Traffic Flow Forecasting Analysis based on Two Methods

Yingxin Liu¹ᵃ, Na Zhang², Xinggang Luo¹ and Min Yang²

¹ College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, China
² Logistic Engineering and management, Shenyang University, Shenyang, Liaoning, China

ᵃemail: yingxin888@126.com

¹ Corresponding author’s e-mail: 327373832@qq.com

Abstract. Along with the rapid growth of China's economy and the continuous improvement of the national living standard, the number of private cars traveling on the road is gradually increasing, and the traffic congestion problem has become the focus of people's attention. In order to solve this big problem, the most important thing is to improve the development of public transportation, conduct scientific and reasonable public transportation scheduling, and realize the green transportation. The prediction of public transportation information is the key point of public transportation scheduling. In this study, the part of public transportation information, traffic flow, is selected for research, and the collected historical information is preprocessed and characterized. Using the weather factor as the input characteristic, the historical traffic flow is divided into 15-minute interval and 1-hour interval data sets, and two models, gradient boosting decision tree and wavelet neural network, are used to predict the traffic flow of road sections. The experimental results show that the gradient boosting decision tree model based on the 15-minute interval has high prediction accuracy and good generalization performance.

1. Introduction

With the rapid development of China's national economy, people's living standards are gradually improving and the environmental requirements for travel are getting higher and higher leading to an increasing number of private cars. According to the reporter learned from the Ministry of Public Security: as of September 2020, the national motor vehicle fleet reached 365 million, an increase of 25 million compared to the same period in 2019. However, there are still some serious problems with road facilities and traffic management in China, which lead to more serious traffic congestion, and it is the realization of scientific and reasonable public transportation scheduling that is the key to solve the traffic congestion problem.

Accurate and efficient prediction of public transportation information not only provides decision makers with the basis for problem solving and timely avoidance of traffic congestion, but also provides valuable reference for passengers to choose a reasonable travel route[1].

The prediction of public transportation information is inseparable from the acquisition of public transportation information. Advanced traffic information collection technology mainly includes GPS-based floating vehicle information collection technology, which is a kind of mobile detector data collection technology that can comprehensively and accurately reflect the whole city traffic condition, and floating vehicle traffic information collection technology can obtain timely and accurate traffic
information all day long by using the advantages of GPS satellite system, and it has the advantages of low installation and arrangement cost, wide coverage, strong information reliability, and Real-time and other advantages. At the same time, it can also collect other traffic information such as the location, instant speed, direction, and time of the floating vehicle in real time[2].

2. Literature Review

In terms of the use of acquisition technology, the use of technology for floating vehicle information acquisition is earlier in foreign countries. Tsuge M, Tokunaga M et al[3] achieved optimization for floating vehicle systems by predicting travel times for time periods, traffic volumes, and weather and other influencing factors on actual travel times, thus achieving the minimum amount of data needed for calculation, in addition to ensuring accuracy. Castro P S, Zhang D and Li S[4] proposed a method to construct a traffic density model using trajectories generated by GPS-equipped cabs, which enables the prediction of traffic conditions while also evaluating the impact of emissions on urban air quality. Wang Z, Lu M, Yuan X et al[5] proposed an interactive system for visualizing and analyzing urban traffic congestion based on GPS trajectories to achieve the detection of traffic congestion events. Anuar K, Cetin M[6] proposed a highway traffic flow estimation method. The estimation of traffic flow in congested and free-flowing conditions was achieved by estimating the flow rate of each state through shock waves. With the popularity of floating vehicles, a lot of research has been done on floating vehicle data applications in China. In 2010, Jiancheng Weng, Wentao Liu, and Zhihong Chen[7] analyzed and preprocessed a large amount of floating vehicle traffic data in order to obtain accurate and timely information on the basic indicators of cab operation characteristics such as operating mileage, operating time, and operating speed, and proposed a cab operation management based on floating vehicle data Indicator calculation model was realized for the analysis of the operation situation. In 2013, Yi Cai[8] used the K-means clustering algorithm to divide the cab GPS data into traffic data and realized the improvement of the boundary division problem. In 2016, Shuai Tang and Qing Li[9] used historical cab GPS data to train the algorithm and implemented a generative learning algorithm to determine traffic congestion conditions.

On the other hand, many current scholars have proposed various traffic flow prediction models and algorithms using the data collected by floating vehicles. Among the linear forecasting methods, the research based on time series models in traffic flow forecasting has been improved and applied based on time series models since it was first proposed by Ahmed Mohamed et al. Xianghai Sun[10] proposed a seasonal ARIMA model (SARIMA ); Kumar[11] optimized the ARIMA model so that the data constraint of the model problem was improved. Among the traditional nonlinear prediction methods, Habtemichael[12] proposed a nonparametric and data-driven nonparametric prediction model for traffic flow based on the identification pattern of traffic flow by reinforcing the K-nearest neighbor algorithm, which was experimentally validated using real data, and the results showed that this model can accurately identify traffic flow for a specific pattern. Yanchong C et al[13] combined wavelet analysis and BP neural network to construct a short-term traffic flow prediction model. Among intelligent nonlinear prediction methods, Zhu J Z[14] developed a traffic volume prediction model based on artificial neural networks considering the interaction of traffic volumes at adjacent intersections. Gui Fu[15] realized the transformation of traffic flow prediction to high dimensionality by using RBF as a kernel function, and established a short-time traffic flow prediction model based on support vector machine regression, using the traffic flow data of Guangzhou city for prediction, which has good prediction accuracy when compared with the results predicted by Kalman filtering model. Among the combined prediction methods, Guan Shuo et al[16] used a combination of K-mean clustering algorithm, least squares method and RBF network to achieve prediction for traffic flow. Xu et al[17] established an ARIMA model for time-series road traffic data based on historical road traffic data, and combined this ARIMA model with Kalman filter to construct a road traffic state prediction algorithm. The optimal parameters of the algorithm are given.
3. Materials and Methods
The historical public transportation information used in this study is traffic flow, and the collection technique used is a GPS-based device for collecting floating vehicle traffic information. The important factor of time interval is added in this study to consider the impact of the differences in the data counted at different time intervals for forecasting.

The main research components are as follows:
- Data pre-processing of public transportation information.
- Characterization and pattern mining of historical public transportation data.
- Prediction of traffic flow of road sections. According to the characterization of traffic flow, the data set of traffic flow is divided into two sample sets with different intervals, and the prediction models are established based on gradient boosting decision tree, and the wavelet neural network prediction model is introduced to compare with the prediction results, and the accuracy of the prediction of the two prediction models is compared.

4. Results & Discussion
From the previous data analysis, it can be seen that there are differences in the time distribution of the traffic flow counted at different intervals. In this chapter, two original traffic flow data sets are counted at two intervals, 15 min and 1 h, and their prediction is analyzed.

4.1. Traffic flow prediction with 15min interval
In this paper, we take the time period of 7:00-12:00 on the road section with linkID 521 for the four weeks of November 2018, and count the traffic flow at 15min intervals. 20 samples per day, so that a total of 400 samples for four weekdays are obtained, and 300 samples for the first three weeks are taken as the training set, and 100 samples for the last week are taken as the test set.

4.1.1. Traffic flow prediction based on GBDT model with 15min interval
GBDT(Gradient Boosting Decision Tree) is a forward-distributed iterative algorithm based on a CART regression tree to initialize a weak learner. Each subsequent step builds a new model in the gradient direction of the residual reduction. The algorithm consists of multiple decision trees, and the results of all tree conclusions added are as the final answer. This paper uses Gradient Boosting Regressor method in scikit-learn to implement GBDT algorithm regression.

The weather factor is used as the input feature, the traffic flow at 15-min intervals is counted, the data of the first three weeks are used as the training set, the data of the last week are used as the test set, the loss function is the mean squared difference function ('ls'), The same grid search method is used to determine the optimal parameters. The final optimal parameters are n-estimators of 500, learning rate of 0.2, max_depth of 3, and loss function of mean squared error ('ls'). The GBDT model is trained with the training data and saved, and the test data $x_i = (w(i), p_{i-7}(t+1), p_{i-14}(t+1), p_{i-21}(t+1), p_{i-1}(t+1), p_{i-2}(t+1), p_{i-3}(t+1), p_i(t), p_i(t-1), p_i(t-2))$, are input to the trained road traffic flow model, and the prediction results are shown in Figure 1.

Figure 1. The comparison between traffic flow predicted by GBDT and actual traffic flow at 15-minute intervals

4.1.1. Traffic flow prediction based on GBDT model with 15min interval
GBDT(Gradient Boosting Decision Tree) is a forward-distributed iterative algorithm based on a CART regression tree to initialize a weak learner. Each subsequent step builds a new model in the gradient direction of the residual reduction. The algorithm consists of multiple decision trees, and the results of all tree conclusions added are as the final answer. This paper uses Gradient Boosting Regressor method in scikit-learn to implement GBDT algorithm regression.

The weather factor is used as the input feature, the traffic flow at 15-min intervals is counted, the data of the first three weeks are used as the training set, the data of the last week are used as the test set, the loss function is the mean squared difference function ('ls'), The same grid search method is used to determine the optimal parameters. The final optimal parameters are n-estimators of 500, learning rate of 0.2, max_depth of 3, and loss function of mean squared error ('ls'). The GBDT model is trained with the training data and saved, and the test data $x_i = (w(i), p_{i-7}(t+1), p_{i-14}(t+1), p_{i-21}(t+1), p_{i-1}(t+1), p_{i-2}(t+1), p_{i-3}(t+1), p_i(t), p_i(t-1), p_i(t-2))$, are input to the trained road traffic flow model, and the prediction results are shown in Figure 1.

Figure 2. Traffic flow error predicted by GBDT at 15 minute intervals
In Figure 1, the blue line with blue circles represents the trend of actual traffic flow and the red line with red triangles represents the trend of GBDT predicted traffic flow. Take a look at the fluctuation range of the error through the error plot, as shown in Figure 2. From Figure 2, the error of GBDT predicted traffic flow fluctuates in the interval [-34,35].

4.1.2. Traffic flow prediction based on wavelet neural network model with 15-min interval
The number of nodes in the input layer is determined by the number of features in the input, and the number of neurons in the input layer is set to 13. The number of neurons in the hidden layer is selected according to the number of neurons in the output layer is set to 1. On this basis, the wavelet neural network prediction model of traffic flow with 15-min interval statistics is constructed. The results of the wavelet neural network prediction traffic flow with 15min interval are shown in Figure 3.

From Figure 3 and Figure 1, it can be seen that the local fit of the wavelet neural network prediction traffic flow results is not as good as that of the GBDT prediction traffic flow, and the change of the error is observed through the error plot, as shown in Figure 4. From Figure 4, it can be seen that the error of predicting traffic flow by wavelet neural network fluctuates in the interval [-78,45]. By comparing with Figure 2, it can be seen that the error of predicting traffic flow by wavelet neural network fluctuates in a larger range than the error of predicting traffic flow by GBDT.

4.1.3. Comparison of the two models
Two evaluation indicators, mean square error (MSE) and mean absolute error (MAE), were also selected to compare the prediction results of the models, as shown in Table 1. The smaller the values of these two indicators, the higher the prediction accuracy of the models.

| Evaluation Index    | MSE   | MAE   |
|---------------------|-------|-------|
| GBDT                | 319.38| 15.92 |
| Wavelet Neural Network | 389.16| 24.16 |

From the table 1, we can see that the GBDT-based traffic flow prediction model has a great improvement in prediction accuracy compared with the wavelet neural network traffic flow prediction model with 15min interval, and the mean square error and the average absolute error are reduced by 69.78 and 8.24, respectively, which shows that the GBDT-based traffic flow prediction model has excellent performance in predicting traffic flow at 15min interval. It can be seen that the GBDT traffic prediction model has excellent performance in predicting the traffic flow of statistical road sections at 15-min intervals, and has high prediction accuracy, stability and good applicability.

4.2. Traffic flow prediction with 1h interval

4.2.1. Traffic flow prediction based on GBDT model with 1h interval
In the traffic flow prediction with 1h interval GBDT model, the weather factor is also used as a feature input, the traffic flow at 1 hour interval is counted, the data of the first three weeks are used as the
training set and the data of the last week are used as the test set, the loss function is the mean squared difference function ('ls'), and the grid search method is used to determine. The final optimal parameters are n-estimators of 500, learning rate of 0.2, max_depth of 2, and the loss function of mean squared error ('ls'), and the training data samples are used to train the GBDT model and saved; the test data 
\[ x_i = (w(i), p_{i-7}(t + 1), p_{i-14}(t + 1), p_{i-21}(t + 1), p_{i-1}(t + 1), p_{i-2}(t + 1), 
\]
\[ p_{i-3}(t + 1), p_i(t), p_{i}(t - 1), p_{i}(t - 2) ) \]
are input to the trained GBDT traffic flow prediction model, and the prediction results are shown in Figure 5.

Figure 5. The comparison between traffic flow predicted by GBDT and actual traffic flow at 1 hour intervals

In Figure 5, the blue circles and lines represent the trend of actual traffic flow, and the red triangles and lines represent the trend of GBDT predicted traffic flow, and observe the fluctuation range of the error through the error map, as shown in Figure 6. From Figure 6, it can be seen that the error of GBDT prediction of traffic flow with 1h interval is more fluctuating in the interval [-102,87] for local samples, but the error fluctuation is more stable for most of the samples.

4.2.2. Traffic flow prediction based on wavelet neural network model with 1h interval

In the traffic flow prediction model based on wavelet neural network with 1h as time interval, a three-layer wavelet neural network is used for traffic flow prediction, and the input data is the same as that of 15min interval, and the same number of neurons in the input layer is 13, the number of neurons in the hidden layer is 14, and the number of neurons in the output layer is 1. On the basis of this, the traffic flow wavelet neural network prediction based on 1h as time interval The prediction results are similar to the actual traffic flow. The comparison between the prediction results and the actual traffic flow is shown in Figure 7.

Figure 7. The comparison between traffic flow predicted by wavelet neural network and actual traffic flow at 1 hour intervals

From the comparison between Figure 5 and Figure 7, it can be seen that the wavelet neural network predicts the traffic flow in the local fit is not as good as the fit of GBDT. The variation of the error is visualized by the error plot, as shown in Figure 8. As shown in Figure 8, the fluctuation range of the wavelet neural network prediction traffic flow error with 1h interval is in the interval [-107,99], which is large compared with Figure 6, and the overall fluctuation range is also larger than that of GBDT prediction traffic flow.
4.2.3. Comparison of the two traffic flow forecasting models

The same mean square error and mean absolute error were used to evaluate the models, and the evaluation indexes of the two models are shown in Table 2.

| Evaluation Index      | MSE   | MAE   |
|-----------------------|-------|-------|
| GBDT                  | 577.69| 37.76 |
| Wavelet Neural Network| 598.53| 53.44 |

From Table 2, it can be seen that the GBDT-based traffic flow prediction model has greatly improved in prediction accuracy and fit compared with the wavelet neural network traffic flow prediction model, and the mean square error and mean absolute error are reduced by 20.84 and 15.68, respectively, which shows that the GBDT-based traffic flow prediction model has excellent performance in the problem of roadway traffic flow prediction. It has the characteristics of high prediction accuracy, good stability and wide applicability.

4.3. Comparison of traffic flow prediction results for two interval methods

In sections 4.1. and 4.2., for better comparison, the traffic flow data are split into two data sets with 15min and 1h time interval statistics respectively, and then two models of GBDT and wavelet neural network are used for traffic flow prediction respectively, and the evaluation indexes of prediction results are shown in Table 3.

| Model              | Different spacing methods | MSE   | MAE   |
|--------------------|----------------------------|-------|-------|
| GBDT               | 15 min                     | 319.38| 15.92 |
|                    | 1 h                        | 577.69| 37.76 |
| Wavelet Neural Network | 15 min                | 389.16| 24.16 |
|                    | 1 h                        | 598.53| 53.44 |

As can be seen from Table 3, the MSE and MAE of the same model with different intervals can visually show that the model with 15min interval predicts the traffic flow of the period with higher accuracy than the model with 1h interval. It can be seen that the shorter the interval of statistical traffic flow, the better the prediction effect of the model. In addition, the prediction results of GBDT predicting time slot traffic flow in the same time interval have higher accuracy and better generalization performance than wavelet neural network prediction.

5. Conclusions

In this study, the traffic flow data were counted in 15min and 1h intervals, and the data sets of these two intervals were input into GBDT model and wavelet neural network model respectively. The GBDT model is suitable for traffic flow prediction.

Acknowledgments

This research was financially supported by the National Science Foundation of China (NSFC Proj. 71601126, 71771070 and 71571037), and Liaoning Province Natural Fund Guidance Plan (Proj. 20180550423).

References

[1] Yin, T.T. (2015) Research on Bus Scheduling Rules Based on Big Data. https://kns.cnki.net/KCMS/detail/detail.
[2] Zhang, J.H. (2013) Research on the method of obtaining real-time road speed based on GPS floating vehicle. https://kns.cnki.net/KCMS/detail.
[3] Tsuge, M., Tokunaga, M., Nakano, K. (2010) On estimation of link travel time in floating car systems. International Journal of Intelligent Transportation Systems Research, 8(3):175-187.
[4] Castro, P.S., Zhang, D., Li, S. (2012) Urban traffic modelling and prediction using large scale taxi GPS traces. International Conference on Pervasive Computing, Springer-Verlag.

[5] Wang, Z., Lu, M., Yuan, X. (2013) Visual traffic jam analysis based on trajectory data. IEEE Transactions on Visualization & Computer Graphics, 19(12):2159-2168.

[6] Anuar, K., Cetin, M. (2017) Estimating freeway traffic volume using shockwaves and probe vehicle trajectory data. Transportation Research Procedia, 22:183-192.

[7] Weng, J.C., Liu, W.T., Chen, Z.H. (2010) Research on cab operation management based on floating vehicle data. Journal of Beijing University of Technology, 36(6):779-784.

[8] Cai, Y. (2013) The division of traffic cells in Nanjing based on cab GPS data. Science and Technology Innovation Herald, 2013(23):218-218.

[9] Tang, S., Li, Q. (2016) Real-time analysis of urban traffic conditions based on cab GPS big data. Business, 2016(30):219-219.

[10] Sun, X.H., Liu, T.Q. (2008) Research on short-term traffic flow prediction of urban roads based on nonlinear time series model. Journal of Civil Engineering, 2008(1):104-109.

[11] Kumar, S.V., Vanajakshi, L. (2015) Short-term traffic flow prediction using seasonal ARIIVIA model with limited input data. European Transport Research Review, 7(3):21.

[12] Habtemichael, F.G., Cetin, M. (2016) Short-term traffic flow rate forecasting based on identifying similar traffic patterns. Transportation Research Part C, 66:61-78.

[13] Chai, Y.C., Huang, D.R., Zhao, L. (2016) A short-term traffic flow prediction method based on wavelet analysis and neural network. Control & Decision Conference. IEEE.

[14] Zhu, J.Z., Cao, J.X., Zhu, Y. (2014) Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections. Transportation Research Part C, 47: 139-154.

[15] Fu, G., Han, G.Q., Tu, F. (2013) Short-time traffic flow prediction model based on support vector machine regression. Journal of South China University of Technology, 41(9): 71-76.

[16] Guan, S., Gao, J.W., Zhang, B. (2014) Traffic flow prediction based on K-mean clustering algorithm RBF neural network. Journal of Qingdao University (Engineering Technology Edition), 29(2):20-23.

[17] Xu, D.W., Wang, Y.D., Li, M.J., et al. (2017) Real-time road traffic state prediction based on ARIMA and Kalman filter. Frontiers of Information Technology & Electronic Engineering, 18(2):287-302.