Research of Passenger Flow Forecast of Urban Rail Transit Based on Data Mining

Jinbing Ha, Weidong Chen *

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: 534705872@qq.com

Abstract. With the development of China's economy and society, urban construction is constantly improving, urban rail transit is becoming more mature, and people's demand for travel quality is getting higher and higher. However, the imperfect operation and management leads to the contradiction between supply and demand of urban rail transit. Passenger flow data of rail transit is the basis of operation scheduling, and accurate prediction can effectively improve the utilization rate of operating energy. In this paper, through data mining of passenger flow data, the law of passenger flow in time dimension is analyzed, and three different forecasting models are established for rail transit passenger flow data. Finally, the forecasting effects of each model are compared. The characteristics of passenger flow are analyzed in the time dimension, which shows the different changing rules of passenger flow on working days and rest days. In the discussion of the three forecasting methods, firstly, the time series forecasting method is realized by SPSS software, and the final model parameters are determined by unit root test, autocorrelation analysis, partial autocorrelation analysis and Bayesian information criterion. After that, the regression prediction model of support vector machine and BP neural network model are established by MATLAB. The former maps nonlinear passenger flow data into high-dimensional space to find linear relationship for prediction, while the latter realizes passenger flow prediction by establishing neural network model. Finally, by comparing the three prediction models, the results show that the average absolute error of BP neural network prediction method is 13%, which is 44% and 10% lower than that of time series method and support vector machine method, respectively, with high accuracy.

Keywords: Time series; Support vector machine; BP neural network; Passenger flow forecast; urban rail transit.

1. Introduction

With the development of information age, mobile Internet and Internet of Things have brought huge urban traffic data information. These data collection costs are low, the amount of data is large, and the coverage is wide. Big data technology can make full use of massive residents' living information, mine passengers' travel rules, and help urban rail transit predict passenger flow with advanced data and technology.

The main basis of rail transit control and planning is the prediction of rail transit passenger flow, which can reasonably and scientifically forecast the passenger flow, provide relatively accurate information support to rail transit operators, and timely adjust the train operation plan, which is conducive to improving the operation management and service level of urban rail transit and making full use of urban rail transit resources.

For passengers, the combination of passenger flow forecast information can better grasp the changing trend of passenger flow, make reasonable route planning for travel, reduce time cost, improve passenger satisfaction and comfort, and then improve the service level of urban rail transit and enhance the competitiveness of rail transit.

For rail transit operators, through the passenger flow forecast, we can obtain the accurate passenger flow trend, improve the ability to cope with the peak passenger flow, provide the basis for the operators to make decisions, realize efficient operation, and balance the relationship between operating expenses and service level.

2. Research Methods

2.1. Time series algorithm

Time series is a series of attribute values of something in the real world arranged in chronological order. Time series prediction method is often widely used in statistics. As a quantitative prediction method, its main application is to find its changing trend and law according to historical data, and then predict future data. In different research fields, the time of time series can be year, month, day, hour and other time classifications [1]. The most important factor of time series is time, which is an attribute value.

2.2. Support vector machine

The problems faced by support vector machines are more nonlinear problems, such as the passenger flow prediction of rail transit. The basic linear support vector machines are difficult to play a role, so researchers have changed their thinking. Linear problems and nonlinear problems can be converted into linear problems if the nonlinear problems are too complicated to be solved. But at the same time, support vector machines are relative, which requires us to find a line that makes all data as close as possible, that is, the regression analysis of support vector machines requires that the distance between each sample and the hyperplane is the smallest.

In the process of establishing the regression model of support vector machine, the final goal is to establish a model to make the most data fall within the range of error acceptance. Only when the deviation exceeds the value, the error is calculated, so the quality of the final model leads to the total error.

2.3. BP neural network

BP neural network belongs to feedforward neural network, and it is multi-layered. The characteristic of this network is that the signal can be transmitted in the forward direction and in the reverse direction, and there will be errors if it is carried out in the reverse direction. During forward propagation, the signal needs to be hidden before reaching the output layer [2]. The connection efficiency of neurons in each layer will directly affect the subsequent connection of neurons. When the output requirements of the output layer are not met, the transmission direction can be changed to predict the moral deviation in advance, and the network weights and thresholds can be adjusted to make the predicted output value of BP neural network as close as possible to the expected output[3]. Specifically, for the BP neural network model with only one hidden layer, the work of BP neural network can be divided into two parts: the input layer is sent to the hidden layer, and the output layer is sent to the hidden layer. However, the error should propagate in the opposite direction from the output layer, and the weight and deviation from the input layer to the hidden layer should be adjusted.

3. Prediction model of urban rail transit passenger flow based on data mining

3.1. Model data selection

In the process of modeling and prediction research in this chapter, SPSS is selected as the modeling environment of time series prediction method, and MATLAB is used to realize the modeling and prediction of support vector machine and BP neural network.

SPSS was successfully developed in 1968 and realized various statistical analysis functions in a very early period. Until now, the field of statistical analysis has been continuously expanded, including natural science, social science, theoretical science and other fields, and achieved good results. For users, it is easy to operate, simple to program, powerful, can read a variety of formats of files, these advantages are favored by researchers in various fields. This experiment uses IBM SPSS Statistic version 21, which was updated in August 2012 and can meet the modeling needs of time series prediction methods.

MATLAB, as a commercial mathematics software, has a very high status in the application of mathematics science and technology. MATLAB has more powerful functions and a wider range of applications. It has developed a powerful module set and toolbox for many fields, which is easy for users to achieve. In this experiment, MATLAB9.7R2019b version was selected, which was updated in September 2019. It has complete functions and can better realize the modeling of support vector machine and BP neural network.

3.2. Time series

Time series modeling environment using SPSS software.

First, ADF test was performed on the data, as shown in Table 1.

Table 1. ADF test table.

Differential order	t	p	Critical value		
			1%	5%	10%
0	-5.475	0.000	-3.606	-2.937	-2.067

ADF test verifies whether the time series is stationary, and its null hypothesis is that the series is not stationary. Generally, p value is less than 0.1(0.05 can also be the standard), indicating that the null hypothesis is rejected at the level of 0.1, that is, the series is stable; If the sequence is not stationary, the ADF test can be performed after the first or second order difference until the sequence is stationary [4]. If the second order difference is still not stationary, it is suggested to take the second order as the final difference order.

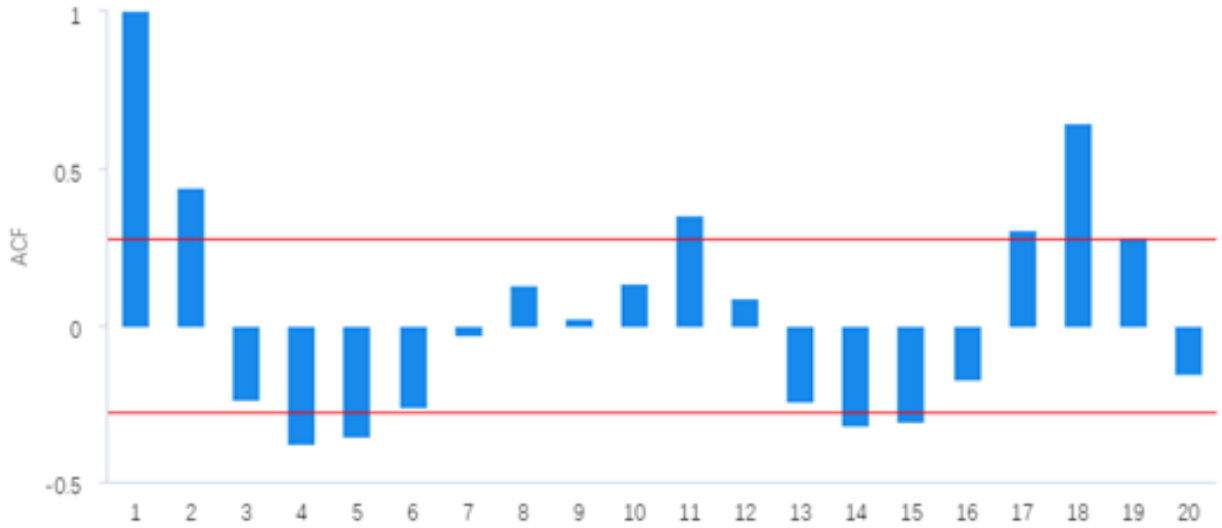

Fig 1. Autocorrelation diagram.

As can be seen from the above table, for passenger flow, the T-statistic of ADF test of this time series data is -5.475, the p value is 0.000, and the critical value is -3.606, -2.937 and -2.607 at 1%, 5% and 10%, respectively. p=0.000<0.01, the null hypothesis is rejected with more than 99% certainty, and the series is stable at this time.

Autocorrelation and partial autocorrelation were analyzed, as shown in Figure 1 and Figure 2.

Fig 2. Partial autocorrelation diagram.

3.3. Support vector machine

MATLAB software was used to implement support vector machine algorithm modeling. After importing the dataset newdata. Mat, the dataset was divided into input indicators and output indicators, and the mapminmax function was used to normalize the data. 80% of the data set is selected as the training sample, and the remaining samples are the test sample. The penalty coefficient C value and kernel radius G value are determined by parameter optimization. Figure 3 and 4 show the 3D view after parameter optimization and contour map after projection respectively. The X-axis is the value of$log_2 g$, the Y-axis is the value of$log_2 c$, and the Z-axis is the value of root mean square error MSE. It can be seen that after parameter optimization, the C value is 1048576, and the G value is 0.0013811, the corresponding $log_2 g$ is 20 and $log_2 c$ is -9.499967.

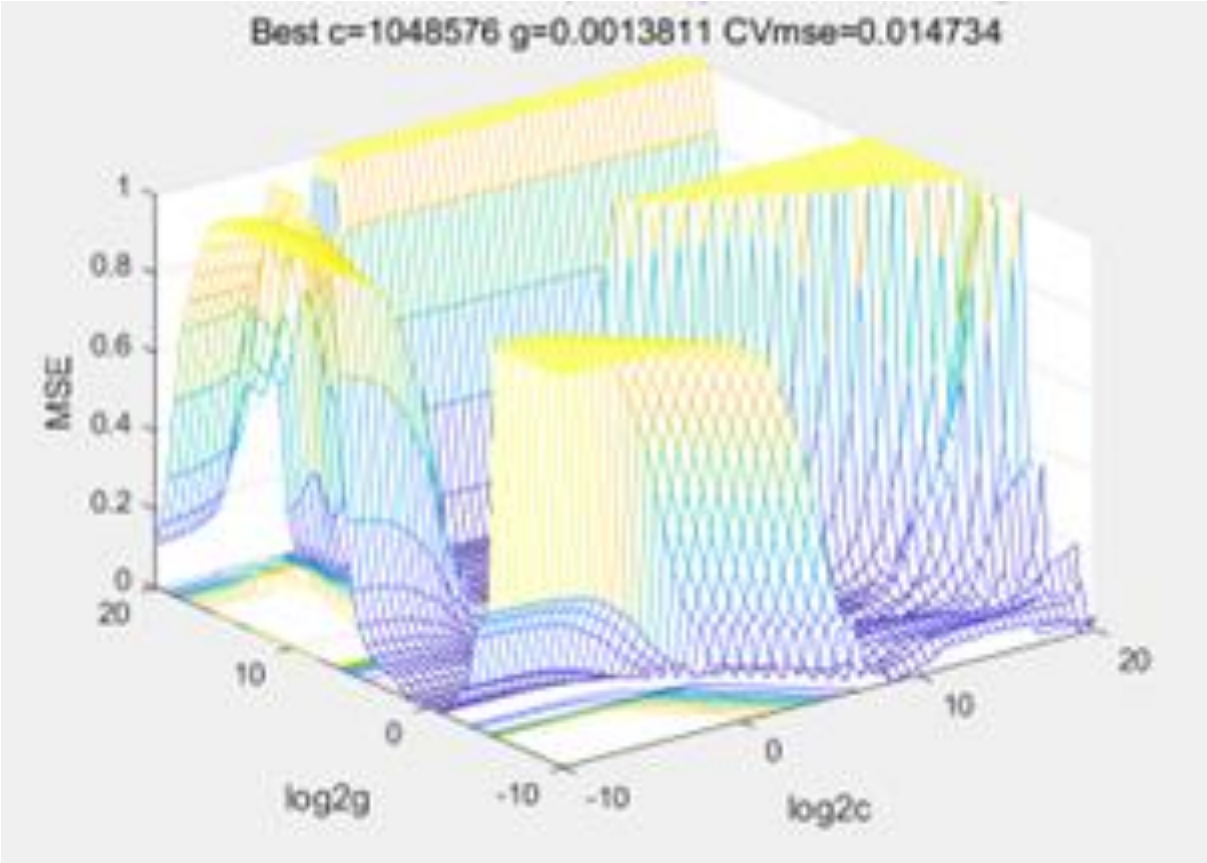

Fig 3. Parameter optimization (3D view).

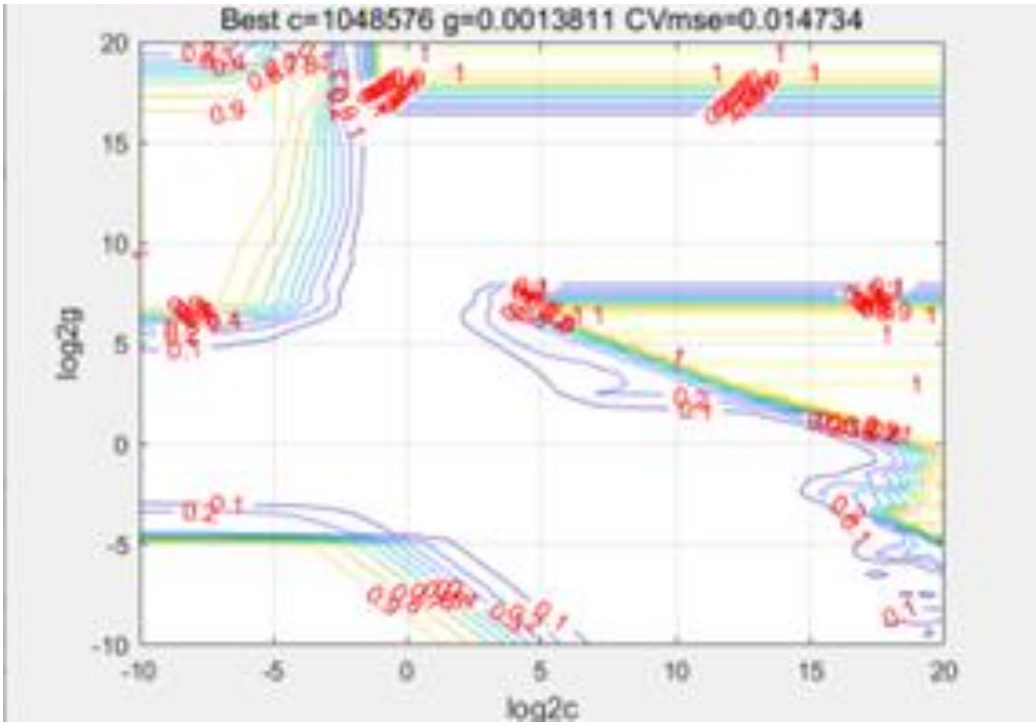

Fig 4. Parameter optimization (contour map).

3.4. BP neural network

MATLAB software is chosen to establish the BP neural network model. Through the BP neural network, we take the passenger flow data of every two days as the input index, and the passenger flow data of the third day as the output index, so as to realize the unified summary of the input index input1 and output index output1 using mapminmax function standard. And extracted 80% of the data as the sample tran_data, the rest of the data as the test sample check_data.

According to the empirical formula of BP neural network to determine hidden layer nodes, the range of hidden layer nodes is between 3 and 11. In this process, we use sample data to test the effect of network model under different hidden layer nodes, and determine the best number of hidden layer nodes by comparing the MSE of different hidden layer nodes. The experimental results are shown in Table 2.

Table 2. BP hidden layer node selection experiment.

Number of hidden layer nodes	MSE
3	9.27e-3
4	9.28e-3
5	9.92e-4
6	8.45e-4
7	5.02e-4
8	7.11e-4
9	6.55e-4
10	6.07e-4
11	5.89e-4

The tansig function is the hidden layer function, the trainlm function is the training function, the learngdm function is the weight learning function, and the mse function is the performance function. Note that the above functions are by default. The total number of training times is 1000, the final goal is 0.00001, and the calculated learning rate is 0.1. Figure 5 shows the convergence of this error. It can be seen that the model established at this time converges rapidly before 200 times of training, and gradually tends to a value after 200 times of training without drastic changes.

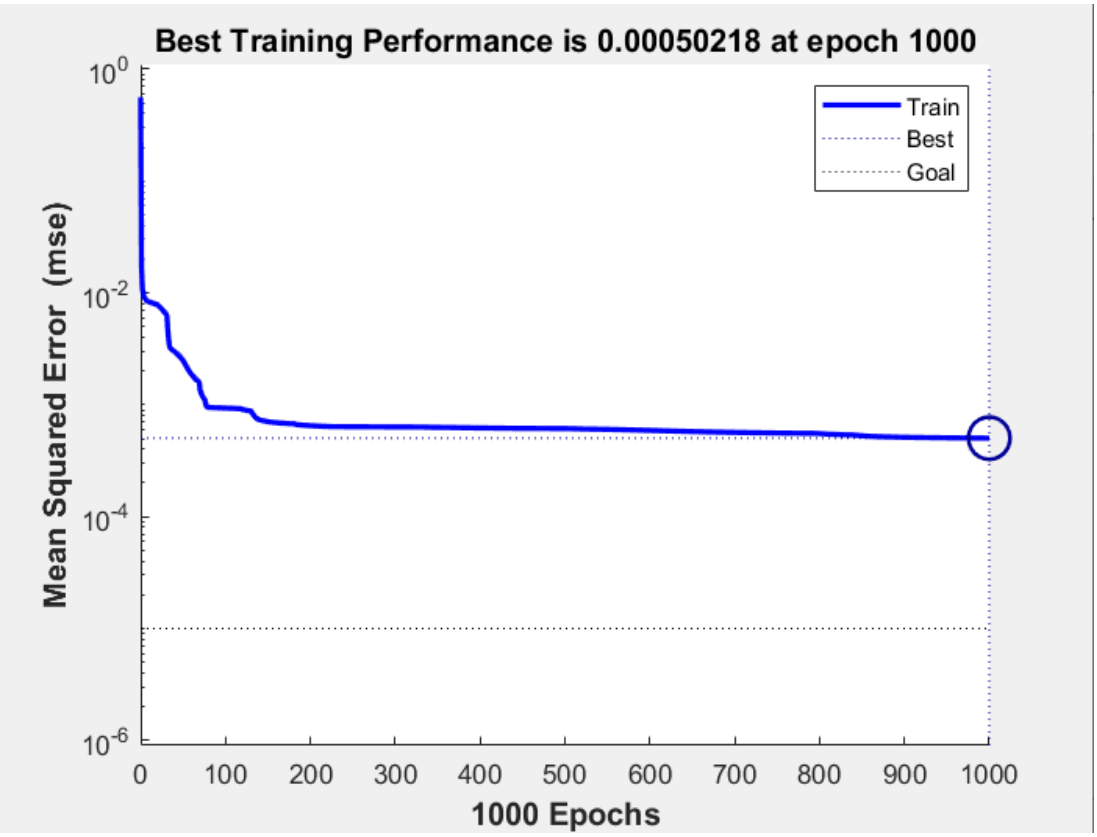

Fig 5. Training error.

4. Summary

The current rail transit passenger flow data are sorted out, and the study is carried out on the basis of time distribution by data mining. After data comparison, the change rules of passenger flow on working days and rest days are mastered, which lays a foundation for the next passenger flow prediction.

Using time series prediction method, support vector machine prediction algorithm, BP neural network prediction algorithm three methods of passenger flow prediction research, comparing the three prediction methods, the results show that BP neural network prediction has a high accuracy, the average absolute prediction error is 13%. Compared with the time series prediction method and the support vector machine prediction method, the error is reduced by 44 and 10 percentage points respectively, and the prediction effect is the best.

References

[1] Voort M V D, Dougherty M, Watson S. Combining Kohonen maps with ARIMA time series models to forecast traffic flow [J]. Transportation Research Part C Emerging Technologies, 1996, 4(5):307-318.

[2] McCulloch, WS & Pitts, W. H. (1942). A logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 5. 115-133.

[3] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[4] K.-H. Kim and H.-S. Kim, “KTX Passenger Demand Forecast with Intervention ARIMA Model,” Journal of the Korean society for railway, vol. 14, no. 5, pp. 470–476, Oct. 2011.