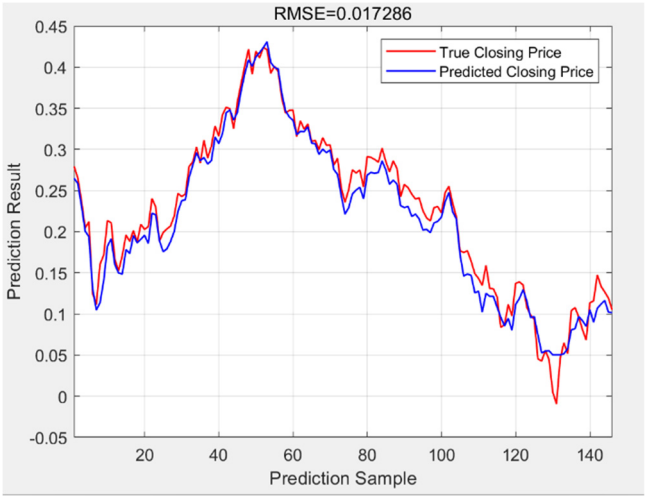

Figure 3. Prediction curve of I-LSTM model

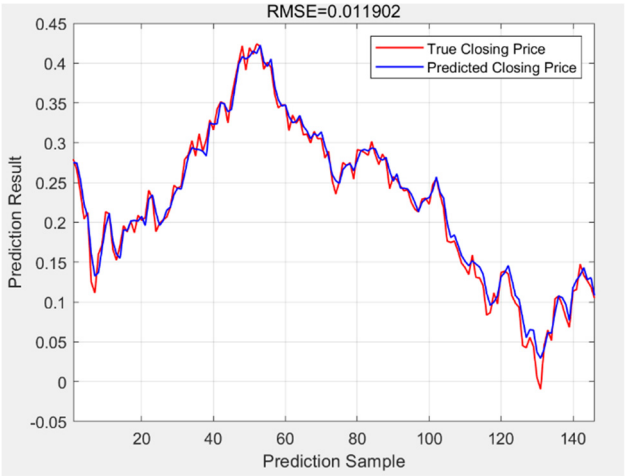

Figure 4. Prediction curve of P-I-LSTM model

4. Evaluation of prediction results and conclusions

From the analysis of the prediction results of the test set of the experiment, compared with the original LSTM model, the prediction accuracy of the I-LSTM model and P-I-LSTM model for stock characteristics is gradually improving, which confirms the superiority of the improved model. From the results of the prediction error analysis of the test set, for example, the coefficient of determination $R^2$is raised to 0.97 for the prediction results of the I-LSTM model compared with 0.93 for the original LSTM, while $R^2$for the P-I-LSTM model is even higher, which is 0.985. For RMSE, MAE, and MAPE, the values of these three error analysis indexes

are shrinking with the improvement of the model, which further demonstrates the continuous improvement of the model prediction accuracy.

This study first collected the basic stock trading data of the CSI 300 index through the official website of finance and economics as well as the tushare financial data interface package in Python. By calculating eight basic data indicators can further obtain a variety of stock price analysis indicators such as KDJ, BOLL, etc. Subsequently, an initial improved I-LSTM model was proposed, and these analysis indicators and the basic stock data were used as the training set data at the same time, so that the number of characteristic variables used for training had reached as many as 21. The essence of the I-LSTM model improvement idea is to improve the model prediction accuracy by increasing the number of analytic metrics of the characteristic variables in the training set. Subsequently, this study further proposed the P-I-LSTM model, and by applying the Pearson correlation coefficient and principal component analysis (PCA), the characteristic variables were selected while ensuring information integrity as much as possible, and the eight analyzed indexes with Pearson correlation coefficients all higher than 0.9 were selected when the cumulative contribution rate reached the 97% criterion, which not only reduced the training difficulty but also removed the problems of data redundancy and noisy data, which further improved the prediction accuracy of the model.

Reference

[1] Hochreiters, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9 (8): 1735-1780.

[2] Si Y W, Yin J. OBST-based segmentation approach to financial time series[J]. Engineering Applications of Artificial Intelligence. 2013, 26(10): 2581-2596.

[3] YU H H, CHEN R D, ZHENG G P, A SVM Stock Selection Model within PCA[J]. Procedia Computer Science, 2014, 31.

[4] MU Y S,LIU X D,WANG L D.A Pearson's correlation coefficient based decision tree and its parallel implementation[J].Information Sciences,2017,435:40-58.

[5] MA Y M,YANG J Y,FENG J W,et al.Load data recovery method based on SOM-LSTM neural network[J].Energy Reports,2022,8(S1):129-136.

[6] WANG X H,ZHOU P,PENG Z G,et al.Fault location of transmission line based on CNN-LSTM double-ended combined model[J].Energy Reports,2022,8(S5):781-791.

[7] ZENG A, NIE W J. Stock recommendation system based on deep bidirectional LSTM [J]. Computer Science, 2019, 46(10): 84-89