Research on Prediction Algorithm of College Entrance Examination Filing Line Based on ARIMA and LSTM

Yifan Yan [1], Zuxu Dai[2]

{1442094315@qq.com[1], zxdai@wit.edu.cn[2] }

School of Optical Information and Energy Engineering, School of Mathematics and Physics, Wuhan Institute of Technology, Wuhan 430205, Hubei,China

Abstract: To improve the prediction accuracy of college entrance filing line, this study uses ARIMA-LSTM combined model to predict the rank of college entrance filing line based on score-to-rank conversion table. The model forecasts the rank of filing line for colleges, upon which the admission filing line is predicted. The ARIMA model is utilized to analyze linear relationships in the data, and its autoregressive coefficients set the time steps for the LSTM model, which addresses the nonlinear aspects of the forecast. The predictive results of the combined model are compared with those of the standalone ARIMA and LSTM models. The experimental results show that at the 90 % confidence level, the prediction error confidence interval of the ARIMA-LSTM combined model is (0.2, 3.6), which surpasses the ARIMA model's interval of (3.5, 6.6) and the LSTM model's interval of (-6.3, -2.7). This demonstrates the combined model's efficiency and accuracy in forecasting college entrance filing line.

Keywords: filing line; ARIMA model; LSTM model; ARIMA-LSTM combined model

1 Introduction

In the higher education entrance examination system, the filing line refers to the lowest score line required for candidates to enter the admission stage of a specific university after completing the voluntary application. This is the decisive factor for students to be admitted, and with the release of the college entrance examination results, the total number of candidates, the provincial control line and other relevant information will be released accordingly.

At present, for the analysis and prediction of the college entrance examination filing line, there are research methods such as line difference method, exponential smoothing method, autoregressive integrated moving average model (ARIMA), artificial neural networks (ANNs) and so on. For example, reference[1] used an improved gray prediction method to predict the filing line; the literature[2] used the attention mechanism and information fusion method to predict; literature[3] combined with the TensorFlow framework to use artificial neural networks and complete the prediction of the college entrance examination filing line.

Based on the existing research methods, this study proposes an ARIMA-LSTM combination model[4]. By analyzing the corresponding relationship between the rank of the filing line and the filing line, this model uses the ARIMA model to model the relevant information such as the rank of the filing line, the provincial control line, the lowest score, the highest score, and the number of candidates, and obtains the autoregressive coefficients of the model residual and the

EIMT 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347733

score difference as the input and time step of the LSTM model. First, the ARIMA model is used to predict the rank of the filing line. Then, the LSTM model is used to train and predict the residuals of the rank values, so as to correct the prediction results of ARIMA. The experimental results show that the combined model shows higher prediction accuracy than the single model and the single variable model in predicting the filing line.

2 Model construction

2.1 ARIMA model

Autoregressive Integrated Moving Average Model (ARIMA) is a time series prediction model proposed by American statisticians Box and Jenkins in 1970[5]. It consists of three parts : Autoregressive Model (AR), which is used to describe the linear relationship between variables; moving Average Model (MA) is used to describe the dependence of random errors. And differential processing is used to convert non-stationary time series into stationary series[6].

After the random sequence is subjected to d-order difference and reaches a stationary state, the ARIMA model obtains the autoregressive order p through the AR part and the moving average order q through the MA part, and then obtains the prediction model of the time series. The formula of ARIMA model is:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t\,, \quad t,d\in N \tag{1}$$

In the formula, φ_i and θ_i represent the coefficients of the AR model and the MA model; ε_t is the error term; p and q respectively represent the order of the model; L is the lag operator, d is the lag order of the model[7]. The ARIMA model is based on the Autoregressive and Moving Average Model (ARMA), and the non-stationary time series is processed by adding a differential process to achieve a stationary state[8].The mathematical expression of ARMA model is:

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q} \tag{2}$$

In this expression, a_i and b_i are the parameters in the autoregressive and moving average processes, respectively.

2.2 Long Short-Term Memory Network

Long Short-Term Memory Network (LSTM) is a variant of Recurrent Neural Network (RNN)[9].It not only inherits the feedback neurons of RNN model and the ability to process time series data, but also effectively solves the limitations of RNN in dealing with long-term dependence problems by introducing gating mechanism[10], so that it shows higher applicability in various application scenarios.

Through the introduction of gating mechanism, the LSTM model selectively retains or deletes information during training, thereby improving the accuracy of prediction[11]. The gating mechanism of the LSTM model includes input gate $\boldsymbol{i}_t$, forgetting gate $\boldsymbol{f}_t$ and output gate $\boldsymbol{o}_t$. LSTM also adds a new internal state $\boldsymbol{c}_t$ to achieve cyclic forward propagation during information transmission. The calculation formula of each door is as follows:

$$\boldsymbol{i}_t = \sigma(W_i X_t + U_i h_{t-1} + b_i) \tag{3}$$

$$\boldsymbol{f}_t = \sigma\left(W_f X_t + U_f h_{t-1} + b_f\right) \tag{4}$$

$$\boldsymbol{o}_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \tag{5}$$

W and *U* are weight matrices, *b* is the offset, X_t is the input value of the current time, which represents all the eigenvalues of the t[th] year; the calculation formula of the internal state c_t of LSTM is[12]:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

Among them, $\tilde{\mathrm{c}}_\mathrm{t}$ is the candidate state, which is calculated by the nonlinear function tanh:

$$\tilde{\boldsymbol{c}}_t = tanh(W_c X_t + U_c h_{t-1} + b_c) \tag{7}$$

Finally, the output state h_t is the external state calculated by the internal state $\mathrm{c_t}$ and the output gate $\mathrm{o_t}$, which is responsible for transmitting information to the next layer:

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t) \tag{8}$$

In this process, the forgetting gate f_t determines the extent to which the internal state from the previous moment c_{t-1} can be retained in the current moment c_t; the input gate i_t and the candidate state $\tilde{c}_t$ decide how much information from the current input value X_t can be preserved in c_t. Meanwhile, the output gate o_t controls the amount of information from the internal state c_t hat can be output to the external state h_t[13].

2.3 ARIMA-LSTM combined model

In this study, an ARIMA-LSTM combined model was constructed by combining the linear prediction ability of the traditional ARIMA model on time series data and the nonlinear memory characteristics of the LSTM model[14], and the model was applied to the multi-dimensional feature prediction analysis of the college entrance examination filing line.

Let matrix:

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T1} & \cdots & x_{TN} \\ x_{(T+1)1} & x_{(T+1)1} & \cdots & x_{(T+1)N} \end{bmatrix}, Y = (y_1, y_2, \dots, y_T)^T$$

be a normalized observation data matrix related to the filing line of college entrance examination, where $X_t = (x_{t1}, x_{t2}, \dots, x_{tN})$ represents the independent variable, and $x_{tn}(n = 1,2, \dots, N)$ denotes the *N* characteristic data of the provincial batch filing control line, batch average score, rank value and so on in the *t* year, y_t is the dependent variable corresponding to X_t, which denotes the characteristic data of the location value of the filing line of the province in that year.

Considering that each column vector in matrix M represents the observed values of a specific feature over consecutive years, it can be regarded as a time series. The ARIMA model is established by columns for matrix *M* and column vector *Y* respectively, and the ARIMA model expression set $F = \{f_1, f_2, \dots, f_n\}$ of each column of matrix *M* and the ARIMA model expression f_y of column vector *Y* are obtained. Using the model $f_i, i = (1,2, \dots, N)$ to

recalculate each feature column of the matrix M, the predicted values of the feature matrix *M* and the column vector *Y*:

$$M' = \begin{bmatrix} x'_{11} & x'_{12} & \dots & x'_{1N} \\ x'_{21} & x'_{22} & \dots & x'_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x'_{T1} & x'_{T2} & \dots & x'_{TN} \\ x'_{(T+1)1} & x'_{(T+1)2} & \dots & x'_{(T+1)N} \end{bmatrix}, Y' = (y'_1, y'_2, \dots, y'_T)^T$$

and the predicted value of the next year 's pitch line rank y_{T+1} are obtained. In addition, according to the expression f_y, the regression coefficient p of the rank value of the filing line can also be obtained.

The difference between matrix *M* and column vector *Y* and prediction matrix and prediction column vector:

$$\Delta M = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1N} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta x_{T1} & \Delta x_{T1} & \dots & \Delta x_{TN} \\ \Delta x_{(T+1)1} & \Delta x_{(T+1)1} & \dots & \Delta x_{(T+1)N} \end{bmatrix}, \Delta \mathrm{Y} = (\mathrm{y}_1 - \mathrm{y}'_1, \mathrm{y}_2 - \mathrm{y}'_2, \dots, \mathrm{y}_\mathrm{T} - \mathrm{y}'_\mathrm{T})^\mathrm{T}$$

Represent the residual values of the ARIMA model for matrix *M* and column vector *Y*, respectively. Let matrix：

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1N} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta x_{T1} & \Delta x_{T1} & \dots & \Delta x_{TN} \end{bmatrix}$$

It represents the residual independent variable, where $\Delta X_t = (\Delta x_{t1}, \Delta x_{t2}, \dots, \Delta x_{tN})$ represents the characteristic residual data of the t^{th} year, which is the dependent variable of ΔY residual. According to the judgment of ARIMA model, when the linear relationship in the sequence is completely extracted, the residual of the model will be expressed as a white noise sequence, that is:

$$\Delta X_n \sim (0, \sigma^2), n = (1,2, \dots, N), \ \Delta Y \sim (0, \varepsilon^2)$$

where σ^2 represents the variance of each column of the residual matrix ΔX, and ε^2 represents the variance of ΔY.

In the same time series, the time correlation should be consistent, so the time step of the LSTM model should be consistent with the autoregressive coefficient p of the ARIMA model. The LSTM model is established with the residual matrix as the independent variable and the residual column vector as the dependent variable, and the relationship between and is:

$$\Delta y_t = f_{lstm}(\Delta X_t, \Delta X_{t-1}, \dots, \Delta X_{t-p+1}) \tag{9}$$

That is, Δy_t is related to the residuals of the independent variables at the first p moments. Let the input test matrix be:

$$x_{test} = \begin{bmatrix} \Delta x_{(T-P)1} & \cdots & \Delta x_{(T-P)N} \\ \vdots & \ddots & \vdots \\ \Delta x_{(T+1)1} & \cdots & \Delta x_{(T+1)N} \end{bmatrix}$$

Then the predicted value $\Delta y_{T+1} = f_{lstm}(x_{test})$ of the next year's rank value, so the predicted value of the rank value is：

$$y_{T+1} = y'_{T+1} + \Delta y_{T+1} \tag{10}$$

Finally, using the predicted rank value of the filing line, combined with the known one-point table, the final filing line of a university in the year can be obtained.

2.4 Evaluation indicators

In order to comprehensively evaluate the prediction effect of the proposed model, this paper introduces the ARIMA model and the LSTM model as the research objects of the comparative experiment. The ARIMA model is mainly used for the prediction of one-dimensional data in time series analysis. Therefore, this paper uses the ARIMA model to predict and analyze the rank value *Y* of the filing line, and the final prediction result of the filing line is recorded as A_{arima}. At the same time, in order to explore the application effect of deep learning methods in this field, this study also uses the LSTM model to predict the rank value of the filing line, and the resulting filing line prediction results are recorded as A_{lstm}. Through such comparative analysis, we can more deeply understand the applicability and accuracy of different models in the prediction of the college entrance examination filing line.

In order to comprehensively evaluate the accuracy of the prediction model, this study uses a variety of statistical indicators such as mean error, confidence interval and correlation coefficient to verify, so as to improve the scientificity and persuasiveness of the experimental results. The formulas are as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \delta \tag{11}$$

$$(c_1, c_2) = \left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{12}$$

$$\rho_{\hat{Y}Y} = \frac{Cov(\hat{Y},Y)}{\sqrt{D(\hat{Y})}\sqrt{D(Y)}} = \frac{E(\hat{Y}-E\hat{Y})E(Y-EY)}{\sqrt{D(\hat{Y})}\sqrt{D(Y)}} \tag{13}$$

In the formulas, $\bar{X}$ is the average value of the prediction error, δrepresents the error between the real value and the predicted value of each university. $\hat{Y}$ represents the predicted value of the filing line for 458 colleges, and *Y* is the real value sequence. (c_1, c_2) is the confidence interval and evaluates the uncertainty range of the prediction error. $Z_{\frac{\alpha}{2}}$ represents the critical value of the Z distribution at the significant level α, σ is the standard deviation, and *n* represents the number of samples[15]. $\rho_{\hat{Y}Y}$ represents the correlation coefficient, which reflects the degree of correlation between the predicted value and the true value. $Cov(\hat{Y}, Y)$ denotes the covariance of $\hat{Y}$ and *Y*, *E* is the mathematical expectation, and *D* is the variance[16].

3 Empirical research

3.1 Data selection and description

Based on the data provided by official platforms such as Hubei Provincial Education Examination Institute, this study involves the historical data records of 458 colleges and universities from 2012 to 2022, including 13 key characteristic data of 458 colleges and universities in 11 years. These data features include important indicators such as the location value of the filing line, the total number of applicants for the college entrance examination, and the provincial control line. Compared to the admission score line, which is subject to fluctuations due to multiple factors, the rank of the filing line has a higher tolerance for error due to the significant span between adjacent score segments. Therefore, the core goal of this study is to estimate and analyze the trend of the college entrance examination filing line by accurately predicting the rank value of the filing line.

In terms of data processing, in order to achieve unity and comparability between eigenvalues, this study used the Min-Max standardization method to preprocess the data set. This method effectively maps all data features to the numerical interval of [0, 1], and effectively eliminates the possible effects of different dimensions and numerical ranges, thus ensuring the consistency and accuracy of the data in the subsequent analysis process[17]. This data preprocessing strategy not only provides standardized processing for input data, but also provides a more stable and effective data basis for model training.

3.2 Experimental results and analysis

Taking the data of Hubei college entrance examination as the research object, this study obtains the filing line information of 458 undergraduate colleges and universities enrolled in Hubei Province from 2012 to 2022 through the relevant platforms such as Hubei Provincial Education Examination Institute, and combines the provincial control line and the total number of candidates for the college entrance examination. The digital characteristics such as the number of applicants are used to predict and analyze the trend of the college entrance examination filing line of candidates in Hubei Province. Among them, the filing information of undergraduate colleges and universities enrolled in Hubei Province from 2012 to 2021 is used as the training set, and the relevant filing information in 2022 is used as the test set to verify the predictive ability of the model.

In order to enhance the model 's ability to capture the characteristics of time series data and improve the prediction accuracy, this study uses a bidirectional long-term and short-term memory network model. The model can learn the context relationship in time series data more comprehensively by processing both historical and future information at the same time[18]. The structural design of the bidirectional LSTM model is shown in Figure 1. Its structural feature is that it transmits data forward and backward to capture the bidirectional dependencies in time series data, thereby improving the performance of the model in complex time series prediction tasks.

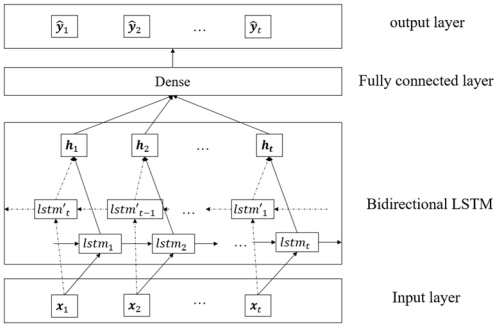

Fig.1 structure of bidirectional LSTM

To prevent information loss during its transfer through the output layer, linear connections are employed in the construction of fully connected layers. This decision stems from the primary function of fully connected layers, which is to integrate feature information from the previous layer and directly transmit it to the output layer. Therefore, avoiding the use of activation functions can maintain the integrity of the feature information and ensure the accuracy of the output results Furthermore, in the construction process of the LSTM model, the Adam optimizer is introduced with a learning rate set to 0.0001. The maximum number of iterations is established at 2000, and an Early Stopping function is incorporated to prevent overfitting. These measures assist in achieving rapid and effective convergence of the model.

In this study, after Min-Max normalization of the data, ARIMA-LSTM combined model was used for model training and prediction analysis. Through the training and prediction of the model, this study obtained the predicted value of the college entrance examination filing line, and compared it with the actual value. In order to verify the accuracy of the combined model prediction, this study compared and analyzed the prediction errors of ARIMA model, LSTM model and ARIMA-LSTM combined model. The prediction error distribution of the three models is shown in Figure 2, and the evaluation indexes such as the error interval of each model under the confidence level of 90 % are given. The specific comparison results are summarized in Table 1.

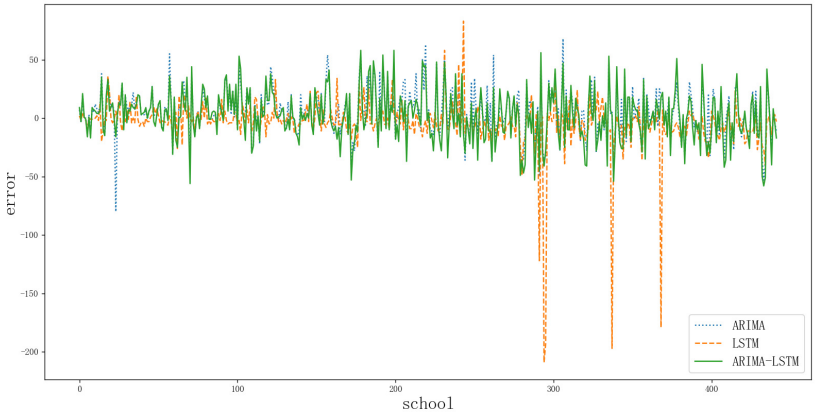

Fig.2 model prediction error diagram

Tab.1 Comparison of model prediction error results

model	Mean Error	Confidence Interval	Correlation Coefficient
ARIMA model	5.0	(3.5, 6.6)	0.906
LSTM model	-4.5	(-6.3, -2.7)	0.881
Combined model	1.9	(0.2, 3.6)	0.891

Data from Figure 2 and Table 1 demonstrate that the ARIMA-LSTM combined model exhibits a smaller range of error fluctuations, and the average value and confidence interval of its prediction error are lower than those of the ARIMA model and the LSTM model. At the same time, he correlation coefficients among the three models do not differ significantly. Considering these evaluation metrics collectively, it can be concluded that the ARIMA-LSTM combined model surpasses the individual ARIMA and LSTM models in terms of predictive accuracy.

4 Conclusion

This study evaluated the above models through different evaluation indicators. Experiments show that compared with the single ARIMA model or LSTM model, the combined multivariate model of ARIMA and LSTM shows higher accuracy in predicting the score difference between the filing line and the provincial control line, and achieves ideal prediction results, which confirms the reliability of the combined model.

The proposal of this model is of great significance for predicting data with large fluctuations. In real life, time series is affected by many factors, resulting in a large deviation in the prediction of data. By analyzing the correlation between the predicted value and other data, the data fluctuation can be reduced and the roundabout prediction can be realized. This method not only avoids the problem of sequence fluctuation, but also improves the accuracy of prediction.

References

[1] Wang Shu, Li Hao, Zhong Ke ,et al. Application of fractional prediction model based on improved gray algorithm[J]. Electronic Technology and Software Engineering, 2020, (11): 212-215.

[2] Hu Ruming. Research on college entrance examination score prediction model and algorithm based on deep learning[D]. Wuhan Institute of Technology, 2022.

[3] Ren Xiangxu. Research on the prediction of college entrance examination score line based on artificial neural network[D]. Jiangxi University of Finance and Economics, 2018.

[4] Jin, Yong-Chao et al. “Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models.” IEEE Access 11 (2023).

[5] Vo, Nguyen and Robert Ślepaczuk. “Applying Hybrid ARIMA-SGARCH in Algorithmic Investment Strategies on S&P500 Index.” Entropy 24 (2022): n. pag.

[6] Zhou Yongdao, Wang Huiqi, Lv Wangyong. Time series analysis and application[M]. Beijing : Higher Education Press, 2015: 192-200.

[7] Vo N, Ślepaczuk R. Applying Hybrid ARIMA-SGARCH in Algorithmic Investment Strategies on S&P500 Index. Entropy. 2022; 24(2):158.

[8] Yao Jinhai, Zou Jiajun. SVM-ARIMA model construction and numerical simulation of CPI prediction[J]. Statistics and Decision, 2022, 38(21): 48-52.

[9] Wang Liya, Liu Changhui, Cai Dunbo, et al. Text sentiment analysis based on CNN-BiLSTM network and attention model[J]. Journal of Wuhan Engineering University, 2019, 41(4): 386-391.
[10] He, M et al. “Application of optimized LSTM in prediction of the cumulative confirmed cases of COVID-19.” Computer methods in biomechanics and biomedical engineering (2023): 1-13 .
[11] Li Shuxian, Zhang Xiaojun, Huchengyu. Policy effect prediction model based on LSTM and its application[J]. Statistics and Decision, 2023, 39(23): 34-39.
[12] Yu, Yong et al. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation 31 (2019):1235-1270.
[13] Duan, Wenyang et al. A hybrid EMD-AR model for nonlinear and non-stationary wave forecasting. Journal of Zhejiang University-SCIENCE A 17 (2016)：115-129.
[14] Ci Bicong, Zhang Pinyi. Financial time series prediction based on ARIMA-LSTM model[J]. Statistics and Decision, 2022, 38(11) 145-149.
[15] Hazra A. Using the confidence interval confidently[J]. Journal of thoracic disease, 2017, 9(10): 4125.
[16] Ratner B. The correlation coefficient: Its values range between+ 1/− 1, or do they?[J]. Journal of targeting, measurement and analysis for marketing, 2009, 17(2): 139-142.
[17] Zhang Heng, Wang Wei, Sun Xuelian. Air quality prediction of Dalian city based on ARIMA-LSTM model[J]. Modern Computers, 2022, 28(18): 75-80.
[18] Jiang Yiqi, Zhao Tongzhou, Chai Yue, et al. Method of sports news subject word extraction based on BiLSTM-CRF[J]. Journal of Wuhan Institute of Technology, 2020, 42(01): 102-107.