Computer Modeling in Engineering & Sciences
DOI:10.32604/cmes.2020.011013
Article

Tech Science Press

Short-Term Traffic Flow Prediction Based on LSTM-XGBoost Combination Model

Xijun Zhang* **and Qirui Zhang**

College of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
*Corresponding Author: Xijun Zhang. Email: zhangxijun198079@sina.com

Received: 14 April 2020; Accepted: 02 July 2020

Abstract: According to the time series characteristics of the trajectory history data, we predicted and analyzed the traffic flow. This paper proposed a LSTM-XGBoost model based urban road short-term traffic flow prediction in order to analyze and solve the problems of periodicity, stationary and abnormality of time series. It can improve the traffic flow prediction effect, achieve efficient traffic guidance and traffic control. The model combined the characteristics of LSTM (Long Short-Term Memory) network and XGBoost (Extreme Gradient Boosting) algorithms. First, we used the LSTM model that increases dropout layer to train the data set after preprocessing. Second, we replaced the full connection layer with the XGBoost model. Finally, we depended on the model training to strengthen the data association, avoided the overfitting phenomenon of the fully connected layer, and enhanced the generalization ability of the prediction model. We used the Kears based on TensorFlow to build the LSTM-XGBoost model. Using speed data samples of multiple road sections in Shenzhen to complete the model verification, we achieved the comparison of the prediction effects of the model. The results show that the combined prediction model used in this paper can not only improve the accuracy of prediction, but also improve the practicability, real-time and scalability of the model.

Keywords: Traffic flow prediction; time series; LSTM; XGBoost; deep learning

1 Introduction

In the area of Intelligent Transportation Systems, accurate traffic speed prediction plays an important role in traffic control and management. With the development of urban intelligence, people pay more and more attention to the planning of urban transportation. Especially the traffic flow prediction research has attracted great attention from many researchers in this field. In recent years, because time series analysis is an important aspect of traffic flow prediction, traffic flow prediction methods based on deep learning have been shown strong competitiveness in time series analysis. These methods can be utilized to simulate traffic characteristics, such as flow, occupancy and speed, or travel time, and produce expected traffic conditions [1]. The application scenarios of traffic flow forecasting are very wide, for example, short-term air passenger forecasting [2], inland river traffic flow forecasting [3], intelligent traffic system speed forecasting [4], occupancy forecasting [5], traffic flow forecasting [6] and etc. The above phenomena show that traffic flow prediction is a very meaningful subject.

Using deep learning theory, this paper presents a short-term traffic flow prediction method based on the LSTM-XGBoost combination model. The LSTM model is used to train the data, and then the hidden layer neural unit obtained by the full connection is used as the input feature of the XGBoost model and trained. Finally, the model is optimized through network reconstruction of the combined model.

The main contributions of this article are as follows:

1. First, we construct a multi-layer LSTM prediction model, combine the characteristics of LSTM and RNN networks to improve the model, and use the reconstructed LSTM model for training and prediction. Normalizing the data set in the model is to decrease the model's error value.
2. Second, we introduce the XGBoost model. Because the fully connected layer in the LSTM model is prone to overfitting, the fully connected layer in the LSTM model is amended by replacing the XGBoost model. This not only avoids overfitting problems, but also enhances the generalization ability of the model.

2 Related Work

Various prediction methods are applied in different subject areas. Prediction models are generally divided into linear statistical models, nonlinear models, intelligent theoretical models and combination models.

2.1 Linear Statistical Model

Wang et al. [7] proposed an estimation model based on Naive Bayes method to realize the estimation of the traffic flow speed in the road network which is not covered by samples; Zhang et al. [8] established a real-time prediction system based on extended Kalman filter (EKF) to predict future passenger flow based on historical passenger data and recent values; Kumar et al. [9] proposed a short-term traffic flow prediction scheme based on the seasonal ARIMA (SARIMA) model, which uses only limited input data for short-term traffic flow prediction.

2.2 Nonlinear Model

Dou et al. [10] proposed a method of traffic flow prediction based on wavelet analysis and ARIMA model, used wavelet analysis theory to denoise the data, and then used ARIMA model to predict the traffic flow; Shao proposed a chaotic time series prediction method. The phase space method was reconstructed using mutual information and pseudo near-point method, and the largest Lyapunov exponent was obtained with a small amount of data [11].

2.3 Intelligent Theoretical Model

Sha et al. [12] used LNN-based LSTM and GRU networks to predict future passenger traffic, and studied the prediction effect at time steps; Zhang et al. [13] proposed a new method based on multi-layer LSTM, combining multi-source traffic data and multiple technologies to improve the performance of passenger flow prediction; Mou et al. [14] proposed a traffic flow prediction method based on time information enhancement. This model improves the prediction accuracy by capturing the inherent correlation between traffic flow and time information.

2.4 Combination Model

Zhang et al. [15] proposed a prediction framework based on support vector regression (SVR), used random forest (RF) and genetic algorithm (GA) to determine the optimal prediction model parameters;

Duan et al. [16] used a combination of CNN and LSTM to predict traffic flow; Guo et al. [17] proposed a model based on the fusion of support vector regression (SVR) and long-short-term memory (LSTM) neural networks, combined the results of SVR and LSTM into the final output of the prediction model; Duan et al. [18] adopted a hybrid deep neural network prediction model based on grid and road nested convolutional LSTM, which can effectively predict the OD traffic of urban taxis.

At present, most researches on traffic flow prediction are using machine learning and deep learning methods, such as using CNN [19,20], RNN [21], LSTM [3,22] and other models to solve traffic problems in prediction. Fu et al. [23] used LSTM and GRU neural network methods to predict the traffic flow. This is the first time that GRU is applied to traffic flow prediction; Liu et al. [24] combined the convolution and LSTM network to analyze the historical traffic flow data of the predicted points by using the bidirectional LSTM module. Although this method has improved the prediction accuracy to some extent, the data preprocessing has not been studied in depth; Yao et al. [25] proposed a deep multi-view space-time network (DMVST-Net) framework to simulate space-time relationships. Although this method analyzes the spatiotemporal characteristics of the data, it uses three methods to study them separately, which does not reflect the advantages of the combined model; Wang et al. [26] and others proposed a short-term traffic flow prediction model based on LSTM-RNN. Although this model can be adaptively updated according to the prediction accuracy, it does not solve the overfitting problem of this model; Wang et al. [27] and others proposed a short-term traffic flow prediction model based on the CNN-XGBoost hybrid model. Although this model studies the temporal and spatial characteristics of traffic flow, the disadvantage of the CNN prediction model compared to the LSTM model is that it is difficult to perform traffic flow multi-step prediction.The grey prediction model can be used to predict traffic flow and real-time and dynamic data. Yang et al. [28] established a coupling model based on the ARIMA and RSDGM (1,1) models to predict and analyze longitudinal and cross-section traffic flow data; Duan et al. [29] proposed an inertial non-uniform discrete gray model and studied the relationship between the inertial model and the state of traffic flow.

In summary, in view of the shortcomings in the above models, this paper proposes a combined model based on two networks which are LSTM and XGBoost. Through the construction of multi-layer LSTM network to achieve the training of time series data. At the same time, in order to avoid overfitting of the fully connected layer in the LSTM model, we introduce the XGBoost model and use the XGBoost model to replace the fully connected layer in the LSTM model. The output of the hidden layer in the multi-layer LSTM model is used as the input of the XGBoost model and use XGBoost model for training to improve the accuracy of the predicted value.

3 Data Preprocessing

In this section, we mainly introduce the preprocessing ways of time series data to improve the quality of the data and ensure the accuracy of the prediction effect.

3.1 Data Exception and Missing Handling

For the poor quality data samples, there are generally four methods to deal with them [30].

1. If the amount of data is large and there is less abnormal data, we can use the direct deletion method for processing;
2. If there is less missing data and there is no continuous missing phenomenon, we can use the data from the previous moment for processing;
3. If the data is continuous and gradual, we can use the weighted value of data from the previous period for processing;

4. If the data is continuously missing and the historical data is huge, we can use the historical average value at the current time for processing.

Because the data samples in this paper have the characteristics of time series [31], and continuous loss exists in the data, this paper uses the historical average to process the data set. According to the characteristics of the data, the data is divided into two periods: weekdays and weekends. Studying the obtained data sample information, the preprocessing process is as follows: for the data anomalies and missing issues during the weekday, we can use the historical average of the current time for processing; for the data anomalies and missing issues during the weekend, we can use the normal value of the historical moment to fill them. The data before and after preprocessing are shown in Figs. 1 and 2. respectively. By comparing the two Figures, we can see that the data after processing tends to be stable, which is consistent with the value change in this period.

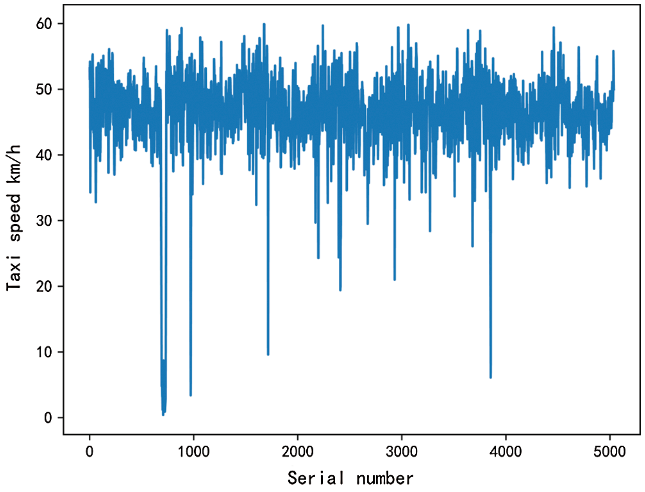

Figure 1: Data information before preprocessing

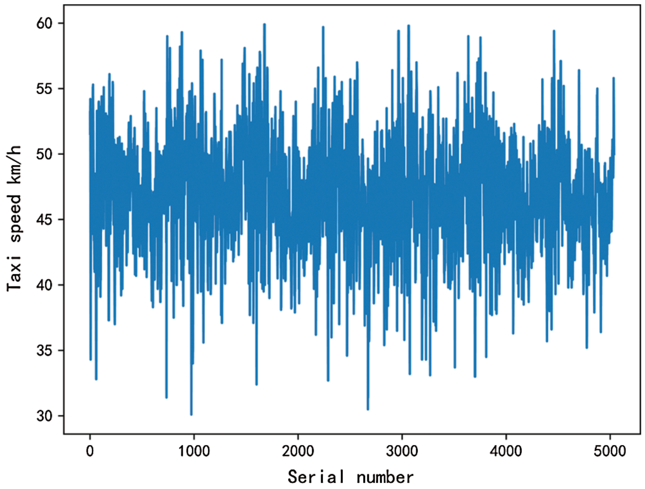

Figure 2: Data information after preprocessing

4 LSTM Model and XGBoost Model

In this section, we mainly introduce the working principles of the LSTM model and the XGBoost model. Among them, Section A introduces the internal structure of LSTM. Section B describes the algorithm mechanism of the XGBoost model.

4.1 LSTM Model

LSTM is an improved recurrent neural network. Because the hidden layer in the original RNN has only one state, it is very sensitive to short-term inputs, and there are also problems of gradient descent and gradient disappearance [32]. LSTM effectively avoids the problems of gradient disappearance and long-term dependence in the RNN model by introducing three different function gate structures [30]. The LSTM model is shown in Fig. 3.

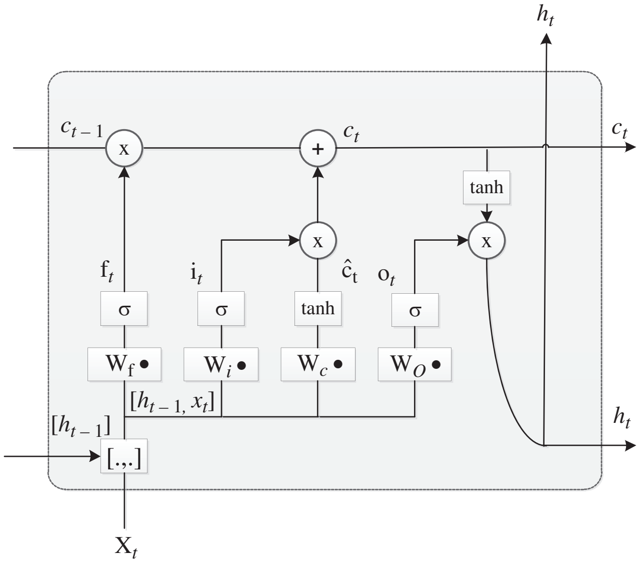

Figure 3: LSTM model structure

The forgot gate f_t determines the number of the memory unit C_{t-1} at the previous moment, which is retained in the current moment C_t. $[h_{t-1}, x_t]$ means joining two vectors into one longer vector, the input of the forgot gate is short-term memory h_{t-1} and the current input x_t. Then we use the forgot gate's weight matrix W_f and bias terms b_f to process. Finally, we introduce the σ function to control, where σ represents the *sigmoid* function [33]. The calculation process of the forgot gate is shown in formula (1):

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

The input gate i_t determines how much the input x_t of the current network moment is saved to the unit state C_t. The input value is firstly processed through the weight matrix W_i and bias term b_i of the input gate, the degree of information retention is determined by the σ layer, and then the input value is processed through the weight matrix W_c and bias term b_c of the calculation unit state. The output $\hat{C}_t$ from the tanh layer is used as the current memory, and finally the current memory $\hat{C}_t$ and long-term memory C_{t-1} are composed into a new state C_t. The calculation process of the input gate is shown in formula (2):

$$
\begin{aligned}
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\hat{C}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t
\end{aligned} \tag{2}
$$

The output gate o_t determines how much the state of the control unit C_t is output to the current output value h_t of the LSTM. The input value is firstly processed through the weight matrix W_o and bias term b_o of the output gate, then introduce the σ function to control. Finally, using the element output C_t by the tanh layer is to form a new output value h_t through multiply output gate o_t by element [34]. The calculation process of the output gate is shown in formula (3):

$$
\begin{aligned}
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \cdot \tanh(C_t)
\end{aligned} \tag{3}
$$

Based on the LSTM model, this paper studies the model structure of different depths, as shown in Tab. 1, and compares the error levels of different models, as shown in Fig. 4. It can be seen that when the depth of the LSTM model reaches four, the prediction effect of the model reaches the optimal value.

Table 1: Different depths for LSTM

Depths	Model structure
1	Input→LSTM→Dense→Output
2	Input→LSTM→Dropout→Dense→Output
3	Input→LSTM→Dropout→LSTM→Dense→Output
4	Input→LSTM→Dropout→LSTM→Dropout→Dense→Output
5	Input→LSTM→Dropout→LSTM→Dropout→LSTM→Dense→Output
6	Input→LSTM→Dropout→LSTM→Dropout→LSTM→Dropout→Dense→Output

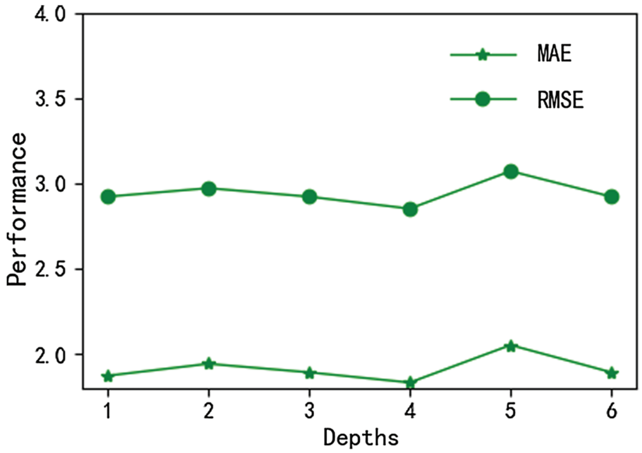

Figure 4: Performance of different depth models

4.2 XGBoost Model

The XGBoost model is improved based on the GBDT model. In the traditional GBDT model, only the first-order Taylor expansion is used. The model is relatively complicated and prone to overfitting. The improved XGBoost model uses the second-order Taylor expansion, and also adds a regularization term, which makes the model simpler and reduces the occurrence of overfitting [35]. The principle of XGBoost model prediction is as follows:

A dataset with m features for n samples $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in R)$. Recorded $\hat{y}_i^{(t)}$ is as the predicted value of the sample x_i in the round t [29], the final predicted value of the sample x_i is shown in formula (4):

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

where, $\hat{y}_i^{(t-1)}$ is the predicted value of the front t wheel, $f_t(x_i)$ is a newly added function. In order to prevent overfitting caused by adding too many nodes, introduce a penalty term to reduce the risk of overfitting. The penalty function $\Omega(f_t)$ is shown in formula (5):

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{\mathrm{T}} \omega_j^2 \tag{5}$$

where, γT is the punishment, $\frac{1}{2}\lambda \sum_{j=1}^{\mathrm{T}} \omega_j^2$ is a penalty term. γ, λ is the coefficient, T is the number of leaf nodes, j is the number of samples, ω_j is the weight [36]. The objective function $obj^{(t)}$ consists of a loss function L and a regularization penalty term Ω, which is defined as shown in formula (6):

$$obj^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant} \tag{6}$$

where, $L(\cdot, \cdot)$ is the loss function, $\Omega(\cdot)$ is a penalty term. Constant is a constant term.

XGBoost algorithm uses the second-order Taylor expansion to optimize the objective function [37]. The expansion formula is shown in formula (7):

$$obj^{(t)} \simeq \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{constant}$$

$$g_i = \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)}), h_i = \partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)}) \tag{7}$$

Then remove the constant term, that is, the difference between the real value and the predicted value of the previous round. The objective function depends only on the first and second derivatives of the error function for each data point. The final simplified form is shown in formula (8):

$$obj^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{8}$$

5 LSTM-XGBoost Combination Model

In this section, we analyze the advantages and disadvantages of the above two models and use the advantages of the two models to design a combined model to make up for each other's shortcomings. Among them, Section A introduces the network structure of the combined model. Section B describes the input matrix based on the multi-step prediction problem. Section C introduces the training process of the combined model.

5.1 Network Structure of Combination Model

For the prediction of traffic speed data, because the nature of the data is a time series problem, this paper analyzes and predicts it from the perspective of time series. Time series data has characteristics of non-linearity and periodicity. It is difficult for traditional time series models based on statistical analysis to solve such problems and to make multi-step predictions on the data. Therefore, this paper chooses the LSTM prediction model for modeling and analysis. However, a single LSTM prediction model has a disadvantage in the fully connected layer, that is to say, there are too many parameters in the fully connected layer. In order to solve the problem of overfitting, the LSTM-XGBoost combined prediction model is introduced. Based on the performance analysis of different deep network models, this paper designs a combination model with a depth of four. First, read the preprocessing time-series speed data by text and store the read data into array A. Then we use 10 consecutive data as the number of neurons in the hidden layer, and process the data in the array by normalizing and reshaping. Second, we use the first four layers of the LSTM model to obtain a residual matrix. Finally, we use the XGBoost model to replace the Dense layer in the LSTM model. At the same time, use the residual matrix as the input of the XGBoost model to train again and output the result into array $\tilde{A}$. This not only solves the overfitting phenomenon in the fully connected layer, but also enhances the generalization ability of the prediction model. The structure of the LSTM-XGBoost combination prediction model is shown in Fig. 5.

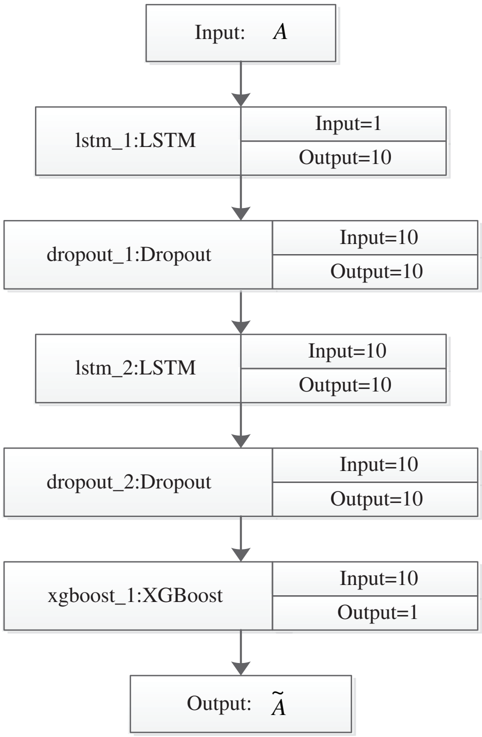

Figure 5: Structure chart of combination prediction model based on LSTM-XGBoost

5.2 Input Matrix Based on Multi-Step Prediction

In order to predict the speed value at multiple times in the future, it is necessary to refer to multiple consecutive historical data. In terms of time, the speed value in the short-term generally does not produce sudden changes, the speed value at the next moment can be regarded as a continuation value at one or more moments. Therefore, in this paper, the data set is divided into two parts, matrix M and matrix N. We set matrix M as the training set and matrix N as the prediction set. The corresponding input matrix is shown in formula (9):

$$M = \begin{pmatrix} x_t & x_{t+1} & \cdots & x_{t+T} \\ x_{t+1} & x_{t+2} & \cdots & x_{t+1+T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t+T} & x_{t+T+1} & \cdots & x_{t+2T} \end{pmatrix} N = \begin{pmatrix} x_{t+T+1} \\ x_{t+T+2} \\ \vdots \\ x_{t+2T+1} \end{pmatrix} \tag{9}$$

where, M is a multi-dimensional matrix composed of the previous continuous velocity data, and N is a one-dimensional matrix composed of the velocity data at the next moment, X_t is the traffic speed at time t, T is the time interval length.

5.3 Training Process of Combined Model

The training process of the LSTM-XGBoost combination prediction model is shown in Fig. 6. First, we perform anomalous and missing data on the weekend and weekday data to obtain an accurate data set. Then, we divide the data into a training set and a test set. We use the data from the training set to enter the LSTM model to obtain the training model. The training data is reprocessed by calling the first four layers of the LSTM model. Next, the processing results of the first four layers of LSTM models are input into the XGBoost model through remodeling and feature extraction, and further processing is performed to finally complete the training process of the combined model.

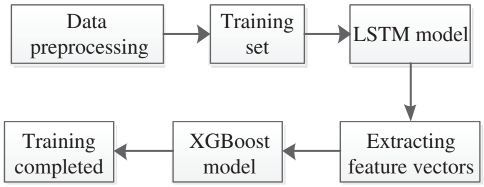

Figure 6: Combination model training process

6 Experiments and Results

In this section, we describe the combination model proposed in this paper in detail, and use the collected data set to verify the model. Among them, Section A describes the metric in the experiment. Section B describes the data set used in this experiment. Section C describes the parameters set in the combined model. Section D compares and analyzes the prediction results of the 4 models.

6.1 Metrics

In order to measure the prediction effect of the combined model, this paper mainly uses the following four performance evaluation indicators to evaluate the model which are the Mean Square Error (Mean Square Error), Root Mean Square Error (RMSE), MAE (Mean Absolute Error), and Mean Absolute Percentage Error (MAPE). The formulas are shown in (10)–(13):

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \left(|y_i - \hat{y}_i|\right)^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(|y_i - \hat{y}_i|\right)^2} \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{13}$$

where, y_i is the real value, $\hat{y}_i$ is the predicted value for the corresponding time. The smaller of the MSE, RMSE, MAE and MAPE values, the better of the model fitting effect and the prediction accuracy.

6.2 Data Set

This paper selects the speed time series of 10 adjacent sections in the "Shenzhen Cup" in 2018 from March 25 to March 31 in 2018 as a data sample. The time interval of this data sample is 2 min. According to the characteristics of the data, the data is divided into two types: weekdays and weekends. The nine sections data and 67% of the last section data are selected as the training set, and the remaining 33% of the last section is used as the testing set.

6.3 Model Parameter Setting

In this paper, the parameters of the LSTM-XGBoost combination model is analyzed through specific experimental procedures. Among them several commonly used cycle lengths were selected for comparison in the experiment, and the results are shown in Tab. 2. The short-term traffic flow is selected within 0~30 min. The comparison results of different batch sizes are shown in Tab. 3. For the Dropout layer, different parameter values have different prediction effects on the model. The comparison results under different dropout rates are shown in Tab. 4.

Table 2: Performance of different Epoch

Epoch	50	100	150	200
MAE	1.90	1.83	1.84	1.85
RMSE	2.94	2.85	2.92	2.97

Table 3: Performance of different batch-size

Batch-size	1	5	10	15
MAE	1.86	1.83	1.93	1.94
RMSE	2.91	2.85	2.99	3.01

Table 4: Performance of different dropout

Dropout	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MAE	1.85	1.83	1.84	1.91	1.95	1.97	2.08	2.22	2.54
RMSE	2.98	2.85	2.93	2.99	3.01	3.07	3.15	3.17	3.55

According to the analysis results of various parameters, the parameter settings of the LSTM-XGBoost model in this paper are as follows: layers is set to 4, epoch is set to 100, batch-size is set to 5, dropout is set to 0.2, activation is set to ‘relu’, optimizer is set to ‘rmsprop’, Loss is set to ‘mse’, and other parameters are set to default values.

6.4 Experimental Results

In order to verify the validity of the combined model, seven models that are CNN model, LSTM model, XGBoost model, LSTM-RNN model, [30], [36] and LSTM-XGBoost model are selected to compare by using the two types of data sets and the parameter setting of the seven models remain the same. The prediction performance indicators of the testing set are shown in Tabs. 5 and 6 and Figs. 7 and 9. Finally, the prediction results of the testing set are shown in Figs. 8 and 10.

Table 5: Workday forecast performance index

Model	MSE	RMSE	MAE	MAPE
CNN	15.33	3.92	3.01	7.31%
LSTM	10.53	3.25	2.16	10.92%
XGBoost	10.46	3.23	2.04	5.10%
LSTM-RNN	9.81	3.13	1.95	10.81%
Literature [30]	9.69	3.11	2.00	10.94%
Literature [36]	9.99	3.16	2.03	10.61%
LSTM-XGBoost	7.55	2.75	1.86	4.19%

Table 6: Weekend forecast performance index

Model	MSE	RMSE	MAE	MAPE
CNN	14.75	3.84	3.00	6.97%
LSTM	4.43	2.11	1.58	7.08%
XGBoost	4.16	2.04	1.50	3.32%
LSTM-RNN	4.40	2.10	1.57	6.93%
Literature [30]	4.30	2.07	1.56	6.91%
Literature [36]	4.47	2.11	1.58	7.05%
LSTM-XGBoost	4.10	2.03	1.49	3.3%

As it can be seen from Tab. 5 and Figs. 7 and 8, we can conclude that the proposed LSTM-XGBoost combination model is better than other models. Compared with the CNN model, the prediction accuracy is improved by 3.12%. Compared with the LSTM model, the prediction accuracy is improved by 6.73%. Compared with the XGBoost model, the prediction accuracy is improved by 0.91%. Compared with the LSTM-RNN model, the prediction accuracy is improved by 6.62%. Compared to the [30], the prediction accuracy is improved by 6.75%. Compared to the [36], the prediction accuracy is improved by 6.42%.

As it can be seen from Tab. 6 and Figs. 9 and 10, we can conclude that the proposed LSTM-XGBoost combination model is better than other models. Compared with the CNN model, the prediction accuracy is

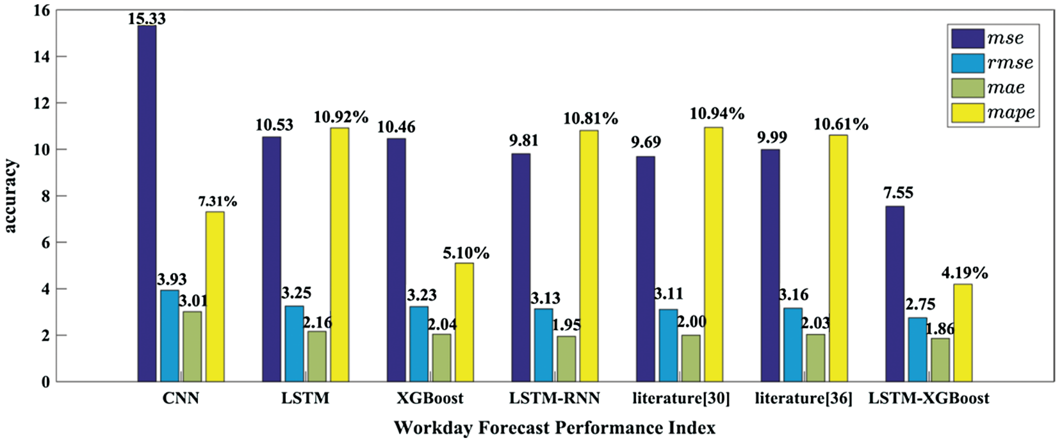

Figure 7: Workday forecast performance index

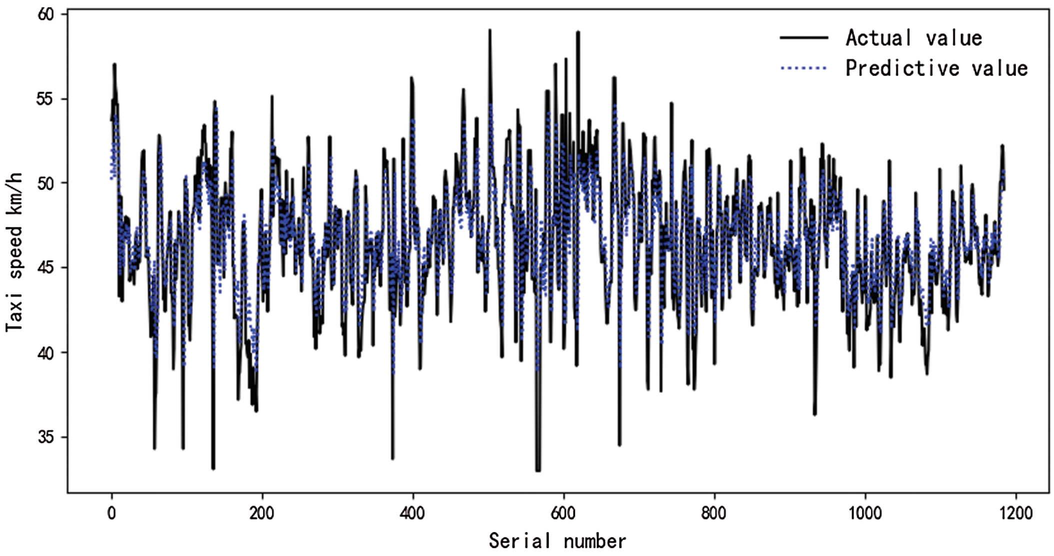

Figure 8: LSTM-XGBoost model workday prediction effect

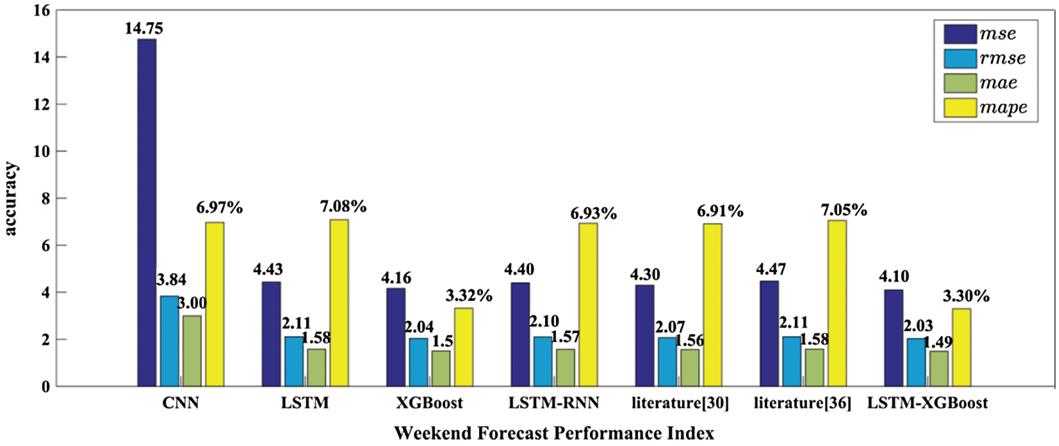

Figure 9: Weekend forecast performance index

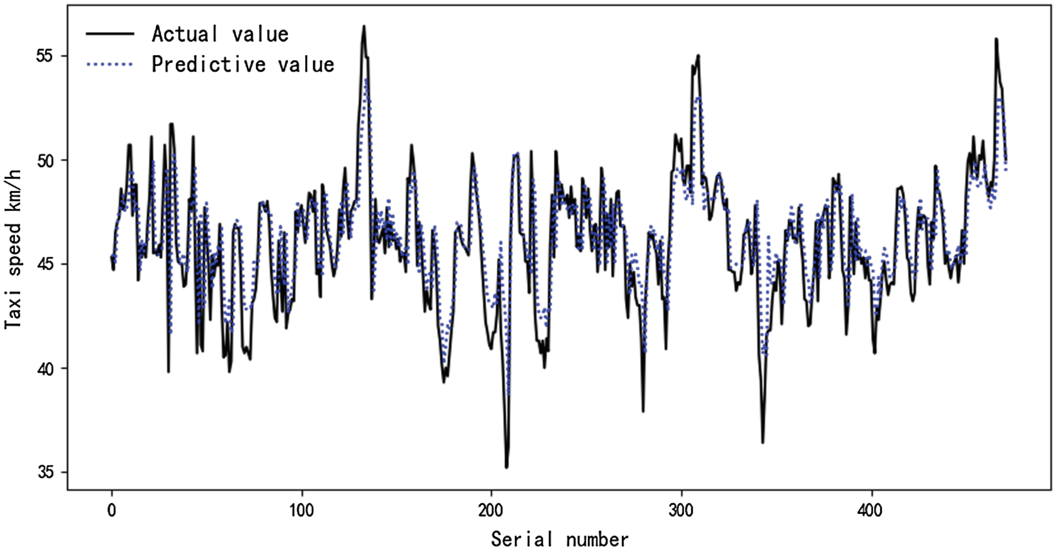

Figure 10: LSTM-XGBoost model weekend prediction effect

improved by 3.67%. Compared with the LSTM model, the prediction accuracy is improved by 3.78%. Compared with the XGBoost model, the prediction accuracy is improved by 0.02%. Compared with the LSTM-RNN model, the prediction accuracy is improved by 3.63%. Compared to the [30], the prediction accuracy is improved by 3.61%. Compared to the [36], the prediction accuracy is improved by 3.75%.

From the above conclusions, it can be concluded that the proposed LSTM-XGBoost combined prediction model can not only improve the prediction accuracy, but also perform multi-step prediction, which is an effective method for traffic speed prediction.

7 Conclusion

Based on the theoretical framework of deep learning, this paper implements short-term accurate prediction which is based on the LSTM-XGBoost combination prediction model. Through the use of Python to complete data preprocessing, time series reconstruction, and normalization operations. Then we introduce the XGBoost model into the LSTM model to avoid overfitting in the fully connected layer. It can improve the generalization ability and prediction accuracy using the model. From the results we can see the combination model proposed in this article is not only suitable for the prediction of short-term traffic flows, but also can be applied to other related multivariate time series prediction fields to solve different practical problems. Therefore, the method in this paper is feasible and efficient.

Acknowledgement: The authors would like to thank the National Natural Science Foundation of China (61461027) and National Natural Science Foundation of China (61465007) for financial support.

Funding Statement: This study was jointly supported by National Natural Science Foundation of China (61461027, 61465007).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Vlahogianni, E. I., Karlaftis, M. G., Golias, J. C. (2014). Short-term traffic forecasting: where we are and where we're going. *Transportation Research Part C: Emerging Technologies, 43,* 3–19. DOI 10.1016/j.trc.2014.01.005.

2. Xie, G., Wang, S., Lai, K. K. (2014). Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches. *Journal of Air Transport Management, 37,* 20–26. DOI 10.1016/j.jairtraman.2014.01.009.
3. Xie, Z., Liu, Q. (2018). LSTM networks for vessel traffic flow prediction in inland waterway. *IEEE International Conference on Big Data & Smart Computing*, pp. 418–425.
4. Li, L., Qu, X., Zhang, J., Wang, Y., Ran, B. (2019). Traffic speed prediction for intelligent transportation system based on a deep feature fusion model. *Journal of Intelligent Transportation Systems, 23(6),* 605–616. DOI 10.1080/15472450.2019.1583965.
5. Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A. (2018). A deep learning model to predict vehicles occupancy on freeways for traffic management. *International Journal of Computer Science & Network Security, 18,* 246–254.
6. Tan, Z., Li, R. (2018). A dynamic model for traffic flow prediction using improved drn. arXiv preprint arXiv: 1805.00868.
7. Wang, X. M., Peng, L., Chi, T. H. (2016). Speed estimation of urban road network traffic flow based on sparse floating car data. *Journal of Surveying and Mapping, 45(7),* 866–873.
8. Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C. et al. (2017). A real-time passenger flow estimation and prediction method for urban bus transit systems. *IEEE Transactions on Intelligent Transportation Systems, 18(11),* 3168–3178. DOI 10.1109/TITS.2017.2686877.
9. Kumar, S. V., Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review, 7(3),* 1–9. DOI 10.1007/s12544-015-0170-8.
10. Dou, H. L., Liu, H, D., Wu, Z. Z., Yang, X. G. (2009). Traffic flow prediction method based on wavelet analysis and ARIMA model. *Journal of Tongji University (Natural Science), 37(4),* 486–489.
11. Shao, X. Q., Ma, X. M. (2012). Based on Chaos theory's network traffic flow prediction of time series research. *Proceedings of the 2012 Second International Conference on Electric Information and Control Engineering,* pp. 392–395. IEEE Computer Society.
12. Sha, S., Li, J., Zhang, K., Yang, Z., Wei, Z. et al. (2020). RNN-based subway passenger flow rolling prediction. *IEEE Access, 8,* 15232–15240. DOI 10.1109/ACCESS.2020.2964680.
13. Zhang, Z., Wang, C., Gao, Y., Chen, Y., Chen, J. (2020). Passenger flow forecast of rail station based on multi-source data and long short term memory network. *IEEE Access, 8,* 28475–28483. DOI 10.1109/ACCESS.2020.2971771.
14. Mou, L., Zhao, P., Xie, H., Chen, Y. (2019). T-LSTM: a long short-term memory neural network enhanced by temporal information for traffic flow prediction. *IEEE Access, 7,* 98053–98060. DOI 10.1109/ACCESS.2019.2929692.
15. Zhang, L., Alharbe, N. R., Luo, G., Yao, Z., Li, Y. (2018). A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction. *Tsinghua Science and Technology, 23(4),* 479–492. DOI 10.26599/TST.2018.9010045.
16. Duan, Z., Yang, Y., Zhang, K., Ni, Y., Bajgain, S. (2018). Improved deep hybrid networks for urban traffic flow prediction using trajectory data. *IEEE Access, 6,* 31820–31827. DOI 10.1109/ACCESS.2018.2845863.
17. Guo, J., Xie, Z., Qin, Y., Jia, L., Wang, Y. (2019). Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM. *IEEE Access, 7,* 42946–42955. DOI 10.1109/ACCESS.2019.2907739.
18. Duan, Z., Zhang, K., Chen, Z., Liu, Z., Tang, L. et al. (2019). Prediction of city-scale dynamic taxi origin-destination flows using a hybrid deep neural network combined with travel time. *IEEE Access, 7,* 127816–127832. DOI 10.1109/ACCESS.2019.2939902.
19. Gutha, S. (2019). *A deep learning approach to real-time short-term traffic speed prediction with spatial-temporal features*. University of Windsor, Windsor.
20. Zhang, W., Yu, Y., Qi, Y., Shu, F., Wang, Y. (2019). Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. *Transportmetrica A: Transport Science, 15(2),* 1688–1711. DOI 10.1080/23249935.2019.1637966.

21. Fandango, A., Wiegand, R. P. (2018). Towards investigation of iterative strategy for data mining of short-term traffic flow with Recurrent Neural Networks. *Proceedings of the 2nd International Conference on Information System and Data Mining*, pp. 65–69, New York.
22. Wang, M. M. (2017). *Research on short-term traffic flow prediction method based on machine learning*. Chang'an University, Xi'an, China.
23. Fu, R., Zhang, Z., Li, L. (2016). Using LSTM and GRU neural network methods for traffic flow prediction. *31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328.
24. Liu, Y., Zheng, H., Feng, X., Chen, Z. (2017). Short-term traffic flow prediction with Conv-LSTM. *9th International Conference on Wireless Communications and Signal Processing*, pp. 1–6.
25. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y. et al. (2018). Deep multi-view spatial-temporal network for taxi demand prediction. *Thirty-Second AAAI Conference on Artificial Intelligence*. arXiv preprint arXiv:1802.08714.
26. Wang, X. X., Xu, L. H. (2018). Research on short-term traffic flow forecasting based on deep learning. *Journal of Transportation Systems Engineering and Information Technology, 18,* 81–88.
27. Wang, Q. S., Xie, X. S., She, Y. (2019). Short-term traffic flow prediction based on CNN-XGBoost hybrid model. *Measurement and Control Technology, 38(4),* 42–45+72.
28. Duan, H. M., Xiao, X. P., Pei, L. L. (2017): Forecasting the short-term traffic flow in the intelligent transportation system based on an inertia nonhomogenous discrete gray model. *Complexity, 2017,* 1–16.
29. Yang, J. W., Xiao, X. P., Mao, S. H., Rao, C. J., Wen, J. H. (2016). Grey coupled prediction model for traffic flow with panel data characteristics. *Entropy, 18(12),* 454. DOI 10.3390/e18120454.
30. Wang, Q. S. (2019). *Research and application of timing optimization of short-term traffic flow at urban intersections*. University of Science and Technology of China, Hefei, China.
31. Yao, H., Tang, X., Wei, H., Zheng, G., Li, Z. (2019). Revisiting spatial-temporal similarity: a deep learning framework for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence, 33,* 5668–5675. DOI 10.1609/aaai.v33i01.33015668.
32. Liu, J. X., Chen, S. C. (2019). Non-stationary time series prediction based on hybrid gate unit. *Computer Research and Development, 56(8),* 1642–1651.
33. Luo, X., Li, D., Yang, Y., Zhang, S. (2019). Spatiotemporal traffic flow prediction with KNN and LSTM. *Journal of Advanced Transportation, 2019(PT.1),* 537–546.
34. Qing, Z. (2019). *Research on recurrent neural network algorithm for short-term traffic flow analysis and prediction*. Xi'an University of Technology, Xi'an, China.
35. Xiong, Y. (2019). Research on prediction of the use of electronic coupons based on XGBoost. *Computer Science and Application, 9(5),* 1029–1035. DOI 10.12677/CSA.2019.95116.
36. Chen, F., Chen, Z. D. (2019). Application of weighted combination model based on XGBoost and LSTM in sales forecasting. *Journal of Computer System Applications, 28(10),* 226–232.
37. Jiang, Y., Tong, G., Yin, H., Xiong, N. (2019). A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters. *IEEE Access, 7,* 118310–118321. DOI 10.1109/ACCESS.2019.2936454.