Study on the Prediction Model of Short-term Bus Passenger Flow Based on Big Data

Cheng Wang[1, a], Zhiying Cao[1], Xiuguo Zhang[1], Weishi Zhang[1], Huawei Zhai[1]

[1]*School of Information Science and Technology,Dalian Maritime University,Dalian,China*

Abstract. Prediction of short-term bus passenger flow can help bus managers timely and accurately get the changes of the passenger flow and make scientific and reasonable vehicle scheduling to meet passengers' needs. In this paper, a SLMBP model is constructed to predict the bus passenger flow. The SRCC(Spearman rank correlation coefficient) method is used to determine the factors that have significant influence on passenger flow changes. The Levenberg-Marquardt algorithm is used to optimize the BP neural network to avoid getting stuck in local optimal solutions and prompt the convergence speed. A SLMBP neural network parallel algorithm is constructed to perform multiple stations prediction. The experimental results show that the SLMBP neural network parallel algorithm can not only guarantee the accuracy of short-term passenger flow prediction, but reduce the time spent on model learning and prompt the prediction speed.

1 Introduction

Practical experiences have shown that encouraging people to take public transportation can solve the problem of congestion faced by cities effectively. Through the analysis of the previous bus passenger flow and the prediction of future passenger number, it is possible for bus companies to take effective measures in time and ensure the allocation of bus resources to meet passengers' demand. It is of great significance to optimize urban bus system and assist decision-making. So it is important to use proper model algorithms to accurately predict short-term bus passenger flow.

Most short-term passenger flow prediction model algorithms run in stand-alone mode and appear to be inefficient in multiple stations prediction with large-scale data. This paper designs a SLMBP neural network parallel algorithm based on Hadoop, which can greatly prompt the prediction speed.

2 Related Works

Yue et al. [1] proposed a wavelet neural network prediction model based on bat algorithm optimization of adaptive t distribution variation, but the algorithm was more complicated, and the prediction accuracy was not improved much. Mahendran et al. [2] used the SARIMA model to predict monthly passenger flow. The model only highlighted the role of time factors and ignored the impact of other factors on passenger flow for it was a time series model. The short-term passenger flow at bus station was predicted by parallel BP neural network algorithm based on Hadoop platform [3]. Although this method has greatly improved the efficiency in processing and training massive passenger flow data, its convergence speed is slow and can easily fall into local optimal solutions.

In this paper, Levenberg-Marquardt [4] is used to optimize the BP neural network [5] to construct the SLMBP passenger flow prediction model, which solves the problems that BP neural network is easily trapped in the local optimal solutions and the convergence speed is slow. Because the SLMBP algorithm in stand-alone mode can't predict short-term passenger flow in multiple bus stations in a short time, this paper designs a parallel algorithm of SLMBP neural network based on Hadoop, which satisfies the demand of storage and real-time processing of massive data in short-term passenger flow predicting.

3 SLMBP bus passenger flow prediction model

Firstly, SRCC method is used to determine the factors that have a significant influence on the change of passenger flow. These factors and historical passenger flow data are input into the SLMBP prediction model to conduct training and prediction.

3.1 Analysis of influencing factors of passenger flow

Through analysis of lots of historical data, it is found that temperature, whether it rains or snows, whether it is weekday and other factors will affect the number of

[a] Corresponding author: 1341116439@qq.com

passenger flow.When preprocessing the raw data, the factors should also be quantified, as shown in table 1.

Table 1. Quantitative Results of Each Influencing Factor

Factors	Quantitative Results
temperature	measured temperature
rain and snow conditions	Sunny, rain and snow were quantified as 1, 0
Whether it is weekday	Working days and non-working days are quantified as 1,0
Wind force	Measured wind force (grade 1-12)
Humidity	Measured humidity

In this paper, the SRCC method [6] is used to determine the factors that have a significant impact on passenger flow changes, and factors that have less influence on passenger flow are removed to simplify the input of the prediction model.

Through lots of researches, the temperature, rain and snow conditions have a significant impact on the change of passenger flow, so these factors are input into the prediction model.

3.2 Establishment of SLMBP Bus Passenger Flow Predicting Model

3.2.1 Design the input layer of the SLMBP prediction model

In this paper, MSE (mean square error) is used as evaluation standard to determine the number of optimal input nodes in the input layer and the formula of MSE is shown in equation (1). The predicted value of passenger flow is Y, while the actual value is $Y'$and n is the number of predicted time periods.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}(\mathrm{Y}-\mathrm{Y}')^2 \tag{1}$$

Many experiments have shown that when the input is the passenger flow at the same time and the last time in the two weeks before the predicting time, and the passenger flow in the first two moments of the predicted time, the rain and snow conditions on the predicting day, the minimum temperature, the highest temperature and whether it's weekday, the MSE of the predicting model is the smallest, so the above data is used as the input of the prediction model, a total of 12 input nodes.

3.2.2 Design the output layer and hidden layer of SLMBP prediction model

The number of nodes of the output layer in this experiment is 1. After lots of predictions, it is verified that when the hidden layer is 11, the MSE is the smallest, so the number of hidden layer nodes is determined to be 11. In this paper, logarithmic Sigmoid function is used as the excitation function of SLMBP passenger flow prediction model.

3.2.3 Training SLMBP prediction model

The data of historical passenger flow and factors affecting passenger flow constitutes the training sample $\mathrm{X} = (x_1, x_2, \dots, x_n)^T$, x_i represents the input data of the SLMBP prediction model; $\mathrm{Q} = (q_1, q_2, \dots, q_n)^T$ is the expected output corresponding to X. q_i represents the expected output corresponding to each x_i. The actual output vector of the prediction model is $\mathrm{Y} = (y_1, y_2, \dots, y_n)^T$ and y_i represents the actual output of each x_i after the calculation of the model. The threshold of the j-th neuron in the output layer is θ_j.

1) The threshold of neurons in initial output layer is θ_j; the target error value is E'; the learning rate is η; maximum number of training is n and initial value is μ.

2) Input training sample data at the input layer to start the operation and output y_i at the output layer, then calculate the error function vector $\mathrm{e} = [e_1, e_2, \dots, e_n]$ of the training sample.

$$e_i = y_i - q_i \tag{2}$$

The sum of squares of errors is E.

$$\mathrm{E} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}_\mathrm{i}^2 \tag{3}$$

The update equation of the overall weight and threshold after the k-th iteration is

$$\mathrm{w}(\mathrm{k}+1) = \mathrm{w}(\mathrm{k}) - [\mathrm{J}^{\mathrm{T}}\mathrm{J} + \mu\mathrm{I}]^{-1}\mathrm{J}^{\mathrm{T}}\mathrm{e} \tag{4}$$

$$\theta(\mathrm{k}+1) = \theta(\mathrm{k}) - [J^T J + \mu I]^{-1} J^T \mathrm{e} \tag{5}$$

In equations above, J is a Jacobian matrix, μ is a very small proportional coefficient and I is a unit matrix. μ is set to be dynamically to reduce the number of loops in the iteration [7].

In equations (4) and (5), it is very time consuming because it involves the inverse operation of the matrix. In this paper, the LU decomposition method [8] is used to determine the amount of change Δw of the weight, and the inverse matrix is not needed, which greatly improves the algorithm calculating speed.

3) Repeat step 2) to iteratively calculate and update the weights and thresholds. The algorithm is considered to be convergent when the highest iteration number n or E is less than or equal to the target error value, and the prediction model is trained well.

4) The trained SLMBP prediction model is applied to passenger flow prediction.

3.3 Design SLMBP Parallel Algorithm Based on Hadoop Platform

In this paper, the SLMBP model algorithm is applied on Hadoop [9] platform for MapReduce decomposition, and its implementation is data running parallel. This method is mainly composed of Map function, Reduce function and driver function. The main steps are as follows:

1) Map function: The entire process of learning operations of the SLMBP prediction model is defined in

each Map function. Enter the training sample data of a certain bus station corresponding to each Map function, which is represented by data. The Map function reads the weights and thresholds corresponding to data in shared directory on HDFS, represented by value.

The output of Map function is key-data value pairs <data, ErrorFunction>. The data is the input training instance data, and the ErrorFunction is the correction value related to the weight and threshold of each data.

2) Reduce function: The Reduce function receives the <data, ErrorFunction> data output from each Map function and starts to calculate the correction amount of the weight and threshold of each SLMBP prediction model in each Map function, and then each weight and threshold are updated and written into the shared directory for use in the next iteration.

3) Driver function: Responsible for starting the Hadoop task and controlling the execution of the task, as well as initializing the weight of the SLMBP algorithm and controlling the iteration of SLMBP algorithm during the training process.

4 Experimental results and analysis

4.1 Algorithm convergence speed and prediction accuracy test

This paper selects the passenger flow survey data of the 12th buses in Dalian from 2017-08-08 to 2017-09-04 and the passenger flow influencing factors data on the predicting day for the training of predictive models.

The MSE is used as a criterion for the prediction accuracy of the model. The MSE, the total time-consuming and training times in convergence of the SLMBP neural network prediction model, the standard BP neural network prediction model and RBF neural network prediction model for predicting the above the station are shown in Table 2.

Table 2. MSE and Total Time-consuming and Convergence Speed of Three Models

Model	MSE	Time-consuming	Training times in convergence
Standard BP	54.87	15s	Divergent
SLMBP	24.75	11s	150
RBF	39.24	11s	137

4.2 Mass Data Running Test

In this experiment, the 10th bus route was selected and the data of 24 hours for 15 minutes interval was placed on the Hadoop ten-machine cluster for training with SLMBP model. Table3 is the comparison of total time-consuming and total MSE in cluster and stand-alone.

Table 3. Comparison of Total Time-consuming and Total MSE in Cluster and Stand-alone

Model	MSE	time-consuming(s)
stand-alone	1157.24	199
cluster	1159.18	15s

5 Conclusion

In this paper, SLMBP prediction model was constructed for short-term passenger flow prediction. Experimental results showed that the SLMBP prediction model can effectively predicted the short-term bus passenger flow. And SLMBP parallel algorithm based on Hadoop platform was designed simultaneously. Results of experiments showed that the parallel process reduced the time spent in model learning and training compared with the prediction model in the stand-alone state, greatly improving the operation speed of the prediction model.

In addition, the model must be trained with a large amount of historical data to reach high prediction accuracy, so it is necessary to ensure that sufficient and real-time passenger flow data and influencing factors data are obtained.

References

1. X.Yue, Y.Zheng, J.Lin. The Forecast of Urban Rail Transit Passenger Flow Based on Improved WNN[J]. Computer Engineering and Applications, 52(11): 227-232(2016)
2. Shitan, Mahendran & Kumar Karmokar, Provash & Yung Lerd, Ng.Time series modeling and forecasting of ampang line passenger ridership in Malaysia. Pakistan Journal of Statistics. 30. 375-386(2014)
3. Z.Li. Research and Forecast of Bus Passenger Flow Based on Hadoop Platform [D]. Northeast Normal University(2015)
4. Y.Li, J.Lei, J.Yang, et al. Classification of Tieguanyin Tea with an Electronic Tongue and Pattern Recognition[J]. Analytical Letters, 47(14):2361-2369(2014)
5. T.Guo, X.Liu, S.Hao, et al. Prediction of Equivalent Electrical Parameters of Dielectric Barrier Discharge Load Using a Neural Network[J]. Plasma Science and Technology, 17(3):196-201(2015)
6. Bickel D R. Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically[J]. Bioinformatics, 19(7):818(2003)
7. B.Shi, X.Zhu. Research on Improved Algorithm of LMBP Neural Network[J]. Control Engineering, 15(2): 164-167(2008)
8. Y.LI, H.Huang. Research on convergence speed improvement of LMBP algorithm in neural network[J]. Computer Engineering and Applications, 42(16): 46-49(2006)
9. Z.Yu, W.Zhang, H.Ge,W.Ai,Y,Sun. Log Analysis Model Based on Hadoop Platform[J].Computer Engineering and Design, 37(02): 338-344+428 (2016)