Research on Short-Term Urban Traffic Congestion Based on Fuzzy Comprehensive Evaluation and Machine Learning

Yuan Mei[1(✉)], Ting Hu[2(✉)], and Li Chun Yang[1]

[1] Department of Information and Software Engineering, Chengdu Neusoft University, Dujianyan, Chengdu 611844, Sichuan, China
2551161628@qq.com, 1377759045@qq.com

[2] Department of Information Management, Chengdu Neusoft University, Dujianyan, Chengdu 611844, Sichuan, China
1187352043@qq.com

Abstract. There are many factors that affect urban traffic flow. In the case of severe traffic congestion, the vehicle speed is very slow, which results in the GPS positioning system's estimation of the vehicle speed being very inaccurate, which in turn leads to poor reliability of the estimated congestion time of the road segment. The main contents of this study are: in the case of urban traffic congestion, the prediction and analysis of the degree of traffic congestion and the length of congestion. Taking the dynamic traffic data of Shenzhen on June 9, 2014 as an example, the road section of Binhe Avenue is selected, and the data of traffic flow, average speed of traffic volume and traffic volume density in the current time period are calculated after data preprocessing, as a measure of traffic. The main impact indicators of congestion status. Then we use the fuzzy comprehensive evaluation method to divide TSI as a traffic congestion evaluation index and divide the road congestion into four levels. In this way, we can get the congestion of the road in each time period of the day and the time required to pass. Then we use the random forest, adaboost, GBDT, Lasso CV and BP neural networks in the machine learning algorithm to build models to measure traffic congestion for training and testing. Finally, the BP neural network has the best effect on this problem, and mean square error is 0.0190. Finally, we used BP neural network to predict and congest the road in the next three hours. From the experimental simulation results, this method can effectively analyze and predict the real-time traffic congestion.

Keywords: Random forest · Adaboost · GBDT · Lasso CV · BP neural network · Fuzzy comprehensive evaluation

1 Introduction

At present, there are many solutions for urban road congestion, such as the use of single and double day limit traffic. But the daily congestion on some roads is still not optimistic. There have been many studies on congestion, such as Fu Gui's short-term traffic flow

Y. Tan et al. (Eds.): DMBD 2020, CCIS 1234, pp. 95–107, 2020.
https://doi.org/10.1007/978-981-15-7205-0_9

prediction model based on support vector machine regression [1], Dewey's short-term traffic prediction based on the combination of K nearest neighbor algorithm and support vector regression [2] but its prediction There is a certain deviation between the result and the true value. Kang Danqing's research on short-term traffic flow prediction methods based on deep learning [3] has better results but is more complicated. To this end, this paper compares the effects of various machine learning algorithms on this problem, and proposes that the back propagation based BP neural network has a better effect on the problem of short-term traffic flow prediction and is simpler than deep learning. The specific research content is as follows.

In this paper, the GPS data was collected in real time on a taxi on the Shenzhen Stock Exchange on June 19, 2014. After preprocessing by deleting invalid data, erroneous data, filling missing data, etc. The traffic flow, average speed of traffic volume and traffic volume density are extracted as the measurement indicators of traffic congestion, and the main influencing factors are determined by calculating the correlation between the indicators. Then through the use of various machine learning algorithms (random forest, adaboost, GBDT, LassoCV, BP neural network), through grid search cross-validation, the best parameters of each algorithm and accuracy under the best parameters are obtained. Comparing the effects of various models, it is obtained that the BP neural network has the highest effect in the current situation, and its mean squared error is 0.0190.

Based on the BP neural network, the BinHe avenue was selected, and a model based on fuzzy comprehensive evaluation to measure the time required for vehicles to pass through the congested road section was established. The congestion situation is divided into four levels: very smooth, unobstructed, congested, and severely congested. The congestion situation on the road section was analyzed on that day. Finally, the TSI evaluation index is used to analyze the effect of the model.

2 Research Methods

2.1 Random Forest Algorithm of CART Tree

CART can be used for regression analysis and classification analysis, and expanded some integrated algorithms based on CART. In order to solve the problem of large data volume and large data volume in a big data environment, this study chose CART as the basic random forest algorithm. The CART decision tree has the advantages of being easy to understand and has some nonlinear classification capabilities, but the single decision tree has some disadvantages. In integrated learning, the above defects can be improved through the random forest integration method. Random forest is composed of many decision trees, and there is no correlation between different decision trees, and the model generalization ability is stronger. The algorithm is mainly to randomly sample the samples, train the decision tree, and then classify the nodes according to the corresponding attributes until they are no longer split, and finally build a large number of decision trees to form a forest [4].

2.2 Adaboost Algorithm

Adaptive Boosting is a boosting method that combines multiple weak classifiers into a strong classifier. Its self-adaptation lies in: the weight of the sample that was divided by

the previous weak classifier (the weight corresponding to the sample) will be strengthened, and the sample with the updated weight will be used again to train the next new weak classifier. In each round of training, a new weak classifier is trained with the population (sample population) to generate new sample weights and the speech weight of the weak classifier, and iterates until it reaches a predetermined error rate or the specified maximum number of iterations.

2.3 GBDT Algorithm

Gradient Boosting Decision Tree is a combination of GB (Gradient boosting) and DT (Decision Tree), that is, when a single learner in GB is a decision tree. Decision trees are divided into two categories, regression trees and classification trees. The former is used to predict real values, and the latter is used to classify label values. Through multiple iterations, each iteration generates a weak classifier, and each classifier is trained on the basis of the residuals of the previous classifier. The requirements for weak classifiers are generally simple enough and have low variance and high deviation. Because the training process is to continuously improve the accuracy of the final classifier by reducing the deviation.

2.4 LassoCV Algorithm

Least absolute shrinkage and selection operator is a linear model used to estimate sparse parameters, especially suitable for reducing the number of parameters. For this reason, the Lasso regression model is widely used in compressed sensing. Mathematically, Lasso adds an L1 regular term to the linear model. No matter whether the dependent variable is continuous or discrete, lasso can handle it. In general, lasso has extremely low data requirements, so it is widely used; in addition, lasso can also filter and The complexity of the model is reduced. Variable selection here refers to not putting all variables into the model for fitting, but selectively putting variables into the model to get better performance parameters.

2.5 Neural Network Algorithm

BP neural network is an algorithm that transforms a set of sample input and output problems into nonlinear optimization [5]. It has a three-layer structure and interconnects neurons (Fig. 1).

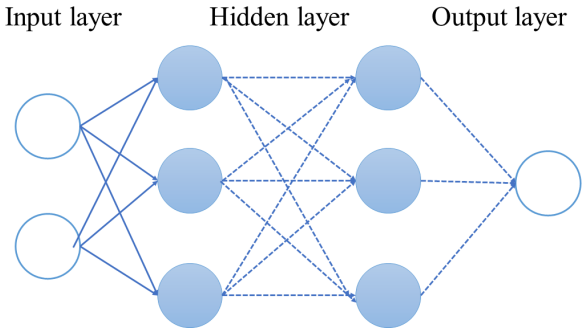

Fig. 1. Neural network structure diagram

The neurons from left to right are the input layer, hidden layer and output layer. The weight of the input layer to the hidden layer (the connection weight of neuron i and neuron j), the threshold; the size of the hidden layer to the output layer may be (the connection weight of neuron i and neuron k).

Usually, two nodes are input, the second layer is hidden, and the output of the three nodes in each layer is one node. The algorithm flow is as follows.

Algorithm: BP neural network algorithm

Seeding: Initialize the weight and bias of network.
repeat
  foreach Each training tuple X in D **do**
    // Forward propagation input
    foreach Each input unit j do
      $O_j = I_j$;// The output of the input unit is its actual input value
    end
    foreach hide or export each cell of the layer j do
      $I_j = \sum_i W_{ij} O_i + \theta_j$; //Regarding the previous layer, the net input of calculation unit j// $O_j = \frac{1}{1+e_j^{-1}}$;// Output of calculation unit j
    end
    // Back propagation error
    foreach each unit of the output layer j do
      $Err_j = O_j(1 - O_j) \sum_k Err_k W_{jk}$;//Calculate the error about the next higher layer k
    end
    foreach every right in the network W_{ij}
      $\Delta W_{ij} = (l) Err_j O_i$;// Weight increment
      $W_{ij} = W_{ij} + \Delta W_{ij}$; // Weight update
    end
    foreach θ_{ij} **each bias in the network**
      $\Delta\theta_i = (l) Err_j$;//Bias increment
      $\theta_i = \theta_j + \Delta\theta_j$;// Bias update
    end
  end
until termination condition
return forecast result

3 Establishing Model

3.1 Selection of Indicators

For the prediction and evaluation of traffic congestion status, we must first select the factor indicators that can accurately and effectively characterize the traffic congestion status. The principle of selection is to have overall completeness, objectivity, operability, and comparability. However, we cannot just pass a certain traffic flow. The parameters evaluate the traffic congestion. The average speed of a vehicle can intuitively represent the state of traffic congestion, but when the vehicle is waiting for a red light at an intersection, although the speed is very small, it does not mean that the road is congested at the moment. Therefore, this paper selects three traffic flow parameters, traffic flow, average speed of traffic flow, and density of traffic flow as factor indicators.

3.2 Calculate Relevance

The average speed of traffic flow refers to the average distance traveled by all vehicles on a road per unit time. This indicator can intuitively reflect the current road traffic congestion. Generally speaking, the greater the speed, the smoother the road; the lower the speed, the more congested the road [6]. Calculated as follows:

$$\overline{v_i} = \frac{1}{N}\sum_{i=1}^{N} v_I \tag{1}$$

$\overline{v_i}$——Average speed of traffic flow (Km/h); N——The number of all vehicles on the road in unit time; v_i——The number of all vehicles on the road in unit time.

Traffic flow density refers to the total number of vehicles on a road per unit length in a unit of time. When the road is congested, the vehicle stalls and the change in traffic flow is almost zero, but the traffic density is very large, so it is decisive for the traffic congestion state. effect. The calculation method is as follows:

$$D = \frac{f}{v} \tag{2}$$

D is the required traffic flow density (vehicles/km), and f is the monitored traffic flow every five minutes; v——Average speed.

Calculate the Pearson correlation coefficient between features. The correlation is a non-deterministic relationship, and the correlation coefficient is the amount of linear correlation between the variables studied.

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \tag{3}$$

Among them, Cov (X, Y) is the covariance of X and Y, Var [X] is the variance of X, Var [Y] is the variance of Y.

3.3 Fuzzy Comprehensive Evaluation

Through analysis and selection, and taking traffic flow and average traffic speed as the basic indicators of traffic state characteristics, a weighted calculation model is applied to the traffic state assessment of road segments and road networks, and a road and road network traffic state characterization model is established.

Calculate the degree of composition of each index state, and then obtain the single-factor fuzzy discrimination matrix of the entire system, as shown below.

$$R = \begin{bmatrix} \mu_1^1 & \mu_1^2 & \mu_1^3 & \mu_1^4 \\ \mu_2^1 & \mu_2^2 & \mu_2^3 & \mu_2^4 \\ \mu_3^1 & \mu_3^2 & \mu_3^3 & \mu_3^4 \end{bmatrix} \tag{4}$$

Among them, the μ_1^1 strongly transitive fuzzy matrix reflects the membership of each traffic state level of each evaluation index pair corresponding to the calculated value of the membership function.

Determining weights Describe the importance of each indicator according to the traffic situation. It is necessary to determine the weight of the basic indicators. Each indicator weight constitutes a weight set, namely:

$$\omega = (\omega_1, \omega_2, \omega_3) \tag{5}$$

Establish an evaluation system-construct a judgment matrix for comparison According to the 9-scale method, the importance between n elements of the same layer can be obtained, thereby establishing a judgment matrix.

$$I = \begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{m1} & \cdots & W_{mn} \end{bmatrix} \tag{6}$$

The elements of the judgment matrix are normalized. The formula is as follows:

$$\alpha_{ij} = \frac{\alpha_{ij}}{\sum_{i=1}^{n} \alpha_{ij}} \tag{7}$$

n——Index factor number of each column judgment matrix

The judgment matrix generated after normalization is added row by row:

$$w_i = \sum_{i=1}^{n} a_{ij} \tag{8}$$

In——The number of index factors for each judgment matrix

Based on the judgment matrix obtained above, normalization is performed to obtain a maximum feature vector, which is the weight of each factor. $w = [w_1, w_2, w_3 \ldots, w_n]$. For a multi-level evaluation system, the weights of the upper layer indicators of each factor indicator can be determined from top to bottom, and finally the weight of each layer factor relative to the target layer is obtained.

Fuzzy comprehensive evaluation, after determining the single-factor discriminant matrix of evaluation index weights, comprehensive evaluation can be conducted through fuzzy transformation.

$$B = (w_1, w_2, w_3)^{\circ} \begin{bmatrix} \mu_1^1 & \mu_1^2 & \mu_1^3 & \mu_1^4 \\ \mu_2^1 & \mu_2^2 & \mu_2^3 & \mu_2^4 \\ \mu_3^1 & \mu_3^2 & \mu_3^3 & \mu_3^4 \end{bmatrix} = (b_1, b_2, b_3, b_4) \tag{9}$$

$$b_j = \sum_{i=1}^{3} \omega_i \mu_i^j, j = 1, 2, 3, 4 \tag{10}$$

b_j——For the final fuzzy comprehensive evaluation index, comprehensively consider the membership degree of all basic indexes in the evaluation set to the jth element. According to the principle of maximum membership, the evaluation element with the largest membership is selected as the evaluation result.

3.4 Classification of Congestion Grade Based on Fuzzy Comprehensive Evaluation

According to the calculation method of traffic congestion in Shenzhen issued by the state in 2011. It quantifies the traffic condition based on the average speed of the statistical interval. The value range is 0–100. Eventually, four traffic congestion states of Smoother, Smooth, Crowded and Blockage. The basic calculation model and the weighted calculation model are suitable for the traffic state assessment of road segments and road networks, respectively. The specific method is as follows:

The key parameter of the TSI calculation model is the formation speed, which is used to evaluate the traffic state of the road section. The formula is as follows:

$$TSI = \frac{V_f - V_i}{V_f} \times 100 \tag{11}$$

V_f——Study actual road speed

V_i——Study free flow speed

Quantify traffic conditions based on the average speed of the statistical interval. The value range is 0–100, which divides the traffic operation status into very stable, stable, congested and congested. The table shows the correspondence between TSI and road network traffic (Table 1).

Table 1. Division of TSI

Traffic status level	Smoother	Smooth	Crowded	Blockage
TSI	[0, 30)	[30, 50)	[50, 70)	[70, 100]

The basic calculation model is relatively simple, and the results can be calculated from the average stroke speed and free flow speed of the link. However, in practical

applications, the evaluation of the traffic congestion status of road sections is not enough, and the traffic status of the road network needs to be carried out. Therefore, considering the influence of the number of road segments and miles on the degree of congestion, the formula for the weighted calculation model is as follows.

$$TSI = \frac{\sum_{i=1}^{l} K_i l_i \left[\frac{V_f - V_i}{V_f}\right]}{\sum_{i=1}^{l} K_i} \times 100 \tag{12}$$

K_i——Lane number i

l_i—Section i Road length, *km*.

l——Total number of road segments

After obtaining the road network TSI, convert it to a traffic status level according to the table. According to the comparison of TSI road travel time, Shenzhen looks at the traffic status of the road network.

3.5 Comparison of Various Algorithm Effects

Through the selected traffic congestion measurement indicators, and then use random forest, adaboost, GBDT, LassoCV, BP neural network algorithms to train and test the data, and compare the effectiveness of various algorithms in dealing with the problem.

3.6 BP Neural Network Regression Prediction

Through the "training" to obtain this input, the appropriate nonlinear relationship between the output. The "training" process can be divided into two stages of forward transmission and backward transmission:

Forward transmission phase:
Take a sample from the sample set P_i, Q_i; Enter P_i into the network; Calculate the error measure E_i and the actual output $O_i = F_L\left(\ldots\left(F_2\left(F_1\left(P_i W^{(1)}\right)W^{(2)}\right)\right)W^{(L)}\right)$; Make adjustments to the weight values $W^{(1)}$ ….$W^{(2)}$ … W^L, and repeat this cycle until $\sum E_i < \varepsilon$.
Backward propagation stage-error propagation stage:
Calculate the difference between the actual output O_P and the ideal output Q_i; Adjust the weight matrix of the output layer with the error of the output layer.

$$E_i = \frac{1}{2}\sum_{j=1}^{m}\left(Q_{ij} - O_{ij}\right)^2 \tag{13}$$

Use this error to estimate the error of the direct leader layer of the output layer, and then use the error estimation of the output layer leader layer. Change the error of the previous layer. In this way, the error estimates of all other layers are obtained. And use these estimates to modify the weight matrix. Form a process of gradually transmitting the error displayed at the output to the output in the opposite direction to the output signal.

The error measure of the network about the entire sample set:

$$E = \sum E_I \tag{14}$$

4 Experiment

4.1 Index Correlation Coefficient

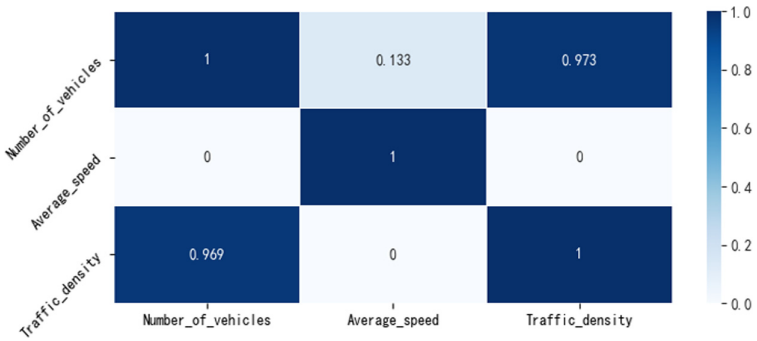

Fig. 2. Correlation coefficients of various indicators

As shown in Fig. 2, the c or relationship coefficient of the number of vehicles, average speed and traffic density extracted according to data processing. It is known from the correlation that although the number of taxis is extracted, there is no difference between the average speed of vehicles on the same road segment and the average speed of non-rented vehicles, so the calculated traffic after observing the extracted data The trend of density change is still representative on the road section.

4.2 Analysis of Congestion in Fuzzy Comprehensive Evaluation

The following figure is obtained by analyzing the relationship between traffic flow and traffic flow density.

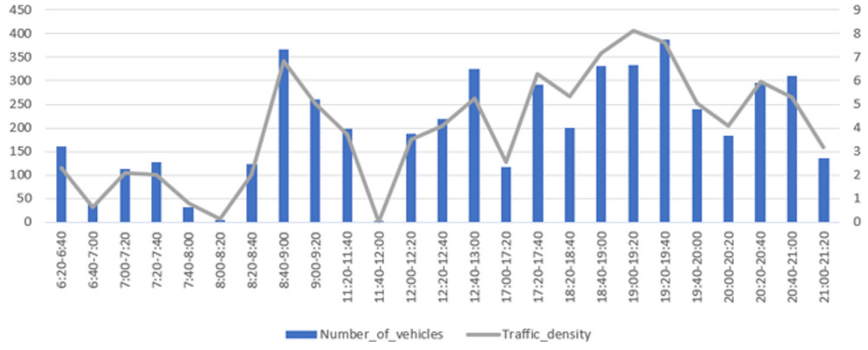

Fig. 3. Relationship between traffic flow and traffic flow density

In the relationship between the number of vehicles and the traffic density in Fig. 3, it can be found that the early, middle, and late high stages are generally between 8:00–9:00, 12:00–1:00, and 19:00–20:00 (Table 2).

Table 2. Congestion level and time prediction during congestion period

Period of time	Number of vehicles	Congestion level	Expected congestion time (min)
8:00–8:20	5	Soomther	5–10
8:20–8:40	124	Crowded	10–20
8:40–9:00	367	Blockage	20–30
12:00–12:20	188	Blockage	20–35
12:20–12:40	218	Blockage	35–50
12:40–13:00	324	Blockage	30–40
19:00–19:20	334	Blockage	35–50
19:20–19:40	184	Blockage	35–55
19:40–20:00	136	Blockage	25–35

Through fuzzy comprehensive evaluation, the degree of congestion is divided into 4 levels, and the analysis of the congestion in each time period is shown in the figure above. It was found that the congestion situation was the most serious in the morning, middle and evening peak congestion time within a day at 8:40–9:00, 12:40–1:00, 19:20–19:40, and the number of vehicles coming out of GPS vehicles was 367, 324, 388, providing conditions for future short-term traffic forecast.

4.3 Comparison of Various Algorithms

We used Anaconda3's Jupiter-notebook software and called the linear model. Lasso CV algorithm module corresponding to the sklearn library and the Random Forest Regressor, Ada Boost Regressor, Gradient Boosting Regressor algorithm modules corresponding to the sklearn. ensemble library, and the Grid Search CV model regulator corresponding to sklearn. model selection Then each algorithm was trained and adjusted, and the relationship between the obtained test set and the real value is shown in Fig. 4.

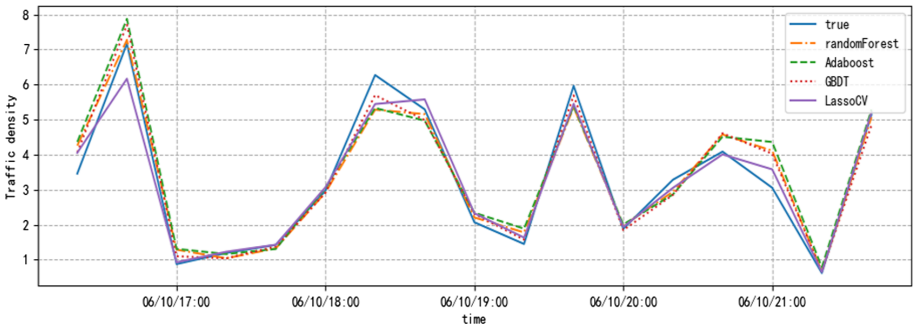

Fig. 4. Comparison of accuracy of various algorithms

As can be seen from Fig. 4, various algorithms have a good effect on traffic congestion analysis, but there are still deviations in certain data. Therefore, we use Anaconda's jupyter notebook software, call the tensorflow module to implement the BP neural network model, and train and test the data. The relationship between the obtained test results and the true value is shown in Fig. 5.

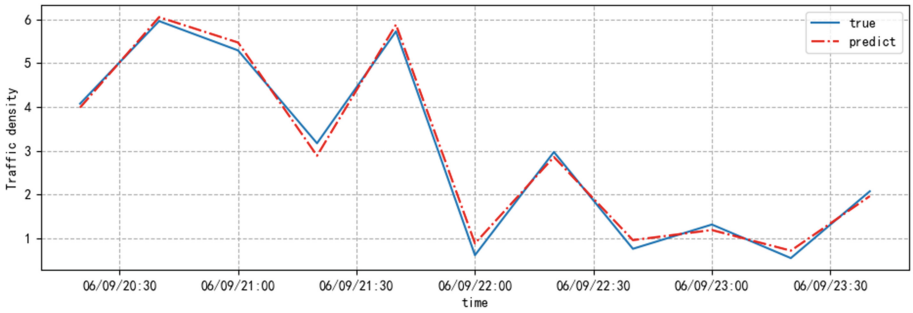

Fig. 5. BP neural network accuracy chart

It can be seen from the prediction results of the BP neural network that the algorithm has a significant deviation in processing a small part of the data in processing this problem, and the overall effect is better. The accuracy of its various algorithms and the optimal parameters for tuning are as follows.

Table 3. Comparison table of accuracy of multiple models

Model	Mean squared error	Hyperparameter
Random forest regression	0.1195	'n_estimators':920
Adaboost regression	0.2127	'n_estimators':1750
GBDT regression	0.1654	'n_estimators':570
Lasso regression	0.1309	Kernel='linear'
BP neural network	0.0190	Number of hidden layer nodes:5

It can be seen from Table 3 that various algorithms have better effects on this problem after training and adjustment. Among them, the BP neural network has the best effect, and the mean squared error has reached 0.0190.

4.4 BP Neural Network Prediction

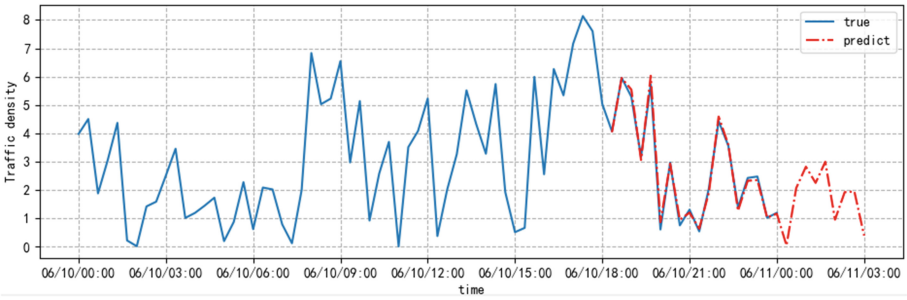

Fig. 6. BP neural network predicts the future time period diagram

By using the data from June 9 to train and test the BP neural network, the mean square error (MSE) of the model is 0.01905, which is relatively reliable. Then predict the congestion of the road within three hours. It can be seen that during the time period from 00:00 to 03:00 on June 10, the road congestion will maintain a small fluctuation trend, but the overall amplitude does not change significantly, that is, the road will not appear serious in the short term Congestion. In addition, you can use the above method-fuzzy comprehensive evaluation in this article to process the prediction results, that is, you can get the congestion situation and congestion time in this period (Fig. 6).

5 Conclusion

It can be seen from the above analysis that the various algorithms after parameter optimization are consistent with the short-term traffic prediction results and recorded values, and the results obtained by these algorithms are all valid. However, the prediction result of BP neural network is more accurate than other algorithms, which shows that BP neural network has stronger data expression ability and can better simulate short-term traffic conditions. Within the acceptable error range, the prediction results can provide effective decision-making information for emergency rescue of urban center road accidents. In the future 5G mobile network, cars with GPS positioning system can accurately analyze the current location of the vehicle and the traffic flow of the next section and the surrounding sections that will pass. Immediately before entering the traffic jam section, the most accurate algorithm is used to recommend a new relatively smooth section for the vehicle. On this basis, the number of vehicles in each section is regulated in a balanced manner to prevent the congestion from further deteriorating, and at the same time provide a guarantee for citizens to travel smoothly and efficiently, laying a foundation for the development of intelligent traffic control systems. Although with the increase of traffic congestion, the prediction results of the model may be deviated to a certain extent, and various traffic parameters will change with time, but it can ensure that the driver occurs in a traffic accident or bad weather. The accident allows the command center to give a rescue plan based on the traffic situation that occurred within a short period of time and arrive at the scene as soon as possible.

References

1. Danqing, K.: Research on short-term traffic flow prediction method based on deep learning. Harbin University of Science and Technology, Heilongjiang, pp. 7714–2015 (2018)
2. Liu, Z., Du, Y., Yan, D., et al.: Short-term traffic flow prediction based on the combination of K-nearest neighbor algorithm and support vector regression. Highway Traff. Sci. Technol. **34**(5), 122–128 (2017)
3. Fu, G., Qiang, K., Lu, F., et al.: Short-term traffic flow prediction model based on support vector machine regression. J. South China Univ. Technol. (Nat. Sci. Ed.) **41**(9), 71–76 (2013)
4. Yan, Y., Bai, L, Wu, Q., et al.: Traffic congestion prediction and evaluation based on multi-index fuzzy comprehensive evaluation. Comput. Appl. Res. **36**(12), 3697–3700, 3704 (2019)
5. Zhu, B., Fu, Z., Yang, S., et al.: Forecasting model of traffic accident spatio-temporal influence based on nonlinear regression and BP neural network. Highway Eng. **43**(6), 134–139 (2018)
6. Li, Y., Liu, L., Wang, Y.: Short-term traffic flow prediction based on combined prediction model. Transp. Syst. Eng. Inf. **13**(2), 34–41 (2013)