International Core Journal of Engineering Volume 6 Issue 5, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

Spatial and Temporal Correlation Prediction of Traffic Flow—a Combined Algorithm Based on k-cnn-lstm

Liang Zhou[1,a], Pan Geng[1,b], Shanjiang Pan[1], Haoju Hu

[1]College of Logistic Engineering, Shanghai Maritime University, Shanghai 201306, China.

[a]578309666@qq.com, [b]pangeng@shmtu

Abstract

Nowadays, with the development of urban modernization, intelligent transportation system is becoming more and more important, especially for the prediction of traffic flow. It is difficult to predict the traffic flow accurately because of its nonlinearity and spatial —temporal correlation. According to the characteristics of traffic flow, this paper adopted a K neighbor algorithm combined with the combination of deep learning algorithm structure, using the K neighbor algorithm selected sites related to the target site traffic, then using a combination of CNN - LSTM deep learning algorithm to dig two site traffic flow data of the relationship between each other to make a preliminary forecast, finally to incorporate all the results through the exponential distribution, the final forecast. The data used in this paper are from Minnesota traffic management system, and the experimental results show that this method has certain advantages compaired with the traditional algorithm.

Keywords

Intelligent transportation, Deep learning, Big data.

1. Background introduction and literature review

Nowadays, the rapid development of intelligent traffic management system facilitates people's daily travel and facilitates the control of urban traffic.Accurate prediction of traffic flow and driving speed is an important part of intelligent traffic.Researchers have been predicting traffic flows for more than 40 years, but the field is still struggling. In recent years, with the development and improvement of traffic infrastructure and related technologies such as data transmission, a traffic information network is forming, which can monitor all kinds of traffic information in real time, and now it is easy to obtain a large amount of traffic data.The sheer volume of data makes it easier for researchers to study traffic flows.Therefore, the research focus in recent years lies in the use of huge data volume to predict the future traffic flow [1].

It is difficult to predict the traffic flow because of its nonlinearity and variability.A lot of research has focused on short-term traffic flow forecasts (5 to 15 minutes).After years of research, three traditional methods are summarized: parametric method, nonparametric method and hybrid algorithm.One of the most extensive traditional research methods is the moving autoregressive mean model (ARIMA).Chen et al. [2] proposed a hybrid model based on ARIMA model, which combines linear ARIMA model with nonlinear GARCH model to capture both the conditional mean and conditional heteroscedasticity of traffic flow sequences.Kumar et al. [3] proposed a seasonal ARIMA model that USES very little data as input to predict future traffic flows.However, in [4], it has been proved that traffic condition data is characterized by randomness. If the data itself cannot meet the characteristics of ARIMA model, the prediction effect is not ideal. Therefore, researchers began to attach importance to some non-parametric and nonlinear methods. ZHAO LIU et al. [5] proved that compared with the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

traditional ARIMA model, the prediction interval performance of KNN model is mixed and not limited by the linear assumption of traffic data.Another nonlinear supervised statistical method, support vector regression (SVM) model, is also favored by many researchers [6,7].H. Chang, y. Lee et al. [7] used the neural network model.Although the above model has been applied in most small-scale applications, it is still difficult to predict due to the obvious limitations of shallow structure depth, which makes it impossible to extract useful information from the huge amount of data.

In recent years, with the continuous improvement of computer processing speed, the deep learning network and its hybrid structure have made a very advanced breakthrough in theoretical research and practical application.Y. Lv, y. Duan [8] et al. used the stacked autoencoder (SAE) model to learn the traffic flow characteristics, and proved the advantages of the SAE model over the multi-layer perceptron.In literature [9], multi-task learning is used to stack automatic encoders on the top layer of neural network.In [10], a deep belief network (DBN) composed of finite boltzmann machine layers was proposed.In [11], an integrated model of four fully connected neural networks is proposed.While most of these models use the full connectivity layer, other types of neural network layers capture spatiotemporal patterns better.

Convolutional neural network (CNN) has been applied to a variety of data, such as images, video and audio.Weight sharing is the main feature of the convolution layer, which reduces the number of trainable variables and better captures locality in data.X.ma, z.dai et al. [12] studied the performance of a CNN model for time series prediction problem, in which spatio-temporal traffic flow data is represented by images.S. Deng, s. Jia et al. [13] proposed a class image representation of spatiotemporal data using convolution layer and integrated learning model.

In addition, in the presence of time data, the recursive neural network shows good performance in time series prediction.Researchers successfully applied the long-short model to the time series prediction problem [14], traffic speed prediction problem [15] and traffic flow estimation without data [16][17].The convolutional neural network is superior in spatial data, and the recursive neural network is superior in temporal data.The question of space and time combines the two.X. Cheng, r. Zhang et al. [18] used the cnn-lstm model to capture the evolution of traffic flow in the traffic network.The convolution layer is followed by the LSTM layer, which applies to the downstream and upstream data respectively.In [19], a structure recursive layer is proposed to transform the road network topology into a recursive layer.In [20], a model of graph convolution from sequence to sequence is proposed for multi-step prediction.

In view of the changing characteristics of traffic flow and the research on related literatures, we proposed a hybrid model, which considered the spatial correlation characteristics of traffic flow and studied the trend changes of data itself, which has certain advantages over some traditional algorithms.

2. Relevant theories

Traffic flow data is defined as the amount of traffic that passes through an area over a period of time.Short-term traffic flow prediction plays an indispensable role in the intelligent traffic information system. It is of great research value both for providing effective guidance to improve the traffic information service system and improving the service base for the development of intelligent city.

2.1 Convolutional neural network (CNN)

CNN is a typical deep neural network.It has the following advantages :1) strong data local feature extraction ability.2) weight sharing.The correlation of traffic flow is reflected in the whole traffic network, which has similar properties and can also prevent the overfitting of the network.3) high scalability.CNNs can also use the back-propagation algorithm to learn network parameters, which facilitates the combination of CNNs and traditional neural network to complete the learning of complex features (such as external features) and achieve the purpose of prediction.A typical CNN consists of input layer, convolution layer, pooling layer, several hidden layers, and output layer.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

2.2 LSTM neural network

RNN is a neural network specially used for processing time series.Unlike traditional networks, RNN allows previously entered "memory" to remain in the internal state of the network. However, traditional neural networks cannot train the time series with a long time lag.In order to overcome the shortcomings of traditional RNN, LSTM is an effective improvement method.LSTM is a special RNN designed to learn sequences with long-term dependencies.The LSTM architecture consists of a set of memory blocks.Each block contains one or more memory cells and three gates, namely the input gate, the forget gate and the output gate.The input gate determines the data that is entered into the network structure.The forgetting gate determines when the previous state is forgotten, thus selecting the optimal delay input sequence.The output gate requires all the results to be computed and the output LSTM cells to be generated.

2.3 K nearest neighbor algorithm

KNN algorithm is a non - parametric method for classification and regression.The KNN algorithm selects the data associated with the current data from the database.In this paper, Manhattan distance is used to represent the relationship between data, as shown below(1)：

$$d_{xy} = \sqrt{\sum_{k=1}^{n} \left| x_k - y_k \right|} \qquad (1)$$

2.4 The proposed methodology

Traditional traffic flow prediction methods focus on the research of historical data and do not pay attention to the spatial correlation of traffic flow.In this paper, we adopt a research method for spatial and temporal correlation of traffic flow data.First, we use KNN algorithm to figure out the k statistical sites associated with the target traffic flow statistical sites, and then use the mixed algorithm of cnn-lstm to figure out the predicted values for the data of k+1 sites, including the target sites, and then calculate the weight of each site according to its exponential distance ratio.Then the weighted sum of the forecast data of the K+1 site is used as the final forecast value of the target site.Because KNN is an unsupervised learning algorithm, we cannot choose the optimal k value at the beginning, so we calculate the optimal value of prediction repeatedly.Previous literature studies have proved that the CNN model can automatically learn and extract features from the original sequence data without scaling or difference. Therefore, we combined it with LSTM and made the CNN model as the front end of the LSTM model.The CNN model is applied to the subsequence of the input data, and the results of these subsequence jointly form a time series. The features extracted from this sequence can be explained by LSTM model. According to the specific data, we will also divide them into different subsequence.The LSTM model has been shown to handle the advantages of long time series.The structure of the entire framework is shown in Figure 1:

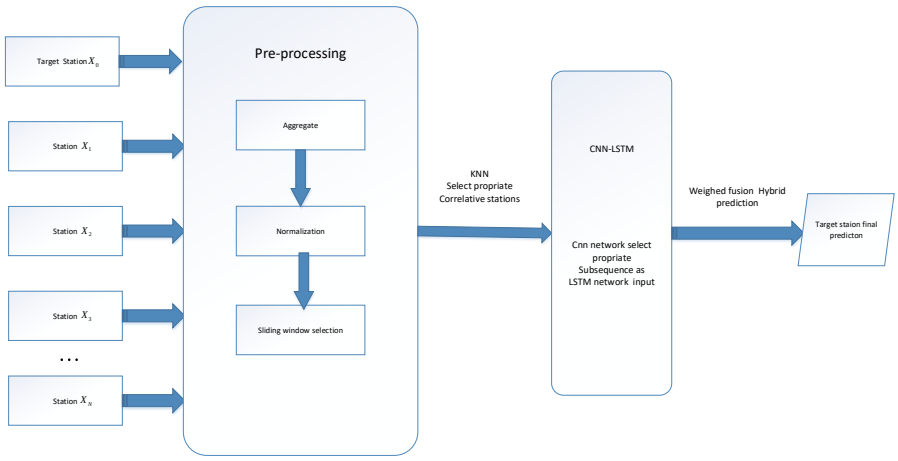

Figure1.Model flow step diagram

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

For the N observation sites considered, if the traffic flow at time t is x_t, m represents the interval of the observed period, as shown in (2) :

$$X_t = \begin{bmatrix} x_1(t) & x_1(t-1) & x_1(t-2) & \cdots & x_1(t-m) \\ x_2(t) & x_2(t-1) & x_2(t-2) & \cdots & x_2(t-m) \\ x_3(t) & x_3(t-1) & x_3(t-2) & \dots & x_3(t-m) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N(t) & x_N(t-1) & x_N(t-2) & \dots & x_N(t-m) \end{bmatrix} \quad (2)$$

For the traffic flow of the target station, is $X_0(t) = [x_0(\mathrm{t}), x_0(\mathrm{t}-1), x_0(\mathrm{t}-2), \dots x_0(\mathrm{t}-\mathrm{m})]$. After that, the appropriate subsequence is selected and input into the cnn-lstm combination model. First, the appropriate number of subsequence is selected, and the features of the subsequence are extracted by CNN, and then input into the LSTM network to predict the traffic flow at time t+1. X_i (i= 0,1,2,3... K) is a traffic station selected by KNN algorithm for the target site. Then the flow at the next moment is (3) :

$$Y_{t+1} = \sum_{i=0}^{k} w_i X_i(t+1) + b \quad (3)$$

w and b are the weight and deviation respectively, and X is the input historical data.After the predicted value of the selected traffic station is obtained, the final predicted value of the target station is obtained by multiplying the predicted value by the weight of the number. The final prediction formula is shown in (4) :

$$Y_{t+1} = \sum_{i=0}^{k} w_i X_i(t+1) \quad (4)$$

w_i is the value of Manhattan distance index ratio obtained by KNN algorithm.According to literature [21], for the weight of k selected relevant traffic stations and test sites, we adopt the rank index method, because it provides a certain degree of flexibility by adjusting the weight dispersion measure (z=2 is used in this study) to assign (k +1) weight.Formula is as follows:

$$w_i = \frac{(k - r_i + 1)^z}{\sum_{i=0}^{k} (k - r_i + 1)^z} \quad (5)$$

Attention: i=0 indicates the weight of the test site.K represents the number of relevant traffic stations selected, and r_i represents the order of the ith station.Z is the measure of weight dispersion, and this text is selected as 2.

3. Experiment

3.1 Data preparation.

The data prepared for the experiment in this paper are from the performance measurement system (PeMS) of California transportation company. Since its establishment, the system has collected historical traffic data of major cities in California every 30 seconds.In each traffic station (VDS), it can be downloaded at different intervals (minimum intervals of 5 minutes).Its data quality is 99.8%. As shown in Figure 2, in our experimental study, we selected 18 stations from the data observation station H to the observation station Y in the following region of Los Angeles, California, USA.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

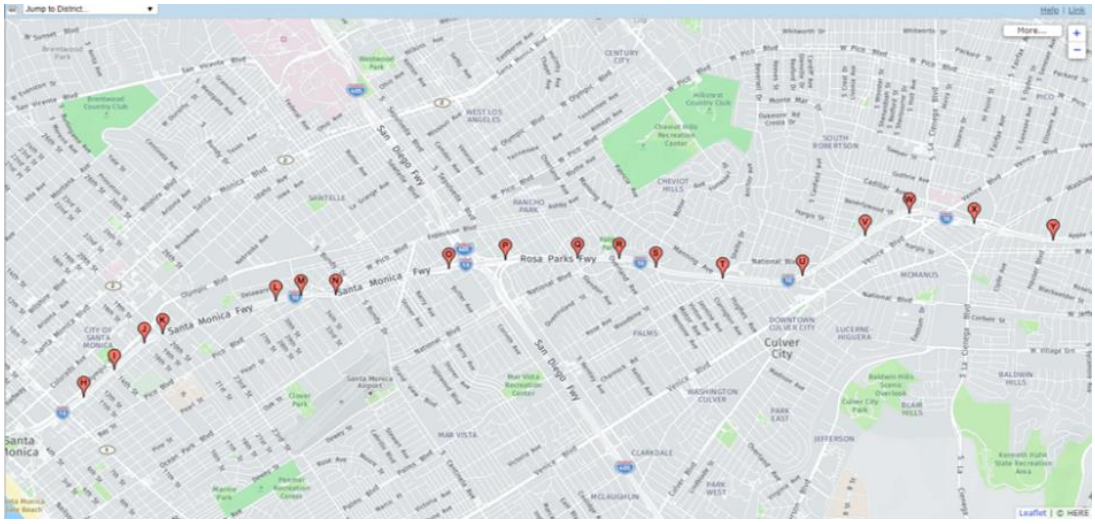

Figure 2.The ID and locations of stations in our experiment.

In experiments, we take the K site as the target, in order to guarantee the similarity of traffic flow data, we chose the weekday traffic flow data as the research object, this article chose the data) on Wednesday, the collected on March 13, 2019 to 2019 of the 1 July all traffic flow data on Wednesday, collect data interval for 5 minutes.We divided the data set into two parts, the first 15 weeks of data as the training set, the last week of data as the test set, the data collected every day 288 groups, each site a total of 16 weeks of 4608 groups of data.

Figure 3 shows the data distribution of 1,152 groups in test station K for four consecutive weeks on Wednesday (from April 24 to May 15). Figure 4 shows the data distribution of test station K and the four adjacent typical stations J,I,P and O on April 3. It is obvious that the data of each day has a similar development trend.

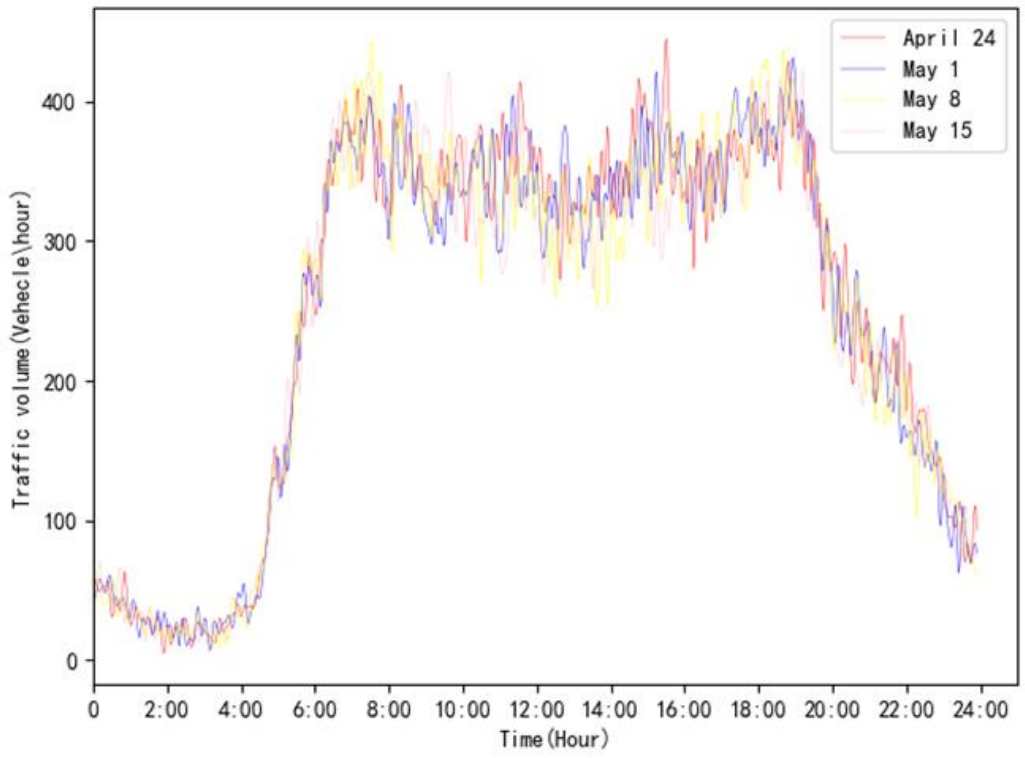

Figure 3.Traffic flows Wednesday for 4 consecutive Tuesdays in the K station .

The two figures prove that the historical similarity of traffic flow data is related to space and time, and also prove the significance of the research direction in this paper.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

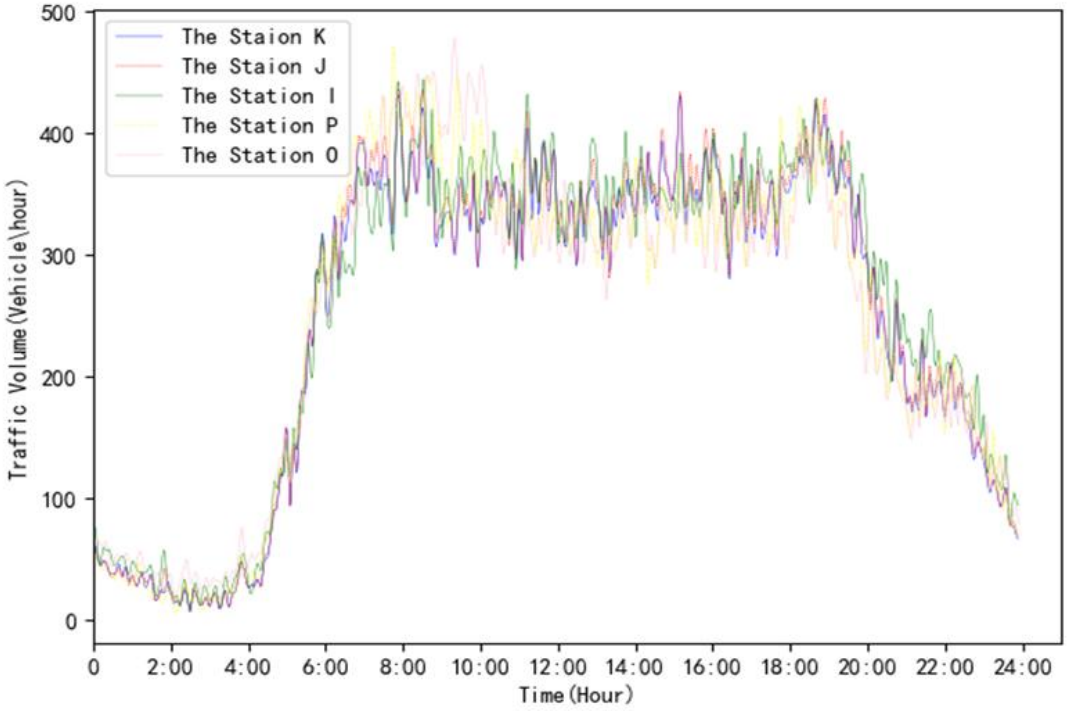

Figure 4 .Traffic flow in the K station and 4 neighboring stations.

3.2 Performance indicators

Generally speaking, the root mean square error (RMSE) is an appropriate evaluation index for the comparison model, but for the sake of accuracy, the paper also chooses the average absolute error (MAE) as the measurement index.

The calculation formula is as follows(6)(7):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{y_i} - y_i)^2} \quad (6)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\overline{y_i} - y_i\right| \quad (7)$$

N is the number of points to predict , $\overline{y_i}$ is the predicted value, y_i is the real traffic flow value.

4. Results analysis and discussion

4.1 Result analysis

In the experiment, we took site K as the test station for analysis.For the cnn-lstm framework selected in this experiment, the time step and the number of subsequences are very important superparameters of the model, which determine the input size of the model and the accuracy of the prediction.After experimental verification, when the time step is set to 20 and the number of subsequences is set to 2, the experimental result is optimal. Two layers of LSTM are set, and the number of cells is 130.In order to verify the validity of the results, we chose several traditional methods for comparison, including support vector regression model (SVR),ARIMA,LSTM, and cnn-lstm.For the support vector regression model, the kernel function is set as radial basis function (RBF), the penalty parameter is set as 300, and the kernel coefficient is 0.008.For the integrated moving average autoregressive model (ARIMA), the coefficient of AR and MA are set to 12 and 2 respectively by the stationarity detection of the data.For the LSTM model, after testing, it is set to two layers of LSTM hidden layer, the number of LSTM cells is set to 130, and the number of iterations is 300.For the cnn-lstm model, the parameter values set are equivalent to the k-cnn-lstm model parameters.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

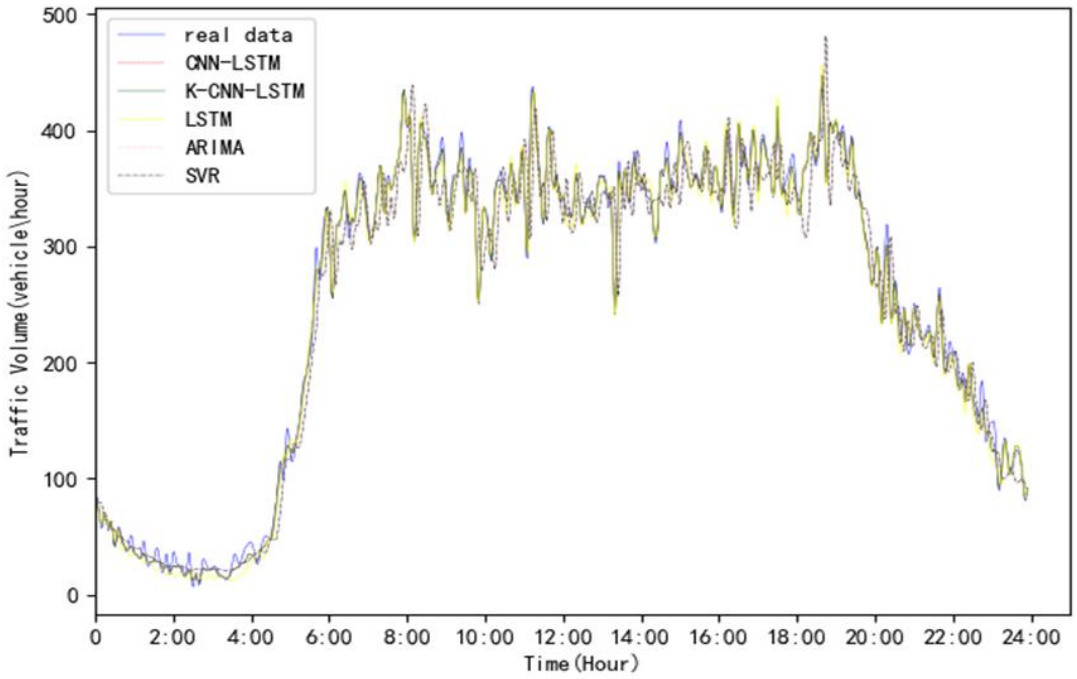

Figure 5 .The real and predicted traffic flow in Station K.

The predicted results and real values of the five models are shown in Figure 5.From the figure analysis, it can be seen that the models measured are in line with the data development trend of real values, while the k-cnn-lstm model is the closest to the real values.

Table 1. Prediction performances of different models.

Models	RMSE	MAE
SVR	23.99	18.639
ARIMA	18.21	14.44
LSTM,	11.18	8.52
CNN-LSTM	9.504	7.358
K-CNN-LSTM	6.21	4.33

RMSE and MAE values of different models are shown in Table1. It can be seen that the two evaluation data of this method are the lowest and the closest to the actual value.Compared with the other four prediction models, RMSE of the improved method increased by 17.78, 12,4.97, 3.294, and MAE by 14.309, 10.11, 4.19, and 3.028, respectively.The traditional SVR model and ARIMA model have poor prediction effects. For the ARIMA model, it assumes that the test data itself is stable, but in the actual test, it does not conform to the non-linear and uncontrollable characteristics of traffic flow data.However, several models of deep learning have better prediction effect.

4.2 Discussion of results

When the historical data of test sites were used for prediction (that is, the cnn-lstm algorithm was used), it was found that RMSE and MAE of k-cnn-lstm algorithm were 3.294 and 3.028 lower than those of RMSE and MAE of k-cnn-lstm algorithm, which proved the spatial correlation of traffic flow data.For the k-cnn-lstm model proposed in this paper, the important point is the choice of K value.Firstly, KNN algorithm is used to calculate the number of sites that are correlated with the test station.The experiment proves that the K value will affect the size of RMSE and MAE, and the evaluation data is the smallest when the K value is set to 4. The four related sites are J,I,P and O.From Figure2 site location, select the relevant stations located in the upstream and downstream of the target

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

site, if according to the traditional theory of space, close the target site to site should be the target site and has more relevance, but from the point of the experimental results, we did not used on the downstream side of the target site choice space distance closer L, M, N site, but a little farther chose the site and P O as related sites, now have a better forecast effect, to some extent illustrates the variability of traffic flow and nonlinear.Experimental results show that the k-cnn-lstm model takes into account both the temporal and spatial characteristics of traffic flow, and has certain advantages over the traditional algorithm model.

5. Conclusion

In this paper, according to the spatial-temporal correlation of traffic flow, we choose an improved k-nearest neighbor algorithm combined with deep learning combination algorithm to predict traffic flow in the future.The KNN algorithm was used to select the site related to the test site to meet the spatial characteristics of traffic flow.Then, the cnn-lstm deep learning combination algorithm is adopted, and the CNN model is used as the front end of the LSTM model. The CNN model is used to extract the characteristics of the input subsequence and then input to the LSTM model. The LSTM model is good at processing long time series.By comparing with other models, the advantages of k-cnn-lstm model are proved.

In this paper, we mainly consider the temporal and spatial characteristics of traffic flow.In the future, we will add weather, road slope and other factors to further improve the accuracy of the prediction.

References

[1] Dongxiao Han1,2, Juan Chen1,3 and Jian Sun " a parallel spatiotemporal deep learning network for highway traffic flow forecasting" International Journal of DistributedSensor Networks 2019, Vol. 15(2).

[2] C. Chen, J. Hu, Q. Meng, and Y. Zhang, "Short-time traffic flow prediction with arima-garch model," Intelligent Vehicles Symposium,pp.607–612, 2011.

[3] S.V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal arima model with limited input data," European Transport Research Review, vol.7, no.3, p.21, 2015.

[4] B.L. Smith, B.M. Williams, and R.K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," Transportation Research Part C Emerging Technologies, vol.10, no.4,pp.303–321, 2002.

[5] ZHAO LIU, XIAO QIN, WEI HUANG, XUANBING ZHU,YUN WEI, ,JINDE CAO, ,JIANHUA GUO." EFFECT OF TIME INTERVALS ON K-NEAREST NEIGHBORS MODEL FOR SHORT-TERM TRAFFIC FLOW PREDICTION". Traffic & Transportation, Vol. 31, 2019, No. 2, 129-139 129.

[6] Y. Zhang and Y. Liu, ''Traffic forecasting using least squares support vector machines,'' Transportmetrica, vol. 5, no. 3, pp. 193–213, Jul. 2009.

[7] M.Castro-Neto,Y.-S.Jeong, M.-K. Jeong, and L. D. Han, ''OnlineSVR for short-term traffic flow prediction under typical and atypical traffic conditions,'' Expert Syst. Appl., vol. 36, no. 3, pp. 6164–6173,2009.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with big data: a deep learning approach, IEEE Transactions on Intelligent Transportation Systems 16 (2)(2015) 865–873.

[9] W. Huang, G. Song, H. Hong, K. Xie, Deep architecture for traffic flow prediction: deep belief networks with multitask learning, IEEE Transactions on Intelligent TransportationSystems 15 (5) (2014) 2191–2201.

[10] T. Kuremoto, S. Kimura, K. Kobayashi, M. Obayashi, Time series forecasting using a deep belief network with restricted boltzmann machines, Neurocomputing 137 (2014) 47–56.

[11] L. Wang, Z. Wang, H. Qu, S. Liu, Optimal forecast combination based on neural networks for time series forecasting, Applied Soft Computing 66 (2018) 1–17.

[12] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction,Sensors 17 (4) (2017) 818.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202005_6(5).0040

[13] S. Deng, S. Jia, J. Chen, Exploring spatial–temporal relations via deep convolutional neural networks for traffic flow prediction with incomplete data, Applied Soft Computing, 2018 (2018).

[14] Z. Zhao, W. Chen, X. Wu, P. C. Chen, J. Liu, Lstm network: a deep learning approach for short-term traffic forecast, IET Intelligent Transport Systems 11 (2) (2017) 68–75.

[15] X. Ma, Z. Tao, Y. Wang, H. Yu, Y. Wang, Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, Transportation Research Part C: Emerging Technologies 54 (2015) 187–197.

[16] Y. Tian, K. Zhang, J. Li, X. Lin, B. Yang, Lstm-based traffic flow prediction with missing data, Neurocomputing 318 (2018) 297–305.

[17] R. Asadi, A. Regan, A convolution recurrent autoencoder for spatio-temporal missing data imputation, arXiv preprint arXiv:1904.12413 (2019).

[18] X. Cheng, R. Zhang, J. Zhou, W. Xu, Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting, arXiv preprint arXiv:1709.09585 (2017).

[19] Y. Kim, P. Wang, L. Mihaylova, Structural recurrent neural network for traffic speed prediction, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5207–5211.

[20] Z. Zhang, M. Li, X. Lin, Y. Wang, F. He, Multistep speed prediction on traffic networks:A deep learning approach considering spatio-temporal dependencies, Transportation Research Part C: Emerging Technologies 105 (2019) 297–322.

[21] F. G. Habtemichael, M. Cetin, and K. A. Anuar, “Methodology for quantifying incident-induced delays on freeways by group ing similar trafc patterns,” in Proceedings of the Transportation Research Record 94th Annual Meeting, Washington, DC, USA,2015.