Stock price prediction based on PCA-LSTM model

Zheng Xinyuan *
Guangdong Mechanical and Electrical Polytechnic,
Guangzhou, 510515
China.lucaszheng@163.com

Xiong Naiping
Guangdong Mechanical and Electrical Polytechnic,
Guangzhou, 510515, China
373057277@qq.com

ABSTRACT

In order to improve the prediction accuracy, this study proposes an new PCA-LSTM neural network stock price prediction model that combines principal component analysis(PCA) and long-term and short-term memory neural network (LSTM). We download time series indicators and technical indicators of PingAn insurance (X601318) form Tushare interface and Wind database. PCA method was used to reduce the technical indicators dimension, LSTM model was used to predict the next day stock closing price. The results show that PCA-LSTM model can greatly reduce data redundancy and obtain better prediction accuracy compared with the simple LSTM model.

Additional Keywords and Phrases: stock price prediction, PCA, LSTM

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural networks**;

ACM Reference Format:
Zheng Xinyuan and Xiong Naiping. 2022. Stock price prediction based on PCA-LSTM model. In *2022 5th International Conference on Mathematics and Statistics (ICoMS 2022), June 17–19, 2022, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3545839.3545852

1 INTRODUCTION

The stock market is essentially a dynamic, non-stationary, noisy and chaotic system, it is difficult to achieve good prediction results by traditional statistical prediction methods such as regression analysis and time series analysis[1]. With the development of artificial intelligence in big data era, deep learning stands out because of its powerful learning ability and prediction accuracy in stock price prediction. The two most basic models of deep learning are CNN and RNN. The long-term and short-term memory neural network (LSTM) is one of the most classic variants of RNN. LSTM has great advantages in dealing with time series data and correlated data, it can effectively store the data characteristics and solve the problem of gradient disappearance through controlling the input and output between hidden layers. The stock market is easy to be affected by all kinds of information, so all related information should be used as model input to the predict. LSTM has no restrictions on the form of input variables, greatly reduces the number of neurons weights and effectively prevents over fitting through the cyclic use of neuron weights[2-6]. Because of too many information, this study proposes a new PCA-LSTM neural network stock price prediction model. Firstly, the PCA analysis is used to reduce the dimension and maximize the correlation between various input data in this experiment. Then, PCA-LSTM model is constructed to analyze the stock data of Ping An insurance from 2010 to 2021 and predict the next day closing price using the deep learning platform keras. Prediction accuracy was evaluated by mean absolute percentage error (MAPE) between prediction results with real value, and stability was evaluated by comparing the simple LSTM model and PCA-LSTM model.

*Corresponding author.

ICoMS 2022, June 17–19, 2022, Paris, France

ACM ISBN 978-1-4503-9623-3/22/06…$15.00
https://doi.org/10.1145/3545839.3545852

2 PCA-LSTM NETWORK MODEL FRAMEWORK

2.1 Principal component analysis(PCA)

Principal component analysis is a dimension reduction statistical method. It replaces many original variables by several linear and uncorrelated combination. These new uncorrelated variables are called principal components and reflects the original variable information as much as possible. Firstly, the original data is standardized and transformed to eliminate the influence of different dimensions, $x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i}$, x_i' represents the data after standardized transformation, x_i represents raw data. Then the correlation coefficient matrix of variables and the eigenvalues and eigenvectors are calculated. And then calculate the variance contribution rate $a_i = \frac{\lambda i}{\sum_{i=1}^{p} \lambda i}$, a_i reflects the percentage of the ith principal component in the original population variable. Finally, the contribution rate of involved variance is calculated $\sum_{i=1}^{m} a_i$. The minimum number of m was determined according to the principle of accumulating more than 80%.

2.2 Long Short-term Memory Networks

LSTM is a special type of RNN, which can learn long-term dependent information. The LSTM unit consists of a storage unit (C_t), a unit status update value ($\widetilde{C_t}$) and three gates, such as forgetting gate (f_t), input gate (i_t) and output gate (o_t), see Figure 1 for details. In which, storage unit $C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$. Hidden node $h_t = o_t \times \tanh(C_t)$. In the back propagation of LSTM model error, some errors can be directly transmitted to the next layer of neurons through the input gate, and some errors can be forgotten through the forgetting gate, so LSTM solves the problem of gradient explosion and disappearance.

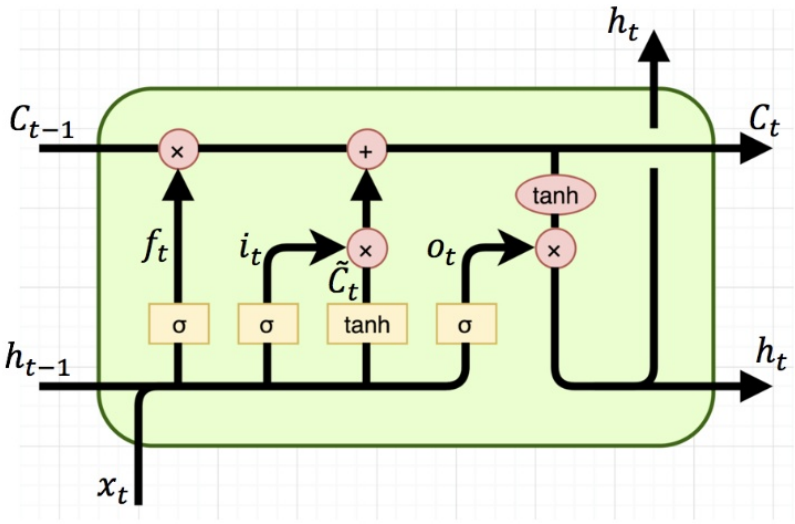

Figure 1: The long short-term memory network block architecture

2.3 PCA-LSTM model framework

Stock price forecasting is a typical time series problem, and the price of a certain time is affected by the price of the previous time and historical multi time, so LSTM model is selected for stock price forecasting. In this paper, PCA and LSTM are fused to construct PCA-LSTM deep learning prediction model. PCA reduces the complexity of input data to prevent over fitting followed by the LSTM layer. The number of layers and neurons in each layer are determined by empirical analysis. the dropout method is used after each LSTM layer to reduce the over fitting phenomenon and improve the generalization ability of the model, and finally is the full connection layer. PCA-LSTM can effectively filter and update the associated data information and better predict the associated data by the forgetting gate and input gate.

3 EMPIRICAL ANALYSIS BASED PCA-LSTM MODEL

3.1 Process of stock price prediction based on PCA-LSTM model

In this paper, PCA-LSTM neural network is used to predict stock data. The design process is as follows: Step #1: Download and preprocess stock data, remove abnormal data and normalize the data; Step #2: PCA dimensionality reduction; Step #3: build a neural network model, determine the number of layers of neural network, the number of neural network nodes in each layer, the optimizer, the learning rate and the definition of loss function; Step #4: conduct model training and optimize preset parameters; Step #5: apply model prediction, evaluate the predicted value, display the loss rate and error, save the optimal model and parameters.

3.2 Downloading and preprocessing of stock data

Indicators in time series of stock trading, such as opening price, highest price, lowest price, closing price of the previous day, closing price, trading volume and turnover, are collected from Tushare interface[7]. There are many influencing factors of stock price fluctuation, we should consider the influencing factors as far as possible and analyze the problems in an all-round way. Trading technical indicators are important parameters calculated from time series indicators. Investors widely use technical indicators to detect trading signals. This paper selects the important 66 trading technical indicators recognized by the market, which can cover the potential information of stock price fluctuation in many aspects, and has strong stock price explanation. After comprehensive analysis, this paper selects 58 sub-indicators of 11 categories, including market indicators, valuation indicators, profitability indicators, cash flow indicators, solvency indicators, operating capacity indicators and growth capacity indicators. These data downloads from the Wind database[8]. See Table 1 for 66 indicators in details.

This paper selects the stock trading time series indicators and technical indicators of PingAn insurance (X601318) from January 4, 2010 to December 30, 2021. Delete all rows with null values in the transaction time series indicators, delete rows with more than 3 null values and columns with more than 10 null values in the technical indicators, and fill the null values with KNN method. There are 2776 remaining sample data and 60 indicators. This paper uses 75% of the sample as the training set and the remaining 25% as the prediction set. that is, the first 2082 days was used as input data to predict the closing price of the next day, the last 694 days as the prediction set to evaluate prediction accuracy.

Table 1: stock trading time series indicators and technical indicators

category	Index and their identifier
Trading time series indicators	Opening price, high price, low price, close price, pre_close price, change, volume, amount
Trend index	BBI, DMA, DMI, EXPMA, MA, MACD, MTM, RICEOSC, SAR, TRIX,
Counter trend index	B3612, BISA, CCI, DPO, KDJ, SLOWKD, ROC, RSI, RPS, SI, SRDM, VROC, VRSI, WR
Energy index	ARBR, CR, PSY, VR, WAD
Volume price index	MFI, OBV, PVT, SOBV, WVAD
Pressure support index	BBIBOLL, BOLL, CDP, ENV, MIKE
Trading volume	Quantityrelativeratio, VMA, VMACD, VOSC, TAPI, VSTD
Overbought and oversold indicators	ADTM
Swing index	MI, RC, SRMI
Relative Strength Index	DPTB, JDQS, JDRS, ZDZB
Volatility index	ATR, MASS, STD, VHF, CVLT

Table 2: Eigenvalues and percentage of explained variance by first eight components by PCA

Component	RC2	RC1	RC3	RC4	RC6	RC8	RC5	RC7
Eigenvalue	11.566	11.133	7.474	3.12	2.936	2.232	1.837	1.463
Proportion (% of variance)	0.222	0.214	0.144	0.06	0.056	0.043	0.035	0.028
Cumulative (%)	0.222	0.437	0.58	0.64	0.697	0.74	0.775	0.803

3.3 PCA dimensionality reduction

Due to the large number, overlapping and strong collinearity of stock technical indicators, it is necessary to reduce the dimension of technical indicators and eliminate the collinearity. PCA is an effective way to reduce the computational burden, improve the convergence speed, and improve the prediction ability of LSTM neural network[9-10].

PCA was carried out by psych R package[11]. use natural logarithm ln (P) transformation to normalize the data to eliminate the dimensional gap between the data. PCA on 52 technical indicators yielded eight principal components explaining sample variance of about 80.3% (Table 2), indicating that these eight main factors can basically reflect most of the information of the original variables, and it is feasible to use them to replace the original index variables for research and analysis. As demonstrated in Figure 2, similar to eigenvalues, 8 components were retained according to where the slope of the scree plot levels off. These components accounted for 80.3% of the total variance.

In order to make the interpretation of principal component analysis more efficient and minimize the data redundancy of LSTM neural network model, we retain the parameter data with a value greater than 0.5. The eight principal components are as follows:

- RC2=0.96 BOLL + 0.96 CDP + 0.96 BBI +. . .+0.52 ATR

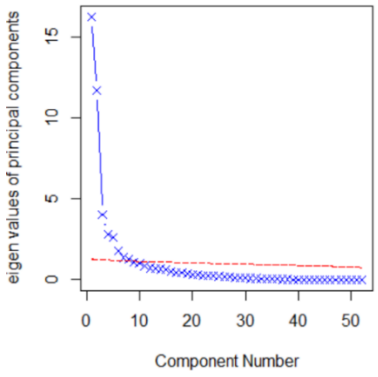

Figure 2: Scree plot of eigenvalues

- RC1=0.93 BIAS + 0.92 KDJ + 0.91 CCI +. . .+0.57 DPO
- RC3=0.93 TRIX + 0.92 DMA + 0.88 PRICEOSC +. . .+0.62 ARBR
- RC4=0.86 STD + 0.77 VMA + 0.74 ATR +. . .+0.60 TAPI
- RC6=0.82 VMACD + 0.82 VOSC + 0.75 CVLT
- RC8=0.65 WAD + 0.57 ZDZB
- RC5=0.56 VHF + 0.53 VR

Table 3: experimental configuration environment

content	configuration parameters
operating system	Windows 10
processor	AMD Ryzen 9 5900HX
memory	32.0 GB
graphics card	NVIDIA GeForce RTX 3070 Laptop GPU
Python	3.8.0
TensorFlow	2.6.0
Keras	2.6.0

Table 4: Prediction results under different LSTM model parameters

LSTM layers	Dense layers	Neurons	minimum MAPE			
			previous 10 days		previous 15 days	
			LSTM	PCA-LSTM	LSTM	PCA-LSTM
1	1	16	14.114	8.122	15.850	8.572
1	1	32	13.009	6.954	15.848	7.239
1	1	64	15.485	6.223	12.463	6.174
1	1	128	13.083	7.428	13.080	8.055
1	2	16	15.345	9.321	14.137	7.365
1	2	32	15.218	6.611	20.033	8.319
1	2	64	13.659	6.382	14.538	7.895
1	2	128	12.858	7.909	15.318	6.586
2	1	16	12.056	8.981	11.881	8.178
2	1	32	11.955	8.612	16.802	9.498
2	1	64	15.704	6.016	14.189	5.818
2	1	128	12.440	6.534	13.904	6.660
2	2	16	11.317	6.992	10.132	9.890
2	2	32	13.675	8.563	24.581	9.645
2	2	64	13.278	4.692	13.859	5.205
2	2	128	15.549	6.583	15.553	8.297
3	1	16	15.000	7.245	16.107	7.996
3	1	32	13.968	7.499	24.058	8.535
3	1	64	11.950	5.795	14.947	7.240
3	1	128	13.585	5.391	12.208	7.419
3	2	16	14.853	9.841	17.518	9.885
3	2	32	15.245	8.741	30.053	8.007
3	2	64	12.976	5.215	11.630	5.225
3	2	128	14.417	6.425	18.869	5.542

- RC7=0.79 Quantityrelativeratio +0.63 MFI

3.4 Construction of LSTM model and parameter optimization

Keras, a popular deep learning framework, is used for the experiment[12], and GPU is used for acceleration, the specific experimental configuration environment is shown in Table 3

Neural network training is a complex process. Different super parameters of LSTM model have significant impact on the prediction ability of LSTM model. For example, if the number of layers and neurons of neural network is too small, so the structure of neural network is too simple, its learning ability and classification ability will be reduced; However, too much layers will make the neural network structure too complex and heavy, the predicted efficiency will be reduced, and the promotion ability will become worse. In this paper, using the close price of the previous days (10 days, 15 days) to predict the close price of the next day. The number of LSTM layers (1 layers, 2 layers, 3 layers), the number of neural units (16, 32, 64, 128), and the number of dense layers (1 layers, 2 layers) of LSTM neural network are optimized, a dropout layer with parameter 0.1 followed each LSTM layer. the activation function is Relu. and the number of iterations is 500. The optimizer selects the adaptive motion estimation (Adam) algorithm. Adam uses the first-order moment estimation and second-order moment estimation of the gradient to dynamically adjust the learning rate

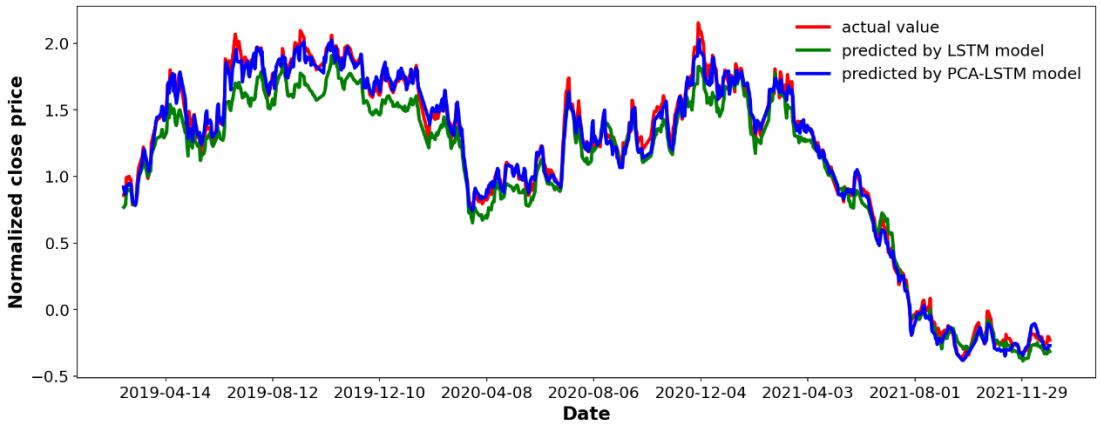

Figure 3: Comparison between predicted value and actual value. Red: actual value of the stock closing price. Green: predicted value of the normalized stock closing price of LSTM model when memory = 15, LSTM layers = 2, Dense layers = 2, Neurons=16; blue: predicted value of the normalized stock closing price of PCA-LSTM model when memory = 10, LSTM layers = 2, Dense layers = 2, Neurons=64.

of each parameter. The main advantage is that after bias correction, the learning rate of each iteration has a certain range, making the parameters more stable. The Mean Absolute Percentage Error (MAPE) is used to evaluate the prediction accuracy. The smaller the value, the more accurate the prediction result is. Compare the prediction efficiency between LSTM model and PCA-LSTM model by MAPE, table 4 shows the experimental results. As in Table 4, the smallest MAPE of LSTM model is 10.132 when model sets previous day=15, LSTM layers=2, neurons per LSTM layers=16, and dense layers=2. The smallest MAPE of PCA-LSTM model is 4.692 when model sets previous day=10, LSTM layers=2, neurons per LSTM layers=64, and dense layers=2. The results show that the MAPE value is the smallest in PCA-LSTM model, PCA-LSTM model can greatly reduce the data redundancy, and provide best prediction accuracy.

Use the training set to train LSTM models and PCA-LSTM models, and then test the performance of the models on the test set. The prediction results of each model on the test set are shown in Figure 3. In Figure 3, the red line represents the actual value of the stock closing price, the green line represents the predicted value of the stock closing price of LSTM model, and the blue line represents the predicted value of PCA-LSTM model, the abscissa is the time, and the ordinate is the normalized close price. As can be seen from the figure, blue line is more coincident with red line, the predicted result of PCA-LSTM model is more accurate.

4 CONCLUSION

Taking the stock trading indicators of PingAn insurance (X601318) from January 4, 2010 to April 30, 2021 as the input characteristics, this paper uses indicators of previous days to predict the closing price of PingAn insurance stock of the next day with LSTM model and PCA-LSTM model. PCA-LSTM model carries out principal component analysis on technical indicators of stocks first, extracts the principal component with the least number, reduces data redundancy, makes the prediction model more efficient and accurate.

REFERENCES

[1] Funt, M. J. Stock market indicators. Pennsylvania dental journal[J]. 2010, 77(3), 37-38.
[2] Han, Taedong. Stock Price Prediction Using LSTM: Focusing on the Combination of Technical indicators, Macroeconomic Indicators, and Market Sentiment[J]. The Society of Convergence Knowledge Transactions.2021, 9 (4) :189-198.
[3] Shouvik Banik, Nonita Sharma, Monika Mangla, *et al.* LSTM based decision support system for swing trading in stock market[J]. Knowledge-Based Systems, 2022, 239:107994.
[4] Yinghao Ren, Fangqing Liao, Yongjing Gong. Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM Model[J]. Procedia Computer Science, 2020, 174:128-140.
[5] Moghar A, Hamiche M. Stock Market Prediction Using LSTM Recurrent Neural Network[J]. Procedia Computer Science, 2020, 170:1168-1173.
[6] Yadav A, Jha C K, Sharan A. Optimizing LSTM for time series prediction in Indian stock market[J]. Procedia Computer Science, 2020, 167:2091-2100.
[7] Tushare. https://tushare.pro/.
[8] Wind Economic Database. https://www.wind.com.cn/en/edb.html.
[9] Nema Salem, Sahar Hussein. Data dimensional reduction and principal components analysis[J]. Procedia Computer Science, 2019,163:292-299.
[10] Hess, A. S., & Hess, J. R. Principal component analysis. Transfusion, 2018, 58(7), 1580-1582.
[11] Revelle W (2022). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.2.3, https://CRAN.R-project.org/package=psych.
[12] Chenjie Sang, Massimo Di Pierro. Improving trading technical analysis with TensorFlow Long Short-Term Memory (LSTM) Neural Network[J]. The Journal of Finance and Data Science, 2019, 5(1):1-11.