Short-term Traffic Flow Prediction Based on Time-space Characteristics

Jinxiong Gao, Xiumei Gao, Hongye Yang
Inner Mongolia University of Technology
Hohhot, China
e-mail: 254716786@qq.com, 2953682879@qq.com, 1172840138@qq.com

***Abstract*—In order to accurately predict short-term traffic flow, alleviate traffic congestion and improve traffic operation efficiency, a short-term traffic flow prediction method based on cnn-xgboost is proposed. Combined with the temporal and spatial correlation of short-term traffic flow data, the historical data of this section and adjacent sections are taken as input for prediction. This paper uses convolutional neural networks (CNN) to extract features to reduce data redundancy. An xgboost model with parameters optimized by Drosophila algorithm is proposed for traffic flow prediction. The results show that CNN can effectively extract the traffic flow data under the combination of time and space; compared with SVR, LSTM and other models, the traffic flow prediction error of the improved xgboost model is significantly reduced.**

Keywords -traffic flow prediction; xgboost; convolutional neural network; Drosophila algorithm

I. INTRODUCTION

The research and application of fast and accurate short-term traffic flow prediction technology has important practical significance [1]. It is defined as the next stage evaluation of traffic flow data with a time interval of less than 15 minutes. Since 1960s, scholars at home and abroad have developed a variety of models and methods for short-term traffic flow prediction, which can be divided into traditional mathematical statistics model and machine learning model under nonlinear theoretical prediction. Among them, mathematical statistical models include human time series prediction model[2] , Kalman filter model[3] , grey theory model[4], autoregressive model[5] , Fourier transform model[6] , etc. In recent years, with the continuous expansion of traffic flow data and the rapid development of artificial intelligence, machine learning model has become the mainstream in the research of traffic flow prediction. Kangyanan[7] uses the improved cuckoo search algorithm to optimize the parameters of BP network, establishes the prediction model of traffic flow, and improves the accuracy of prediction. Fusco et al. [8]use specific Bayesian network and artificial neural network Yang et al[9]proposed a prediction model method based on space least square support vector regression. At the same time, many applications of SVM model in short-term traffic flow prediction have achieved good prediction accuracy[10-11]; Tian et al.[12]applied long-term memory neural network (LSTM) to short-term traffic flow prediction, and proved that this method is superior to most other models.

In the existing research, the factors based on statistical model are relatively simple, which can not accurately reflect the nonlinear and uncertain characteristics of traffic flow changes. Based on the nonlinear theory, most of the machine learning models only consider the temporal or spatial correlation of the traffic flow data, but not combine the two. Based on the above research, this paper proposes a method of traffic flow data prediction under the time and space structure, using CNN neural network to extract the features, combining the optimized machine learning model xgboost to predict the short-term traffic flow The research on the implementation of the test is expected to provide some scientific basis for the intelligent transportation and provide reference for the follow-up research.

II. CNN FEATURE EXTRACTION

A. Data Matrix Construction

Urban traffic flow data has temporal correlation and spatial correlation [13] . The traffic flow in time can be regarded as the continuation of the traffic flow in the last time node. The traffic flow data in long time scale has certain periodicity, but in short time scale has time-varying and correlation [14] . From the perspective of space, the closer the two points are, the higher the degree of acquaintance[15] . In this paper, two kinds of correlation data are combined to predict the traffic flow data, reconstruct the original data, and integrate the data into matrix form as the prediction input. The structure is as follows:

(1) Select the adjacent n road sections to collect the traffic flow data of the first m time node of each road section.

(2) Construct a single node time series data, that is $S_i = \begin{bmatrix} t_{i1} & t_{i2} & \cdots & t_{im} \end{bmatrix}$.

(3) Integrate the data of multiple nodes to form the input data of traffic flow prediction,

$$D = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{mn} \end{bmatrix}$$

The input matrix constructed above fully extracts the time and space characteristics of traffic flow data. However,

the input data is too redundant to make the training time complexity of the model increase and the prediction accuracy decrease. In order to solve this problem, convolution neural network is introduced to extract the input matrix features, which can reduce the data redundancy and computation, and improve the prediction accuracy of the model. Convolutional neural network is a kind of multilayer neural network with deep supervised learning structure, which can be regarded as two parts: feature extractor and trainable classifier. The feature extractor consists of feature layers, which are retrieved from the original data by convolution filtering and down sampling.

B. CNN Model Construction

The multilayer convolution neural network consists of convolution layer, pooling layer, full connection layer and output layer. The neurons in the adjacent layer are connected with each other, but there is no connection between the neurons in the same layer. The convolution neural network can be divided into three parts: input part, feature extraction part and prediction output part. As shown in Figure 1.

In Fig. 1, feature extraction mainly includes convolution and pooling. In convolution layer C, convolution is calculated by multiple convolution check inputs, and multiple convolution eigenvectors are obtained. The role of pool layer is to extract local features. Convolutional neural network can extract input matrix features without relying on expert experience.

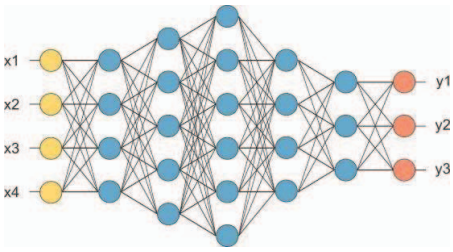

Figure 1. Structure of co nvolutional neural network

Convolution layer: in this paper, the convolution calculation is as follows: (1) input matrix D dimension is $n \times m$, convolution kernel C of $r \times r$ is used for sliding calculation with step size of 1, and the $(n-r+1)(m-r+1)$ dimension eigenvector is obtained by adding bias variable B. In this paper, the feature vector is input into the activation function, and the relu activation function is used. Compared with the activation function such as sigmoid, the relu activation function reduces the computational complexity, and has a faster convergence speed, and the gradient is not saturated. Convolution computing extracts the local characteristics of input matrix by using the characteristics of local connection. In the iterative learning process, the shared weight in convolution kernel is continuously adjusted by using the gradient descent method to maximize the extracted data characteristics.

$$X_{ij} = f\left(\sum_{q=1}^{r} \sum_{p=1}^{r} \left(D_{(i+p)(j+q)} C_{pq} \right) + b_c \right) \quad (1)$$

Pooling layer: after the convolution layer obtains the features, the pooling layer is used for feature aggregation. The purpose is to reduce the amount of calculation and reduce the scale of the convoluted features. The pooling calculation is as follows.

$$X_{ij} = \alpha \cdot \mathrm{Down}\left(X_{ij} \right) + b_p \quad (2)$$

Where α is the pooling weight and b_p is the pooling layer offset weight. Down is a pooling operation. The common pooling operations are mean pooling and Max pooling. When the pool size is l, the range of $l \times l$ in the feature matrix after convolution is selected by sliding operation. Mean pooling is replaced by the average value in the range, and Max pooling is replaced by the maximum value in the range.

III. XGBOOST MODEL AND IMPROVEMENT

A. Drosophila Algorithm

Fruit fly optimization algorithm[16] (FOA) is a global optimization algorithm proposed by Wen Tsao pan inspired by the foraging behavior of fruit flies. Compared with other population optimization algorithms, it is not easy to fall into local optimum. The Drosophila algorithm can be divided into the following steps:

Step1 $S = [s_1 \quad s_2 \quad \cdots \quad s_n]$, $s_i \in DR$, $i = 1,2,3,\cdots,n$,Initial Drosophila population position x_{axis}, y_{axis} .

Step2 random distance and direction of food searching by olfaction

$$x_i = x_{\mathrm{aris}} + RandomValue$$

$$y_i = y_{\mathrm{aris}} + RandomValue \quad (3)$$

Step 3 the Drosophila relies on olfaction to optimize, and calculates the European distance D and olfactory concentration judgment value t of each Drosophila from the origin, such as formula (4) and formula (5).

$$D_i = \sqrt{(x_i)^2 + (y_i)^2} \quad (4)$$

$$T_i = \frac{1}{D_i} \quad (5)$$

Bring the olfactory concentration into the fitness function $f(*)$ to obtain the olfactory concentration $Smell_i$, as shown in formula (6). The best taste concentration $Smell_{best}$ and the

best position of s_{best} in Drosophila population were reserved x_{brst} , y_{best} .

$$Smell_i = f\left(T_i\right) \quad (6)$$

Step 4 visual search, other flies fly to the optimal position according to the vision s_{best} , and get the new population position S again .

$$x_{aris} = x_{best}$$

$$y_{aris} = y_{best} \quad (7)$$

Step 5 iterative optimization, and repeat steps 2 to 4 until the set maximum number of iterations maxgen is reached.

The standard FOA uses the global random search strategy based on the population, and guides the next step of the population search by tracking the information of the current optimal solution, so that the population can carry out the local random search with the current optimal solution as the center, and move towards the better direction.

B. Xgboost model

Xgboost is a parallel regression tree model combined with boosting idea, which is improved by Chen Tianqi et al. [17] on the basis of gradient boosted decision tree (gbdt). Compared with gbdt model, xgboost overcomes the limited calculation speed and accuracy . Xgboost adds regularization to the loss function of the original gbdt to prevent the model from over fitting. In the traditional gbdt, the first-order Taylor expansion is used to calculate the loss function, the negative gradient value is the residual value of the current model, and xgboost is used to carry out the second-order Taylor expansion to ensure the accuracy of the model. And xgboost blocks and sorts each feature, which enables parallel computing when finding the best splitting point, greatly speeding up the computing speed . The model is derived as follows

$$\hat{y}_i = \sum_{k=1}^{n} f_k\left(x_i\right) \quad (8)$$

Where: $\hat{y}_i$ represents n xgboost model of N decision trees $f_k\left(x_i\right)$ series superposition.

$F = \left\{ f\left(x_i\right) = \omega_q\left(x\right) \right\}$, $q : R^m \to T$, $\omega \in R^T$ .Repre sents a function space representing the decision tree, and represents the number of leaf nodes of the decision tree. During the training, a new decision tree function is added to the last prediction value of the original model to minimize the residual with the real value.

C. Parameter Optimization

The selection of parameters directly determines the accuracy of machine learning model. The commonly used methods of parameter adjustment include expert experience method and grid search. The former relies too much on human based judgment, while the latter has the disadvantage that the range of parameter optimization is too narrow and it is not easy to find the optimal parameters. In view of the above problems, this paper proposes to use FOA algorithm to optimize the parameters of xgboost model. Select three groups of parameters -- learning rate (ETA) and maximum depth of tree (max) under the xgboost. Xgbregressor() interface of Python Toolkit_ Depth) and minimum leaf node sample weight (min_ child_ Weight).

When the leaf node is updated, the weight is multiplied by ETA. By reducing the weight of features, the process of lifting calculation is more conservative. The value range is $[0,1]$, and the default value is 0.3.

max _ Depth: control the complexity of decision tree, the larger the value, the more complex the model, when the value is too large, it is easy to appear over fitting. There is no limit and the missing value is 6.

min _ child _ Weight: defines the minimum sum of the weight of the child nodes generated by the observation samples. In the regression model, this parameter refers to the minimum number of samples needed to establish each model. The larger the value is, the more conservative the model is, and the over large model will have the problem of under fitting. The value range is $[0,1]$, the missing value is 1.

Three groups of parameters are used as the optimal solution of FOA algorithm, and the number of Drosophila population is set. The average prediction error of test set is the fitness function, and the parameters of xgboost model are optimized by finite iterations to obtain the optimal model parameters $f\left(*\right)$, max_ depth *min _ child _ weight*.

IV. EXAMPLE VERIFICATION

A. Data

This paper selects the traffic flow data of No.3 and No.32 floating buses in Hohhot as the experimental data. The data collection time is from September 4, 2018 to October 3, 2018. There are 26496 sets of data in each monitoring point. There is a small probability missing in the sample set, and the average value of the upper and lower time nodes of the missing data is used to fill in the experiment. After outlier detection, the input matrix of 35 * 35 is constructed by using the attribute average to replace the data of 175 minutes before the preprocessing of a single sample, which is used to predict the next time node data. If one week of data is taken as test set data every month, 20429 groups of training set samples and 6032 groups of test set samples will be taken.

The core (TM) i5-7500, 8g memory and geforce gtx1050ti computer were used in the experiment. Running in python3.6 environment, the experiment content is written by PyC harm ide.

B. CNN Feature Extraction

Because the input matrix data constructed in this paper is too redundant and dimensionally large, it is not conducive to the training of machine learning model. Different convolution networks are used to extract the features of the data, and the convolution neural network with suitable structure is selected according to the fitting ability of the model. The mean absolute percentage error (MAPE) of training sets under different convolution kernels is shown in Table .

TABLE I. MAP OF TRAINING SET UNDER DIFFERENT CONVOLUTION KERNELS

Serial number	C1	S1	C2	S2	MAPE%
1	3×3	2×2	3×3	2×2	14.7
2	3×3	2×2	5×5	2×2	17.1
3	5×5	2×2	3×3	2×2	14.9
4	5×5	2×2	5×5	2×2	13.6

C. Analysis of Experimental Results

To evaluate the applicability of foa-xgboost model in short-term traffic flow prediction. SVR and LSTM are selected to compare with them.The data of September 15, 2018 and October 3, 2018 were used to compare the prediction results of three models.

TABLE II. COMPARISON OF RMSE AND MAE INDEXES OF THREE MODELS /%

Example	RMSE			MAE		
	FOA-XGBoost	SVR	LSTM	FOA-XGBoost	SVR	LSTM
9.15	9.4	13.7	13.2	8.05	14.4	13.7
10.3	10.3	12.8	12.8	9.87	13.5	13.4

It can be seen from Table II that, compared with SVR and LSTM models, when foa-xgboost model is adopted, RMSE and Mae index values are smaller. It shows that the prediction value of fo-xgboost is more accurate and more reliable than other commonly used models.

V. CONCLUSION

In this paper, the input matrix is constructed based on the spatial and temporal correlation of traffic flow data. Using CNN to extract data features, effectively reduce data redundancy, and greatly improve the efficiency of training calculation under the guarantee of prediction accuracy. In addition, an xgboost model under parameter optimization of FOA algorithm is proposed. Compared with SVR and LSTM, this model is more suitable for traffic flow data prediction under CNN feature extraction. Its RMSE and Mae index are superior to the comparison model, showing more accurate prediction results, and it is an effective traffic flow prediction method.

It should be pointed out that the influence of complex weather on traffic flow prediction is not considered in this paper and the traffic flow prediction model is only for urban roads. In the next stage, the traffic flow prediction in multiple perspectives needs to be considered, which will further strengthen the robustness of the model and make it suitable for the traffic flow prediction in complex situations.

ACKNOWLEDGMENT

R.B.G. thanks Inner Mongolia Autonomous Region Natural Science Foundation "Research on traffic congestion prediction algorithm based on floating bus in Qingcheng, Northern Shaanxi".

REFERENCES

[1] Sun Jingyi, nu ruojin, Liu Yonghua. Study on short-term traffic state forecasting model of SVM considering promotion of large vehicles [J]. Journal of highway and transportation research and development, 2018, 35 (10): 126-132 (in Chinese). [sun Jingyi, Mou ruojin, Liu Yonghua. Study on short-term traffic state prediction model of support vector machine considering large vehicle factors [J]. Highway traffic technology, 2018, 35 (10): 126-132.]

[2] Li Y, Xiao J, Hao Z, et al. Multiple measures-based chaotic time series for traffic flow prediction based on Bayesian theory [J]. Nonlinear Dynamics, 2016， 85（1）：179-194.

[3] Wei H, Cheng Z, Sotelo MA, et al. Short-term vessel traffic flow forecasting by using an improved Kalman model [J]. Cluster Computing, 2017， 2017 （10）：1-10.

[4] Sun Bojun, Yin Weishi. A modified grey model and its app app lication in short time traffic flow [J]. Mathematics in practice & theory, 2016 (23): 201-206 (in Chinese). [sun Bojun, Yin Weishi. Improved grey model and its application in short time traffic flow [J]. Mathematical practice and understanding, 2016 (23): 201-206.]

[5] LI Xiaolei, XIAO Jinli, LIU Mingjun. Vessel traffic flow prediction based on the SARIMA model [J]. Journal of Wuhan University of Technology （Transportation Science & Engineering），2017, 41 (2): 329-332 (in Chinese). [Li Xiaolei, Xiao Jinli, Liu Mingjun study on the prediction of ship traffic flow based on SARIMA model [J]. Journal of Wuhan University of Technology (traffic science and Engineering Edition), 2017,41 (2): 329-332.]

[6] LU Huapu, SUN Zhiyuan, QU Wencong. Repair method of traffic flow malfunction data based on temporal-spatial model [J]. Journal of Traffic and Transportation Engineering, 2015，15（6）：92-100（in Chinese）。 [Lu Huapu, sun Zhiyuan, Qu Wencong. Traffic flow fault data correction method based on time-space model Journal of transportation engineering, 2015,15 (6): 92-100.]

[7] Kang Yanan. Prediction of Expressway flow based on CS Optimized BP neural network [J]. Hightway, 2017 (5): 202-206 (in Chinese). [Kang Yanan, CS Optimized BP neural network highway flow prediction [J]. Highway, 2017 (5): 202-206.]

[8] Fusco G, Colombaroni C, Comelli L, et al. Short-term traffic predictions on large urban traffic net works:Applications of net-work-based machine learning models and dynamic traffic assignment models [C] //International Conference on Models & Technologies for Intelligent Transportation Systems. Budapest，2015：153-158.

[9] Yang Y. A novel prediction method of traffic flow:Least squares support vector machines based on spatial relation[C] //Cota International Conference of Transportation Professionals， 2014：1807-1818.

[10] Zhanquan, Fox, Geoffrey. Traffic flow forecasting based on combination of multidimensional scaling and SVM [J]. International Journal of Intelligent Transportation Systems Research，2014,12（1）：20-25.

[11] Sun Chaodong, Liang Xuechun application of improved flower pollination algorithm optimized support vector machine in short term traffic flow [J]. Computer Engineering & design, 2016,37 (10): 2717-2721 (in Chinese). [sun Chaodong, Liang Xuechun improved flower pollination algorithm optimized SVM application in traffic

flow [J], computer engineering and design, 2016, 37（10）：2717-2721.]

[12] Tian Y, Pan L. Predicting short-term traffic flow by long short-term memory recurrent neural network [C]//IEEE International Conference on Smart City/Socialcom/Sustaincom. IEEE，2015.

[13] QIU Shichong, LU Baichuan MA Qinglu, et al. Traffic flow forecasting based on spatio-temporal characteristic analysis and data fusion [J]. Journal of Wuhan University of Technology，2015（2）：156-160（in Qiu Shichong, Lu Baichuan, Ma Qinglu, et al. Traffic flow prediction based on time-space characteristics analysis and data fusion [J]. Journal of Wuhan University of Technology (information and Management Engineering Edition), 2015 (2): 156-160.]

[14] Zhang Jing, Ren gang. Spatial temporal correlation analysis of urban traffic congestion [J]. Journal of transportation systems engineering and information technology, 2015,15 (2): 175-181 (in Chinese). [Zhang Jing, Ren Gang spatial correlation analysis of urban road traffic congestion [J]. Transportation system engineering and information, 2015, 15（2）：175-181.]

[15] Jiang Yangsheng, Luo Xiaoling, Liu Yuan, et al. Study of optimizing transit network spatial accessibility [J]. Journal of highway and transportation research and development, 2016,33 (4): 102-107 (in Chinese). [J]. Highway transportation technology, 2016,33 (4): 102-107.]

[16] Lei X, Ding Y, Fujita H, et al. Identification of dynamic protein complexes based on fruit fly optimization algorithm [J]. Knowledge-Based Systems, 2016 105（C）：270-277.

[17] Chen T, Guestrin C. XGBoost:A scalable tree boosting system [C] //ACM Sigkdd International Conference on Knowledge Discovery&. Data Mining，2016：785-794.