Journal of Southeast University (English Edition) Vol. 38, No. 1, pp. 92 - 96 Mar. 2022 ISSN 1003—7985

Short-term traffic flow prediction with PSR-XGBoost considering chaotic characteristics

Li Shubin[1,2] Kong Xiangke[1] Li Qingtong[1] Lin Zhaofeng[1] Zhao Zihao[3]

([1]School of Traffic Engineering, Shandong Jianzhu University, Jinan 250101, China)
([2]Department of Traffic Management Engineering, Shandong Police College, Jinan 250014, China)
([3]Beijing Urban Construction Design and Development Group Co., Ltd., Beijing 100017, China)

Abstract: To improve the level of active traffic management, a short-term traffic flow prediction model is proposed by combining phase space reconstruction (PSR) and extreme gradient boosting (XGBoost) algorithms. Firstly, the traditional data preprocessing method is improved. The new method uses hierarchical clustering to determine the traffic flow state and fills in missing and abnormal data according to different traffic flow states. Secondly, one-dimensional data are mapped into a multidimensional data matrix through PSR, and the time series complex network is used to verify the data reconstruction effect. Finally, the multidimensional data matrix is inputted into the XGBoost model to predict future traffic flow parameters. The experimental results show that the mean square error, average absolute error, and average absolute percentage error of the prediction results of the PSR-XGBoost model are 5.399%, 1.632%, and 6.278%, respectively, and the required running time is 17.35 s. Compared with mathematical-statistical models and other machine learning models, the PSR-XGBoost model has clear advantages in multiple predictive indicators, proving its feasibility and superiority in short-term traffic flow prediction.

Key words: traffic prediction; phase space reconstruct; complex networks; model optimization

DOI: 10.3969/j.issn.1003-7985.2022.01.014

As a critical component of intelligent transportation systems, traffic flow prediction plays a vital role in providing traffic state predictions and implementing control measures for traffic management departments. At present, there are mainly two categories of traffic flow prediction methods: model-based and data-driven methods.

Dynamic traffic simulation models can be the typical example of the former, such as DynaMIT-R and Visum-online. However, the model algorithm and software application in this method are very complex, and real-time simulations require substantial computing resources. The latter method is mainly based on historical data and uses statistical analysis and machine learning to make predictions. It has the advantages of a simple algorithm, efficient online operation, and accurate prediction for small-scale road networks. Therefore, this method has also become a research focus in recent years.

Traffic flow prediction methods driven by data can be divided into three categories: 1) linear prediction method based on time series theory and Kalman filter theory, such as the autoregressive integrated moving average (ARIMA) forecasting model[1], and Kalman filter forecasting model [2]; 2) nonlinear prediction method based on chaos theory, such as chaos of traffic system[3] and multi-step prediction algorithm of traffic flow based on chaos theory[4]; 3) intelligent prediction method based on machine learning, such as support vector regression (SVR) model[5], random forest (RF) model[6], and long short-term memory (LSTM) neural network speed prediction model[7-8]. Intelligent prediction methods have gradually become the mainstream method of short-term traffic flow forecasting because of their advantages of accurate forecasting, fast calculation, and good feasibility. However, the current related research mainly focuses on the optimization of model parameters and the exploration of application scenarios. It still lacks the processing of collected data to reflect the traffic situation comprehensively.

In response to the above problems, this paper proposes a short-term traffic flow prediction model combining phase space reconstruction (PSR) and extreme gradient boosting (XGBoost) algorithms, which mainly include four parts: data preprocessing, PSR, complex network analysis, and XGBoost model training and prediction. The advantages of the model in terms of prediction accuracy and calculation speed are verified through examples.

1 Improved Data Preprocessing Method

Data preprocessing is the premise of data analysis, especially the filling of missing and abnormal data. However, traditional data filling methods lack the pertinence of the research object, especially the analysis of the state of

Received 2021-08-25, **Revised** 2021-12-09.

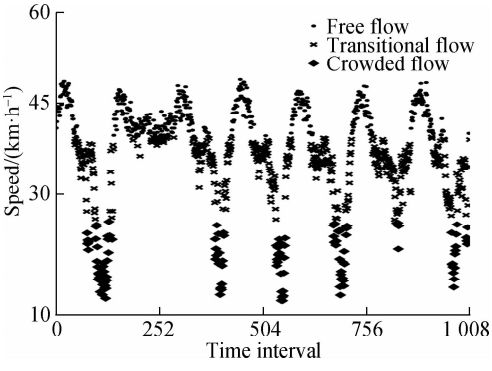
Biography: Li Shubin (1977—), male, doctor, professor, li_shu_bin@153.com.

Foundation items: The National Natural Science Foundation of China (No. 71771019, 71871130, 71971125); the Science and Technology Special Project of Shandong Provincial Public Security Department (No. 370000000159009202100100001, 370000000159009202100120001).

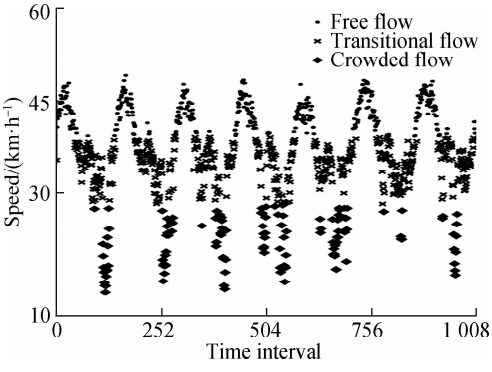
Citation: Li Shubin, Kong Xiangke, Li Qingtong, et al. Short-term traffic flow prediction with PSR-XGBoost considering chaotic characteristics [J]. Journal of Southeast University (English Edition), 2022, 38(1): 92 - 96. DOI: 10.3969/j.issn.1003-7985.2022.01.014.

a transportation system. On this basis, this paper proposes a data filling method based on a hierarchical clustering algorithm and traffic flow state. The data in this paper are the measured speed data of an expressway in Guangzhou from August 1st to September 25th, 2016. The time span is eight weeks, the time interval of data collection is 10 min, and the total number of data sets is 8 064.

The main processes of the hierarchical clustering algorithm used to fill in missing and abnormal values are as follows: 1) Clustering speed data are processed by the hierarchical clustering algorithm. Fig. 1 shows the data clustering results of the first and second weeks in the time series data. The hierarchical clustering algorithm can clearly divide the speed value into three categories. 2) According to the clustering results, the traffic conditions of different categories are identified. According to the basic graph theory, the clustering results can be identified as free flow, transition flow, and congested flow. 3) The traffic state where the missing or abnormal value is located is determined. 4) Fill it with the average value of data in the same state.



(a)



(b)

Fig. 1 Hierarchical clustering results. (a) First week; (b) Second week

2 Data Reconstruction

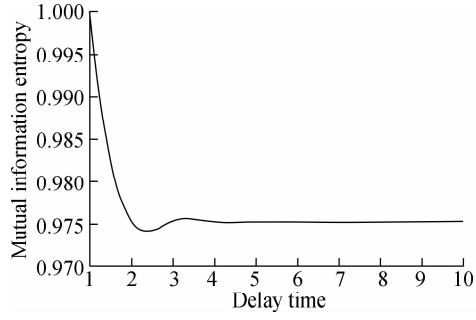
From the perspective of the meso-traffic flow model, the various dimensional states of the transportation system evolve over time to obtain a sequence of multidimensional state parameters to form a phase space. However, in the actual data acquisition process, only a certain dimensional component of the system can be acquired through the detector. Suppose $v = \{v_i \mid i = 1, 2, \ldots n\}$ is the speed sequence of the traffic system, where v_i represents the average speed of vehicles in a certain period of time and n is the length of the time sequence. The one-dimensional component can be reconstructed into a multidimensional data matrix by

$$\boldsymbol{V} = \begin{bmatrix} v_1 & v_{1+\tau} & \cdots & v_{1+(m-1)\tau} \\ v_2 & v_{2+\tau} & \cdots & v_{2+(m-1)\tau} \\ \vdots & \vdots & & \vdots \\ v_{n-(m-1)\tau} & v_{n-(m-2)\tau} & \cdots & v_n \end{bmatrix} \tag{1}$$

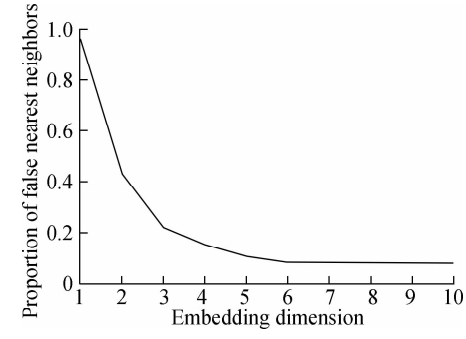
where $\boldsymbol{V}$ is the data matrix after reconstruction; τ is the delay time; m is the embedding dimension.

The selection of parameters is the key to PSR. For the delay time τ, the mutual information method is used[9-10]. For two sequences, mutual information entropy is positively correlated with the degree of correlation between two sequences. Therefore, by calculating the mutual information entropy of the initial sequence and the sequence delay by τ, the correlation between the sequences can be determined. Fig. 2(a) shows the variation of mutual information entropy with delay time, where the best delay time is 2.

For the embedding dimension m, the false k-nearest neighbor method is used[9-10]. The calculation idea of the



(a)



(b)

Fig. 2 Phase space reconstruction parameter selection. (a) Delay time; (b) Embedded dimension

false k-nearest neighbor method is to gradually eliminate false neighbors by increasing the dimensionality of the phase space until the proportion of false nearest neighbors remains unchanged. Fig. 2(b) shows that as the embedding dimension increases, the proportion of the false k-nearest neighbors rapidly decreases and then stabilizes. The acceptable value of the embedding dimension is 6.

3 Complex Network Analysis

The data matrix after PSR reflects the evolution of multidimensional states. When constructing a time series complex network, the vector in the matrix is used as the network node, and the connection between nodes is determined by the node distance and critical threshold. If the node distance is smaller than the threshold, then there is a connection between the nodes.

The choice of the critical threshold is very important to the construction of the network. In this study, the threshold through the network density is examined. In a chaotic system network, there are many clusters of different sizes. As nodes in a cluster are adjacent to one another, the degree will rapidly increase as the threshold within the cluster radius changes. When the threshold is close to the average radius of all clusters, the edge increase will reach the maximum rate, and exceeding the threshold will result in redundant connections between nodes. Therefore, the critical threshold can select the point where the network density grows the fastest. Fig. 3 shows that when the threshold is 25, the network density grows the fastest, so it is used as the network connection threshold.

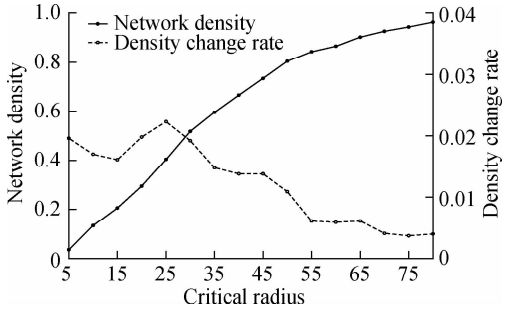


Fig. 3 Selection of the complex network connection threshold

Fig. 4 plots a complex network of the time series of daily period data. From the perspective of the topological structure, the network is mainly composed of clusters of different sizes, without isolated nodes or clusters. From the perspective of quantitative indicators, the Pajek software is used to calculate the degree of nodes in the network. The distribution of node degrees is shown in Fig. 5. Its distribution in double logarithmic coordinates can be fitted with a straight line with a negative slope. The finding shows that the network presents a scale-free characteristic. According to existing research conclusions[11], the network has scale-free characteristics, and the reconstructed time series data has strong robustness and noise resistance and is suitable for prediction.

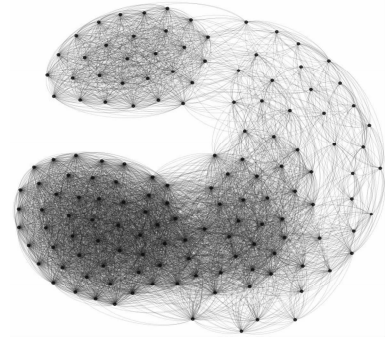


Fig. 4 Complex network topology diagram

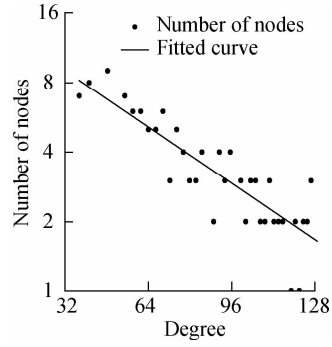


Fig. 5 Degree distribution of network nodes

4 Traffic Flow Prediction

4.1 XGBoost

XGBoost is improved on the basis of the gradient boosting decision tree (GBDT) algorithm[12]. The construction of each tree in XGBoost is completed by fitting the negative gradient of the loss function of the previous model. To make up for the shortcomings of GBDT in terms of prediction accuracy and overfitting, XGBoost introduces regularization in the objective function to quantify the complexity of the tree model. The complexity of the tree model mainly depends on the number and weight of leaf nodes—the lower the complexity of the tree model, the stronger the generalization ability of the model. Intuitively speaking, when the model expects to minimize the objective function, the model is inclined to choose a simple tree model with a strong generalization ability for prediction. The objective function of the XGBoost model can be calculated by

$$L = \sum_j l(y_j, \hat{y}_j) + \sum_k \Omega(f_k)$$
$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2 \quad (2)$$

where L is the objective function; $l(\cdot)$ is the loss function of predicted value and expected value; y_j is the j-th

expected value; $\hat{y}_j$ is the j-th predicted value; f_k is the k-th tree model; $\Omega(\cdot)$ is the complexity of the tree model (i.e., regular term); γ is the complexity cost; T is the number of leaf nodes; λ is the penalty coefficient; ω is the weight of the leaf node.

In addition, an approximate method of the greedy algorithm is used to control the splitting of the tree in the XGBoost model. The approximation algorithm avoids the greedy algorithm to enumerate data, but it is highly suitable for building a tree model with multidimensional data characteristics. This fact also theoretically explains the role of PSR in enhancing the dimension of data in this model.

4.2 Experimental analysis

This experiment uses a computer with an Intel Core i7 processor and 8 GB memory. To verify the feasibility of the PSR-XGBoost method proposed in this paper, five control models were designed, namely, ARIMA, XGBoost, PSR-RF, PSR-SVR, and PSR-LSTM. The parameters of the model are determined by grid search and step-by-step experiments. Fig. 6 shows the fit of the predicted and actual values of the model. The results show that the prediction value of the PSR-XGBoost model fits better than that of other control prediction models.

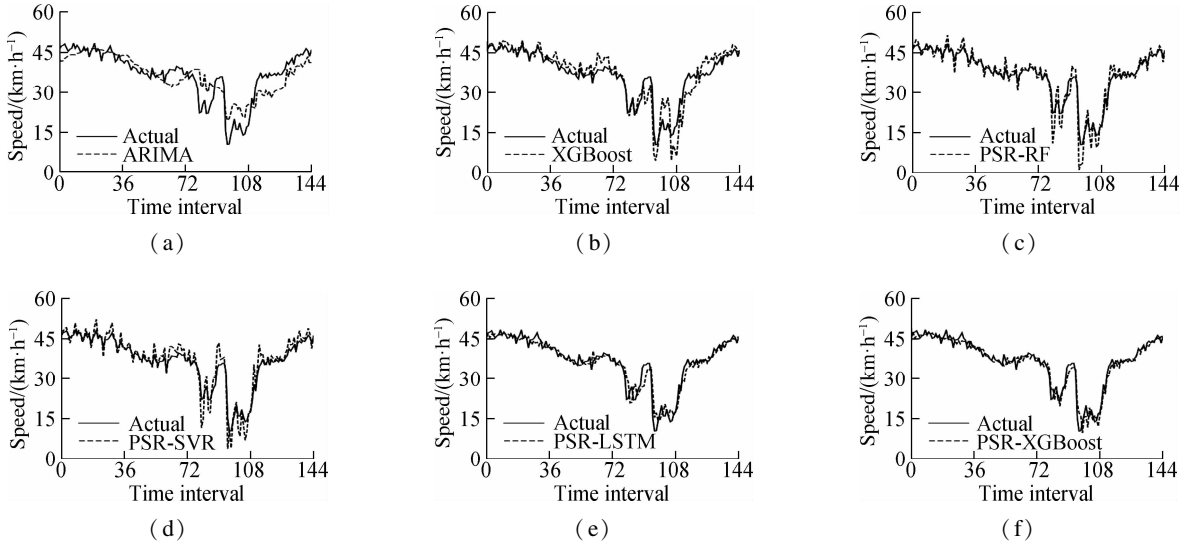


Fig. 6 Model prediction results. (a) Autoregressive integrated moving average (ARIMA); (b) Extreme gradient boosting (XGBoost); (c) PSR-RF; (d) PSR-SVR; (e) PSR-LSTM; (f) PSR-XGBoost

To quantitatively describe the prediction effect of the model, the following evaluation indicators are introduced: mean square error (MSE), average absolute error (MAE), and average absolute percentage error (MAPE). Tab. 1 shows the evaluation index table of the six model prediction results.

Tab. 1 Model prediction performance index table

Categories	Model	MSE	MAE	MAPE/%	Time/s
Experimental model	PSR-XGBoost	5.399	1.632	6.278	17.35
Control model	XGBoost	16.924	3.000	12.005	10.21
	ARIMA	24.004	4.020	14.377	9.58
	PSR-SVR	13.489	2.303	9.173	21.15
	PSR-RF	10.412	1.881	7.670	32.70
	PSR-LSTM	6.436	1.728	6.399	45.83

The experimental results show that the prediction accuracy indicators MSE, MAE, and MAPE of the PSR-XGBoost model are 5.339%, 1.632%, and 6.278%, respectively, and the calculation time is 17.35 s. Compared with the XGBoost model, the PSR-XGBoost model enriches the structure of the tree model by increasing the data dimension and greatly improves the prediction effect of the model. Compared with the mathematical statistics model, the calculation time of the PSR-XGBoost model is longer, but the three accuracy indicators, MSE, MAE, and MAPE, increased by 77.5%, 59.4%, and 56.3%, respectively. Compared with the PSR-SVR, PSR-RF, and PSR-LSTM models, the prediction accuracy of the PSR-XGBoost model improved to a certain extent; the MSE increased by 59.9%, 48.1%, and 16.1%; the MAE increased by 29.1%, 13.2%, and 5.6%; and the MAPE increased by 31.6%, 18.1%, and 1.9%, respectively. Although the prediction accuracy of the PSR-XGBoost and PSR-LSTM models are not much different, the time required for the prediction is significantly reduced.

5 Conclusions

1) The improved data preprocessing method is specific to traffic data. This method can effectively identify the road traffic state through the hierarchical clustering algorithm and fill in missing and abnormal data based on the traffic state.

2) PSR can map one-dimensional data to a multidimensional matrix, effectively solving the problem of low

basic data dimensions, optimizing the model input, enriching the model structure, and improving the prediction accuracy. In addition, complex network methods can analyze and verify the characteristics of reconstructed data.

3) Considering the prediction accuracy and computing efficiency, the PSR-XGBoost model has advantages over mathematical-statistical models and other machine learning algorithms and can be used as an effective method for short-term traffic flow prediction.

References

[1] Han C, Song S, Wang C H. A real-time short-term traffic flow adaptive forecasting method based on ARIMA model[J]. *Journal of System Simulation*, 2004, **16**(7): 1530 - 1532,1535. DOI: 10.3969/j.issn.1004 - 731X.2004.07.042. (in Chinese)

[2] Yang G F, Xu R, Qin M, et al. Short-term traffic volume forecasting based on ARMA and Kalman filter[J]. *Journal of Zhengzhou University* (*Engineering Science*), 2017, **38**(2): 36 - 40. DOI: 10.13705/j.issn.1671 - 6833.2017.02.009. (in Chinese)

[3] Tian Z D. Chaotic characteristic analysis of network traffic time series at different time scales[J]. *Chaos Solitons & Fractals*, 2020, **130**: 109412. DOI: 10.1016/j.chaos.2019.109412.

[4] Jia X C, Chen X M, Gong J L, et al. Multi-step short-term traffic flow prediction based on chaotic theory [J]. *Journal of Transport Information and Safety*, 2013, **31**(6): 27 - 32. DOI: 10.3963/j.issn.1674 - 4861.2013.06.006. (in Chinese)

[5] Wu Q. *Research and application of short-term traffic flow forcasting based on support vector machine regression*[D]. Xi'an: Chang'an University, 2016. (in Chinese)

[6] Zhang L Z, Alharbe N R, Luo G C, et al. A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction[J]. *Tsinghua Science and Technology*, 2018, **23**(4): 479 - 492. DOI: 10.26599/TST.2018.9010045.

[7] Ma X L, Tao Z M, Wang Y H, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data[J]. *Transportation Research Part C: Emerging Technologies*, 2015, **54**: 187 - 197. DOI: 10.1016/j.trc.2015.03.014.

[8] Zhao J D, Gao Y, Yang Z Z, et al. Truck traffic speed prediction under nonrecurrent congestion: Based on optimized deep learning algorithms and GPS data[J]. *IEEE Access*, 2019, **7**: 9116 - 9127. DOI: 10.1109/ACCESS.2018.2890414.

[9] Li Y Y. *Research on the short-term traffic flow forecasting method of based on phase space reconstruction and SVR*[D]. Beijing: Beijing Jiaotong University, 2018. (in Chinese)

[10] Matilla García M, Morales I, Rodríguez J M, et al. Selection of embedding dimension and delay time in phase space reconstruction via symbolic dynamics[J]. *Entropy*, 2021, **23**(2): 221 - 221. DOI: 10.3390/e23020221.

[11] Gao Z K, Jin N D. Complex network from time series based on phase space reconstruction[J]. *Chaos*, 2009, **19**(3): 033137. DOI: 10.1063/1.3227736.

[12] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[C]// *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA, 2016: 785 - 794. DOI: 10.1145/2939672.2939785.

考虑混沌特性的 PSR-XGBoost 短期交通流预测

李树彬[1,2] 孔祥科[1] 李青桐[1] 林兆丰[1] 赵子豪[3]

([1] 山东建筑大学交通工程学院,济南 250101)

([2] 山东警察学院交通管理工程系,济南 250014)

([3] 北京城建设计发展集团股份有限公司,北京 100017)

摘要: 为了提升主动式交通管理的水平,结合相空间重构和 XGBoost 算法提出了一种短时交通流预测模型. 首先,改进了传统的数据预处理方法,通过层次聚类判定交通流状态,并根据不同的交通流状态对缺失、异常数据进行填充. 其次,利用相空间重构将一维数据映射为多维数据矩阵,并利用时间序列复杂网络验证数据重构效果. 最后,将多维数据矩阵输入到 XGBoost 模型以预测未来交通流参数. 结果表明,PSR-XGBoost 模型预测结果的均方误差、平均绝对误差和平均绝对百分数误差分别为 5.399%、1.632% 和 6.278%,所需运行时间为 17.35 s. 相比于数理统计模型和其他机器学习模型,PSR-XGBoost 模型在多项预测指标上均有明显提高,从而验证了其在短时交通流预测方面的可行性和优越性.

关键词: 交通流预测;相空间重构;复杂网络;模型优化

中图分类号: U491.1