Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060504
ISSN 2371-8838 Vol. 6 Num. 5

Research on Stock Price Prediction Based on Orthogonal Gaussian Basis Function Expansion and Pearson Correlation Coefficient Weighted LSTM Neural Network

Shun Lin[1,*], Yuan Feng[2]

[1]School of Fiance, Hunan University of Technology and Bussiness, Changsha, 410205, China
[2]School of Economics, Central South University of Forestry and Technology, Changsha, 410004, China
**Corresponding author: 596046083@qq.com*

Keywords: Stock price prediction, model averaging, orthogonal Gaussian basis function, Bagging

Abstract: For stock price prediction in quantitative finance, deep learning techniques such as LSTM neural network do not need the stationarity assumption of traditional time series models (such as ARIMA and GARCH) and can forecast medium and long-term time series, so they have attracted much attention. This paper proposes an improved LSTM neural network based on orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighting. The proposed method uses the functional features of intra-day prices to fit the residual series predicted by the LSTM neural network. Considering that the underlying model structure between each component of the function eigenvector and the residual series is unknown, we use the Bagging method to capture and trade off the variance and bias of the prediction model. In addition, since the dimension of the predictive variable of the LSTM neural network is a parameter to be estimated, we use the model averaging method based on Pearson correlation coefficient weighting for tuning. The results of actual data **a**nalysis show that the proposed method can significantly improve the prediction accuracy of the original LSTM neural network and has certain robustness. Finally, the proposed method can be further applied to consumer price index (CPI) prediction, daily average temperature prediction, and real-time monitoring of environmental trace elements.

1. Introduction

As China's modern market economy is becoming more and more developed, people's financial management consciousness is increasingly mature, and finance, as the core of the modern economy, gradually becomes people to focus on hot areas. The stock market, as a barometer of the national economy, finance and investment, naturally becomes the most direct; the wave profile reflects the trend of the economy, affects the entire market of nerve fibers, and is one of the key research directions of many experts and scholars. Stock price volatility in the stock market and the influencing factors and more complex; at the same time, the stock price is a big noise, high

dimension, information is not easy to capture the characteristics of time series. Therefore, a problem for academia and industry to focus on how to accurately reveal the changing trend of the time series of share prices to forecast the stock reasonably.

For stock price series prediction, there are many classical time series prediction models. The classical differential autoregressive moving average (ARIMA) model and generalized autoregressive conditional heteroscedastic (GARCH) model have been widely used. For example, the combination of AMIMA and MLPs (multi-layer perceptron) is used to forecast S&P 500 index, Shenzhen Component index, and Dow Jones Index (Rahimi, Z.H.Etter.,2018). [1]; Accuracy of improved GARCH family model in stock market prediction (Wanrui,2022) [2]. However, when the above classical model is established for time series prediction, the time series data must be stable or stable by differencing; otherwise, it will not be able to capture the law. In addition, the model can only capture linear relations but not nonlinear relations and functional relations [3-4]. With the development of machine learning methods, the ARIMA model combined with kernel principal component analysis (KAPC) can achieve nonlinear dimensionality reduction of data, which can mine the nonlinear information contained in the data set and improve prediction accuracy. However, the data structure between the dimensionality reduction variables and the response is unknown (Zheng Hong et al., 2020). [5]. Compared with traditional prediction methods, LSTM (Long short-term Memory) neural network (Bao Yueyan, 2021) [6] BP Neural network (Zeng Lifang et al., 2020) [7] Deep learning methods, such as more good at dealing with time-series data, through them to predict stock price time series, such as no stationarity assumption, need not consider the problem of dimension disaster, but also by capture the nonlinear activation function information, but also they are not able to capture the function information, the predictor variable dimension is not easy to choose (GUI-jun Yang, etc., 2022) [8]. Therefore, both the traditional classical model and the neural network model have their own advantages and limitations. The traditional classical model can well explore the implicit linear relationship in the data, while the neural network model has its own merits in dealing with nonlinear relationships and great dimensional problems. However, stock price prediction needs to combine the advantages of the two types of models to build a combined model. The common idea of composite model construction is to decompose the data, fit the linear and nonlinear parts by statistical model and neural network model, respectively, and then superimpose to get the prediction results. For example, Zhang (2003) [9]. A combined model of ARIMA and neural networks is discussed. In the first stage, ARIMA captures the linear trend in the time series data, and then, based on the output of the previous stage, ANN is used to capture the nonlinear relationship in the residual series. Based on this idea, a large number of scholars have constructed portfolio models to predict financial time series and confirmed the advantages of portfolio models over single models (Anna et al.,2021) [10].

In summary, this paper proposes a combined prediction model based on orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighted LSTM neural network. The proposed method has the following advantages: first, it does not need stationarity assumption and can extract not only linear and nonlinear information but also add functional auxiliary information to the original time series prediction by considering the intraday price and basis expansion method. Second, the latent model structure of the feature components and residual series is determined by the Bagging method. Thirdly, it provides a parameter selection method based on model averaging for the LSTM neural network to select the appropriate dimension of predictor variables. Experimental results show that the proposed method has higher accuracy and robustness than the original LSTM.

2. Theory and Methods

2.1. Gaussian basis function expansion

Assume there is an independent variable, and consider the functionalization of its data without considering the dependence for the sign in the variable, i.e. $\{(x_{\alpha i}, t_{\alpha i}); i = 1, \cdots, N_\alpha, t_{\alpha i} \in \tau \subset R\}$ $\{(x_i, t_i); i = 1, \cdots, N\}$. Assume that the observations in the main body are derived from the regression model:

$$x_i(t) = \sum_{k \geq 1} a_{ik}\ u_k(t) + \varepsilon_i, \qquad \text{i=1,}\cdots\text{,N} \tag{1}$$

Where the residuals obey an independent normal distribution, a_{ik} denotes the coefficients, $u_k(t)$ is a set of orthogonal basis functions, each of which forms a local acceptance domain in the input control, and the specific expression for the Gaussian basis function is

$$u_k(t;\ \mu_k, \eta_k^2, v) = \exp\left\{-\frac{(t-\mu_k)^2}{2v\eta_k^2}\right\} \tag{2}$$

Where μ_k is the location of the decision center, η_k^2 is the discrete parameter, and v is the hyperparameter. A clustering algorithm is first used to determine the centers and discrete parameters of the Gaussian basis functions, and a method of constructing Gaussian unitary orthogonal basis is used in this position, and the procedure is as follows.

$$v_i = \frac{u_i}{\sqrt{\tau_i}}; \tau_i = (\pi \upsilon \widehat{\eta_i}^2)^{\frac{1}{2}}; i, j = 1, 2, 3, \cdots, K \tag{3}$$

$$\langle V_i, V_j \rangle = \left(\frac{2\pi \widehat{\eta_i}^2 \widehat{\eta_j}^2}{\tau_i \tau_j \widehat{\eta_i}^2 + \tau_i \tau_j \widehat{\eta_j}^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{\left(\widehat{\mu_i} - \widehat{\mu_j}\right)^2}{2v(\widehat{\eta_i}^2 + \widehat{\eta_j}^2)}\right\} \tag{4}$$

$$\langle V_i, \Phi_j \rangle_{Hx} = \langle V_i, V_j \rangle_{Hx} - \sum_{k=1}^{i-1} \langle V_i, V_k \rangle_{Hx} \langle V_j, V_k \rangle_{Hx} \tag{5}$$

$$\begin{cases} Q_1 = 1 \\ Qi = 1 - \sum_{j=1}^{i-1} \langle V_i, \Phi_j \rangle_{Hx}^{\ 2}; i > 1 \end{cases} \tag{6}$$

$$\varphi_1 = \frac{V_1}{Q_1}; \varphi_i = \frac{\left(V_i - \sum_{j-1}^{i-1} \langle V_i, \Phi_j \rangle_{Hx}\right)}{Q_i}; i > 1 \tag{7}$$

After orthogonal basis functions, the regularization method is penalized by maximizing the log-likelihood function and the maximum penalized likelihood estimator is:

$$\hat{a} = (\varphi^T \varphi + n\lambda_a \widehat{\sigma}^2 R_a)^{-1} \varphi^T x_i; \widehat{\sigma}^2 = \frac{1}{n}(x_i - \varphi^T \hat{a})^T (x_i - \varphi^T \hat{a}) \tag{8}$$

In practice, the GIC is used for each curve to obtain the optimal number of basis functions.

Finally the coefficient matrix is obtained: $\Lambda = (\hat{a}_1, \hat{a}_2, \hat{a}_3, \cdots \hat{a}_n)^T$.

2.2. Improved LSTM neural network based on Pearson correlation coefficient weighting

In this paper, the intraday price is considered as auxiliary information and then as a function, and the orthogonal Gaussian basis function expansion is used to extract the feature information of the function. Furthermore, since the underlying model structure between each component of the feature vector and the residual sequence is unknown, we use the Bagging method to capture the feature information. The bagging ensemble algorithm is a technique to improve the generalization error by combining multiple base models. In addition, LSTM is a time-recurrent neural network obtained by optimizing RNN. It can effectively overcome the problem of vanishing gradients in RNN and outperforms RNN, especially in long-distance dependent tasks. However, since the selection of the prediction period of the LSTM neural network is a problem to be solved, the model averaging method based on the Pearson correlation coefficient is used in this paper to measure the prediction accuracy and evaluate the importance of each model. In this paper, this combined model prediction method is called LSTM neural network based on Orthogonal Gaussian Basis Function expansion and Pearson correlation coefficient weighting (GBM-LSTM). The specific algorithm process is as follows:

Step.1: Obtain the stock opening price and its corresponding intraday price data;

Step.2: Set multiple forecast periods, and use LSTM neural network for each forecast period to obtain the LSTM neural network predicted value sequence set.

Step.3: By using the difference between the true value of the opening price and the predicted value of multiple LSTM neural networks, multiple residual sequences are obtained

Step.4: Gaussian basis expansion and Bagging regression were used to predict the residual series, and the residual predicted values of each LSTM neural network were obtained, and the residual predicted values were added to the original predicted values.

Step.5: Pearson correlation coefficient is used to calculate the correlation coefficient between the predicted value sequence and the real value sequence on the training set and then normalized to obtain the weight vector of the LSTM neural network set.

Step.6: Inner product of the new predicted value vector obtained from multiple LSTM neural networks with the weight vector to obtain the final predicted value.

3. The data analysis

3.1. Data sources and pre-analysis

The data of this paper comes from Wind database (https://www.wind.com.cn/), and the daily stock prices of three Chinese A-share markets, namely, Trendy Energy (SH600777), Oriental Group (SH600811) and Opai Household (SH603833), are selected as samples. They belong to oil and gas extraction industry, agricultural and sideline food processing industry and custom furniture industry respectively. The reason for selecting these three stocks is that they have different stationarity and complexity degree, so as to test whether the stock price prediction model based on orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighted LSTM neural network has the same excellent improvement effect for stocks with different stationarity and complexity degree. See Figure 1

Specifically, the time series trend of the three stocks is shown in FIG. 1. Firstly, by observing the time series diagrams of the three stocks, we can preliminarily judge that the original numbers are all unstable. The test results show that the absolute values of the ADF test statistic of the three stocks

are -2.723, -1.252 and -2.395 respectively, which are less than the critical value at the significance level of 1%. Therefore, the null hypothesis is not rejected and there is a unit root. Therefore, at the significance level of 0.05, the data of the three stocks can be considered as unstable time series data.

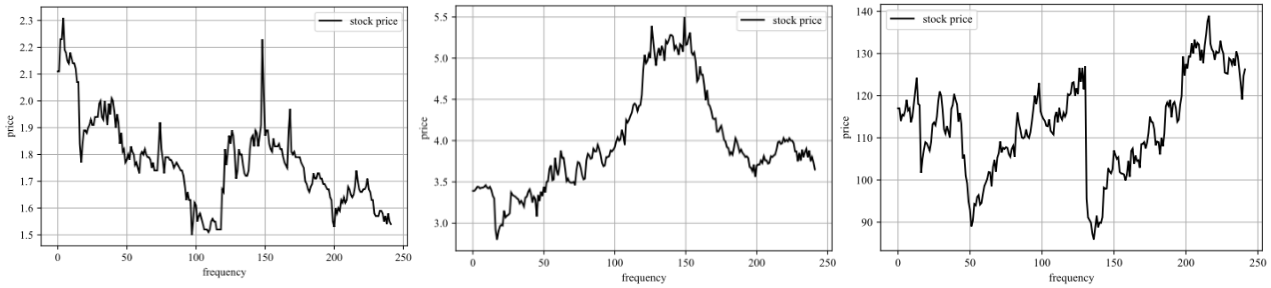

Figure 1: The opening trend of stocks SH600777, SH600811 and SH603833

3.2. Comparison Results

The experiment is implemented based on TensorFlow framework. Select three stocks sample number is 242, frequency intraday price was $240, to test robustness, we select different training periods:

t = [150,155,160,165,170,175,180,185,190,195,200,205,210,215,220],

In addition, The set LT=[12,13,14,15,16] selected for LSTM prediction periods, and the specific results are shown in table 1-4 below. How to objectively evaluate the accuracy of a model needs to introduce three indicators: mean relative error (MRE), mean square error (MSE), posterior error (BE), and their corresponding calculation formulas are as follows:

$$MRE = \frac{1}{n}\sum_{k=1}^{n}\frac{|e_k|}{Y_k} \quad MSE = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \hat{Y}_k\right)^2 \quad BE = \frac{S_2}{S_1} \tag{9}$$

Where e_k is the residual sequence, Y_k is the true value, S_1 is the standard deviation of the original sequence, S_2 is the standard deviation of the relative value sequence. The smaller the three indicators of the model, the higher the prediction accuracy.

Table 1: Comparison of stock SH600777 evaluation indicators

	GBM-LSTM-MSE	GBM-LSTM-MRE	GBM-LSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
150	0.00200	0.01746	0.34356	2.02162	0.81753	0.47208
155	0.00173	0.01640	0.31812	2.04988	0.82314	0.45700
160	0.00180	0.01649	0.32519	2.05869	0.82468	0.45681
165	0.00189	0.01662	0.33187	2.04203	0.82128	0.46085
170	0.00181	0.01676	0.32482	2.03832	0.82070	0.45397
175	0.00169	0.01587	0.31493	2.04710	0.82258	0.44354
180	0.00178	0.01618	0.32396	2.04128	0.82088	0.47430
185	0.00152	0.01502	0.29989	2.05340	0.82379	0.44495
190	0.00159	0.01502	0.30612	2.06157	0.82554	0.44350
195	0.00149	0.01468	0.29791	2.05440	0.82397	0.45229
200	0.00147	0.01462	0.29606	2.06088	0.82499	0.46825
205	0.00157	0.01443	0.30544	2.06959	0.82709	0.46019
210	0.00147	0.01422	0.29592	2.06490	0.82581	0.47961
215	0.00145	0.01389	0.29373	2.06529	0.82593	0.47241
220	0.00139	0.01346	0.28786	2.07713	0.82879	0.45067

Table 2: Comparison of stock SH600811 evaluation indicators

	GBM-LSTM-MSE	GBM-LSTM-MRE	GBM-LSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
150	0.00458	0.01163	0.10244	12.60341	0.88458	0.64550
155	0.00389	0.01121	0.09658	12.64973	0.88590	0.65439
160	0.00323	0.01070	0.08950	12.65407	0.88564	0.66380
165	0.00291	0.01010	0.08643	12.72791	0.88788	0.67362
170	0.00227	0.00908	0.07714	12.75137	0.88877	0.67123
175	0.00210	0.00860	0.07430	12.81115	0.89081	0.67329
180	0.00207	0.00859	0.07384	12.82162	0.89103	0.67727
185	0.00184	0.00806	0.06956	12.85133	0.89224	0.67301
190	0.00197	0.00836	0.07198	12.84335	0.89202	0.67140
195	0.00207	0.00833	0.07384	12.85054	0.89213	0.67524
200	0.00181	0.00787	0.06899	12.85602	0.89262	0.66737
205	0.00189	0.00792	0.07042	12.86463	0.89267	0.67459
210	0.00201	0.00814	0.07260	12.86734	0.89265	0.67774
215	0.00181	0.00769	0.06905	12.86415	0.89268	0.67380
220	0.00140	0.00678	0.06067	12.91695	0.89469	0.66873

Table 3: Comparison of stock SH603833 evaluation indicators

	GBM-LSTM-MSE	GBM-LSTM-MRE	GBM-LSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
150	107.96757	0.06258	0.83143	12707.38803	0.99581	0.98593
155	104.79488	0.06113	0.81471	12708.53685	0.99585	0.98642
160	105.98107	0.06151	0.82093	12709.52560	0.99589	0.98621
165	103.76845	0.06080	0.80610	12707.35481	0.99582	0.98581
170	104.21441	0.06019	0.80430	12707.82570	0.99583	0.98591
175	106.74349	0.06033	0.80968	12708.72348	0.99585	0.98656
180	99.17044	0.05890	0.78851	12707.75114	0.99582	0.98631
185	94.64264	0.05776	0.77912	12707.35755	0.99581	0.98608
190	83.93170	0.05502	0.74349	12707.02293	0.99580	0.98595
195	74.05432	0.05381	0.70782	12707.83289	0.99583	0.98624
200	65.43326	0.05219	0.67375	12706.88763	0.99579	0.98602
205	53.33528	0.04965	0.61708	12705.71130	0.99576	0.98520
210	51.12224	0.04924	0.60837	12704.21630	0.99571	0.98490
215	48.77466	0.04846	0.59684	12704.45025	0.99573	0.98455
220	45.88752	0.04740	0.58033	12716.96368	0.99620	0.98603

Table 4: Difference test

indicators	SH600777-MSE	SH600777-MRE	SH600777-BE
T value	560.46965	1005.62802	28.52318
P values	$7.15783 * 10^{-32}$	$6.18918 * 10^{-46}$	$8.59037 * 10^{-21}$
indicators	SH600811-MSE	SH600811-MRE	SH600811-BE
T value	520.00071	1004.58882	158.34053
P values	$2.06622 * 10^{-31}$	$1.77586 * 10^{-47}$	$9.57109 * 10^{-41}$
indicators	SH603833-MSE	SH603833-MRE	SH603833-BE
T value	1993.70572	674.25353	10.58353
P values	$1.37759 * 10^{-40}$	$5.20293 * 10^{-33}$	$4.60349 * 10^{-08}$

Table 1, Table 2 and Table 3 show the comparison of three stock evaluation indexes, and Table 4 shows the difference test of evaluation indexes. The first three columns of Table 1, Table 2 and Table 3 show the error of using our improved model to predict stock SH600777, while the last three columns show that using the original model to predict stock SH600777. Obviously, the value of the improved model is significantly smaller than that of the original model from the value of the three evaluation indexes. In addition, in Table 4, t represents the Test Statistic, namely the t-statistic; P

stands for p-value, that is, p-value, representing the probability value corresponding to t-statistic; Obviously, all the three models reject the null hypothesis and accept the alternative hypothesis, that is, there is a significant difference between the two models and the improved model is significantly better than the original model.

Investigate its reason, one is the improved model does not need stationarity assumption, it not only can extract the linear and non-linear information, and by considering the intraday price and base expansion method, can increase function auxiliary information of the original time series prediction, but the original model using original data analysis, due to the stock price data is not smooth, which will lead to larger error; Second, the underlying model structure of the feature components and residual series of the improved model can be determined by the Bagging method, which greatly reduces the error of the model structure through the idea of integration. Thirdly, the improved model provides a parameter selection method based on model averaging for LSTM neural network to select the appropriate dimension of predictor variables, which increases the accuracy of the model. In conclusion, we believe that the stock price prediction model based on the improved LSTM neural network has more excellent forecasting ability than the original model. Obviously, the prediction results of the latter two stocks are almost the same as the prediction results of the first stock (SH600777), and the improved model has a better prediction level than the original model.

4. Conclusion and Discussion

In this paper, the LSTM neural network based on orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighting is used to model and predict three stock prices. Compared with the traditional model, it has significant advantages: Firstly, it does not need to deal with the data stationarity like the classical model. Furthermore, it can not only extract the linear and nonlinear information, but also add the function auxiliary information to the original time series prediction through the intraday price and basis expansion method. Secondly, the latent model structure of feature components and residual series is determined by Bagging method. Finally, it provides a parameter selection method based on model averaging for LSTM neural network to select the appropriate dimension of predictor variables. Therefore, although the stock price has the characteristics of high noise, high dimensionality, and difficult to capture information, the stock price prediction model based on the orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighted LSTM neural network can indeed predict the future stock price through historical data, and it is relatively accurate. In this paper, we believe that this prediction method will have certain discussion value in the future: applying the effect of other fields to explore; For high frequency time series data prediction; Whether different extraction methods of function information will affect the results, whether there are differences, etc.

References

[1] Rahimi, Z. H., & Khashei, M. A Least Squares-based Parallel Hybridization of Statistical and Intelligent Models for Time Series Forecasting [J]. Computers & Industrial Engineering, 2018, 118(2): 21-23.

[2] WAN Rui; Institute of Mathematics and Statistics, Changchun University of Technology;. Stock Price Volatility Forecast Based on GARCH Model [J]. Science & Technology Information, 2022, 20(6): 129-132.

[3] Xiong Jinghua; Ru Jing; School of Economics and Finance, South China University of Technology; School of Water Resources and Hydropower Engineering, Wuhan University;. Research on Exchange Rate Forecasting Integrated Model Based on Random Forest and Fuzzy Information Granulation [J]. The Journal of Quantitative & Technical Economics, 2021, 38(1): 135-156.

[4] Sun Shaoyan; SUN Wenxuan; China State-owned Economy Research Center, Jilin University, Jilin University; School of Economics, Jilin University, Jilin University. Research on the Fluctuation Law of RMB Exchange Rate after Joining SDR [J]. On Economic Problems, 2019(2): 42-47.

[5] Zheng Hong, Duan Zhongdong, Wang Zhen, Li Hongwei; School of Civil and Environment Engineering, Harbin

Institute of Technology; Ninbo Shangong Intelligent Security Technology Co., Ltd.;. Structural Damage Detection Based on Autoregressive Model and Kernel Principal Component Analysis [J]. Intelligent Building & Smart City, 2020(5): 15-19.
[6] Bao Yueyan. Covariance matrix prediction model based on LSTM using high frequency data [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2022(4): 1-7.
[7] Zeng Lifang, Li Liping. The Bank of China Stock Price Forecast Empirical Analysis Based on BP Neural Network and GARCH Model [D]. Yunnan: Chinese Master's Theses Full-text Database, 2020.
[8] Yang Guijun, Du Feia, Ji Xiaolei, School of Statistics, Tianjin University of Finance and Economics; China Center for Economic Statistics Research, Tianjin University of Finance and Economics;. Financial Early Warning Model of BP Neural Network Based on the First and the Last Quality Factors [J]. Statistics & Decision, 2022, 38(3): 166-171.
[9] Zhang, G. Time Series Forecasting Using a Hybrid Arima and Neural Network Model [J]. Neurocomputing, 2003, 50(0): 5-7.
[10] Anna, M., Aurelia, R., Artur, D., & Joachim, P. Forecasting of Natural Gas Consumption in Poland Based on Arima-lstm Hybrid Model [J]. Energies, 2021, 14(24): 12-13.