Autoregressive Moving Average Model and Improved LSTM Neural Network Applied in Epidemic Prediction in Zhejiang Province

Lingjie Kong*
College of computer & information science. Southwest university. Beibei Chongqing China 400715
*Corresponding author: lingjiekong@swu.edu.cn

Abstract. Dengue is an acute arbo-borne disease caused by the transmission of Dengue virus by mosquitoes, its incidence is closely related to the local meteorological conditions. Zhejiang Province, as a hotspot of dengue fever, is very suitable for the transmission of dengue fever under climatic conditions. Therefore, fitting meteorological conditions is of great significance for the analysis and prediction of the incidence trend of dengue fever in Zhejiang Province. Based on this, to guide the prevention of reasonable dengue fever, we used the autoregressive moving average model (ARIMA) to analyzed the incidence data of dengue fever in Zhejiang Province in recent 15 years, and filled in the missing items in the data; and through the long short-term memory network (LSTM) model, we established the relationship between the incidence of dengue fever and meteorological factors based on meteorological data of the same period. The experimental results have shown that the MAPE fitted by this method is 43.2, compared with 107.3 of the traditional SIR models, indicating that this method is a fitting method with high accuracy.

Keywords: ARIMA, LSTM, Dengue fever

1. Introduction
Dengue fever (DF) [1] is an acute viral arbo-borne infection which caused by the Dengue virus and transmitted by Aedes mosquitoes. For there isn't currently a specific remedy for DF, severe cases may be bleeding, shock and death [2]. Zhejiang Province is located in the southeast coast of China, has a subtropical monsoon climate. Its natural environment and climatic conditions are very suitable for the growth and reproduction of Aedes albopictus, which is one of the vectors of dengue fever.

In the existing study of epidemic prediction, Wang Yin et al [3] used SIR model to made a preliminary prediction of the development of novel coronavirus. In respect of dengue fever epidemic prediction Wang Qin [4] used ARIMA model to analyze and predict the incidence trend of dengue fever in Guangzhou, it is concluded that the dengue incidence trends into trends and cyclical changes obviously; At the same time, Ge Wenxin [5] et al. found that temperature had the most obvious influence on the incidence of dengue fever among the meteorological factors. However, when analyzing the actual situation, it was found that the SIR model could not include meteorological factors affecting the periodic change of the incidence trend of dengue fever, the ARIMA model could not model the nonlinear
relationship, and there was some error in the mean filling method for the missing values in the data. After our investigation found that Li Xiang et al. [3] used the ARMA model to fill in the missing values of grain yield and fertilization data per unit area in the study on the correlation between grain yield and fertilizer use in the Taihu Lake Basin, which could make the subsequent prediction results more accurate. The long- and short-time memory (LSTM) network model can also solve the problems of nonlinear relationship modeling and adding meteorological factors mentioned above. Therefore, we have adopted the combination of ARMA and LSTM model to conduct time series analysis and prediction of the incidence trend of dengue fever and have obtained relatively accurate results.

2. Data preprocessing

2.1. Source of data
We selected 106 cases of dengue fever incidence data from January 2004 to December 2017 in Zhejiang Province, and 168 cases of mean temperature data. The data were provided by the China Meteorological Science Data Sharing Network and the Public Health Science Data Center.

2.2. Missing data filling based on ARMA model
For the reason that existed large number of missing data, there should be 168 cases of dengue fever incidence data in Zhejiang Province from January 2004 to December 2017, but now 36.9% were missing. Therefore, it’s significance to use scientific methods to fill in the missing data. ARIMA (Autoregressive Integrated moving average model) model is one of the most common statistical models used for time series forecasting which is usually used for stock forecasting, population forecasting and other time series data. we have filled in the missing data based on the prediction of ARIMA model.

ARIMA (p, d, q) model can be expressed as:

\[
(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = (1 - \sum_{i=1}^{q} \theta_i L^i) \epsilon_t
\]

If the homogeneous nonstationary time series of order d \(Y_t\) is set, then \(\nabla^d Y_t\) is a stationary time series. Therefore, it can be set as ARIMA (p, d, q) model:

\[
\lambda(B)(\nabla^d Y_t) = \theta(B) \epsilon_t
\]

For the reason that the time series that can be analyzed and predicted by the ARIMA model must meet the condition of stationary non-white noise series, we have carried the Augmented Dickey-Fuller test (ADF) and white noise test on the data after the first-order difference. We obtained that Test Statistic Value is less than two level values, and p Value is significantly less than 0.05 which indicating the sequence after first-order difference is a stationary sequence; and due to the p Value obtained from white noise is less than 0.05, the sequence after first-order difference is a stationary non-white noise sequence.

Then we have trained the ARIMA model by using the relatively complete data of dengue incidence in Zhejiang Province from 2014 to 2017. And comparing the ARIMA models under various parameters, the highest accuracy ARIMA (0,1,2) has been selected.

\[
y_t = 0.093 - 0.617 \epsilon_{t-1} - 0.383 \epsilon_{t-2}
\]
the incidence of dengue fever in Zhejiang Province was 0.72, indicating that ARIMA (0,1,2) model was accurate in predicting the incidence of dengue fever in Zhejiang Province and could be used to fill in the missing values.

As shown in Figure. 1, 2 and 3.

3. Model building

3.1. The evaluation indexes
In this paper, three indexes including linear regression determination coefficient $R^2$ (also known as coefficient of determination, goodness of fit) which can explain the proportion of all variation in the response dependent variable, Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD) which are the most common indicator of prediction error calculated, were used to judge the prediction performance of various models.

3.2. The establishment of LSTM model combining meteorological factors

3.2.1. Introducing meteorological factors and carrying out correlation analysis. Meteorological factors play an important role in the occurrence and spread of infectious diseases. What's more, dengue fever is an acute mosquito-borne disease, and its outbreak conditions include dengue virus, infectious vectors, susceptible population, suitable environment and climate conditions. In particular, due to the sensitivity of mosquito transmission to weather, more and more studies have pointed out that meteorological factors are important factors affecting dengue fever. At the same time, meteorological factors have many influences on the transmission of dengue fever which not only influencing the media transmission, but also affecting the replication of the virus and the human-mosquitoes transmission. These studies shown that temperature plays a dominant role in dengue transmission. Among them, the rank correlation coefficient (also known as Spearman correlation coefficient) between average temperature and the incidence of dengue fever was 0.172, and the significant level p value of each correlation coefficient was less than 0.05, which was statistically significant. The results have showed that there was a strong correlation between average temperature and the incidence of dengue fever.

3.2.2. LSTM model. Long Short-Term Memory Network (LSTM) is a kind of time cyclic neural network, which is specially designed to solve the long-term dependence problem of general RNN (cyclic neural network). All RNN have a chain form of repeated neural network modules. There are three control gates in the unit structure of LSTM, which are input gate, output gate and forgetting gate, used to protect and control the state of the cell. Each type of gate consists of a Sigmoid neural network layer and a dot multiplication operation. In this study, 20 hidden layers have been used which each layer had 7 hidden layer neurons. And we have adopted one-way LSTM mode, the Beachside was 7, the maximum iteration times epoch was 2000, and the learning rate was 0.0006. This was shown in Figure 4.

The amnesia gate determines what information we discard from the cellular state. It looks at the $h_{t-1}$ (previous output), $x_t$ (current input) and outputs a number between 0 and 1 for each number in the cell state $C_{t-1}$ (previous state). In it 1 represents complete retention, while a 0 represents complete deletion.

The formula of forgetting gate is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Then to determine what information we want to store in the cell state, the input gate is combined with a Tanh layer's candidate vectors $C_t$. Adding the gender of the new topic to the cell state to replace the old object we forgot, and updating the previous status value $C_{t-1}$ to $C_t$. At the same time, multiplying $f_t$ to the value of the previous state to express what you want to forget. And then we added $i_t \ast \tilde{C}_t$ to the value that we get. This is the new candidate value.
\[ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \]  
\[ \hat{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \]  
\[ C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t \]  

Finally, the output gate determines which parts of the cell state we want to export. Then we output the determined parts by letting the resulting cell state through \( \tanh \) (which normalizes the value to between -1 and 1) and multiplying it by the gate output. Combined with the pattern of meteorological factors, we have outputted the final prediction results of dengue incidence through the full linkage layer.

The formula of the output gate is:

\[ O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \]  
\[ h_t = O_t \cdot \tanh(C_t) \]  

4. Solution of the model

4.1. Traditional SIR model
Model 1: Use the SIR model.

Taking the incidence of dengue fever in Zhejiang Province in 2004 as an example, based on the fact that the total population of Zhejiang Province at that time remained unchanged at 49252,000, that is, the impact of births, deaths and population movement was not taken into account, and the infection rate of dengue fever was estimated indirectly by the contact of susceptible persons with infected persons.

The SIR model simulation results for the incidence of dengue fever in Zhejiang Province in 2004 are as Figure 5.

4.2. LSTM model implementation with no ARMA missing value filling
Model 2: An LSTM model with no missing value filling in ARMA is used. Due to the missing value of the dengue incidence data obtained, the missing value has been calculated as the annual mean value to fill it, and then the LSTM model has been used for prediction.

The result was shown in Figure 6.

4.3. LSTM model without combining meteorological factors is implemented
Model 2: LSTM model without combining meteorological factors was adopted. The incidence data have been input into the LSTM network to capture the long-term attributes of dengue incidence data and to make predictive estimates.

The result was shown in Figure 7.

4.4. The implementation of LSTM model combining meteorological factors
In this study, the incidence of dengue fever in Zhejiang Province from January 2004 to December 2017 has been analyzed, with the first 90% data as the training set and the last 10% as the test set. The LSTM neural network has been implemented using Python and TensorFlow framework. In this study, to capture the long-term attributes of dengue incidence data the incidence data have been input into the LSTM network, and then the output results of the LSTM layer have been connected with the corresponding mean temperature data. Finally, the final prediction results have been output through the full connection layer.

The result was shown in Figure 8.
4.5. Comparison of model prediction
Table 1 shows the results of different models, taking the incidence of dengue fever in Zhejiang Province in 2015 as an example. Model 1 is the traditional SIR model, Model 2 is the LSTM model without filling the missing value of ARMA, Model 3 is the LSTM model without combining meteorological factors, and Model 4 is the LSTM model with combining meteorological factors. It can be seen that the prediction accuracy of LSTM model is higher than that of SIR model. And the accuracy of LSTM model improved greatly after the missing value of ARIMA was filled and meteorological factors were combined. Model 2.3.4 was roughly the same as the curve of the true incidence (as shown in Figure 9), but the peak value of the model has shifted to a certain extent after meteorological factors had been added, which was closer to the true incidence of dengue fever in Zhejiang Province in 2015, indicating the influence of meteorological factors on the incidence of dengue fever. Model 2 which without missing value filling of ARIMA was close to the real value curve, but compared with Model 4, the prediction accuracy of the model was slightly improved by using missing value filling of ARIMA.

5. Conclusion
We have taken Zhejiang Province as the research object, using the autoregressive moving average model (ARIMA) to analyze the time series of dengue incidence data in the last 15 years, and to fill in the missing items in the data. Then, the relationship between the incidence of dengue fever and the meteorological factors has been established by using the long short-term memory network (LSTM) model. The experimental results have shown that the ARIMA model combined with meteorological factors and the LSTM neural network model's overall results are optimal and the accuracy is greatly improved, compared with the traditional SIR model, model without ARIMA and model without meteorological factors. In future work, we will further optimize the model by combining other factors related to the incidence of dengue fever, so as to further reduce the error between the predicted data and the true value.

6. Figures and Tables
Tab. 1 MAPE and RMSE parameters predicted by different models

| Model   | MAPE | RMSE |
|---------|------|------|
| Model1  | 107.3| 138.1|
| Model2  | 64.9 | 76.2 |
| Model3  | 72.6 | 83.8 |
| Model4  | 43.2 | 62.3 |

Fig. 1 Temporal changes Fig. 2 Time series change after first-order difference
Fig. 3 ARIMA (0,1,2) predicted value

Fig. 4 Neuronal structure of LSTM

Fig. 5 SIR predicted value results

Fig. 6 LSTM prediction of dengue incidence in Zhejiang province from 2004 to 2017 without ARMA missing value filling

Fig. 7 The LSTM model without combining meteorological factors predicted the incidence of dengue fever in Zhejiang Province from 2004 to 2017

Fig. 8 Comparison of predictions by different models
References

[1] Gong Z Y. Analysis of dengue fever epidemic in Zhejiang Province. Excellent academic papers selection of Zhejiang Provincial Association for Pest Control. 2015.

[2] Wang Zhen, Ling Feng, Liu Ying, Ren Jiangping, Sun Jimin. Epidemiological characteristics of dengue fever in Zhejiang Province, 2015-2019 [J]. Chin J Vector Biol & Control. vol.31(06), pp. 643-647, 2020.

[3] Wang Yin, Luo Shuai, Hu Jun. Research on COVID-19 epidemic prediction and control measures based on SIR model [J]. Journal of Qiannan Normal University for Nationalities. vol.40(04), pp. 58-63. 2020.

[4] Fang Qin, Wang Ying, Liu Zhongming, Wang Kun, Guo Juxuan. Application of time series model in predicting the incidence trend of dengue fever in Guangzhou [J]. vol.31(01). pp. 23-25. 2020.

[5] Li Xiang, David Wei, Gao Hongju, Xu Wenping, Wei Xiaohong. Research on the correlation between grain yield and fertilizer use based on BP neural network [J]. Transactions of the Chinese Society for Agricultural Machinery. Vol.48(S1), pp. 186-192, 2017.

[6] Zhao Jianguo, Jia Qiaojuan, Wang Liying, Ma Wenjun, Xiao Jianpeng, Zhu Guanghu. Research progress on the influence of meteorological factors on the transmission of dengue fever [J]. Modern Preventive Medicine. Vol. 47(22), pp. 4185-4189, 2020.

[7] Tianjia Ma, Tianjia Ji, Guanyu Yang, Yang Chen, Wenbo Xu, Hongtu Liu. Prediction of incidence trend of hand, foot and mouth disease based on short and short time memory neural network [J]. Computer Application, vol. 41(01), pp. 265-269, 2010.