Highway Traffic Flow Forecast Based on Big Data Analysis

Sen Wu

Shanxi Provincial Transportation Technology Research and Development Co., Ltd., Taiyuan, Shanxi, 030000, China

Keywords: Highway traffic flow forecast; Big data analysis; LSTM model; Accuracy of prediction; Time series change

Abstract: The purpose of this article is to study the algorithm model of highway traffic flow prediction based on big data analysis, so as to provide accurate traffic flow prediction information and support the decision optimization of traffic management departments. In order to achieve this goal, this article adopts LSTM (Long-term and short-term memory network) as a forecasting model, and uses its advantage of capturing the long-term dependence of sequence data to forecast highway traffic flow. During the experiment, a representative traffic flow data set is selected, which is used for model training and testing after preprocessing and feature engineering. Through continuous training and optimization, we get a stable and accurate LSTM prediction model. The experimental results show that LSTM model has obvious advantages in highway traffic flow forecasting, which can capture the time series changes of traffic flow and has good forecasting ability for both short-term and long-term traffic flow changes. Especially in peak hours and complex road sections, the prediction accuracy of the model is relatively high.

1. Introduction

In today's era, with the rapid economic growth and the accelerated pace of urbanization, highway traffic, as a key channel to connect urban networks in series and promote regional interaction, has a direct impact on people's travel efficiency, the fluency of logistics distribution and the degree of informatization of urban management [1]. Therefore, the prediction of highway traffic flow has become particularly critical. Accurate traffic forecast can not only help traffic management agencies to make decisions, optimize traffic network resources and reduce congestion, but also provide real-time and practical travel information for the public and improve travel experience [2]. The rise of big data technology has brought broad application prospects for highway traffic flow forecasting. Through in-depth analysis of a large number of traffic data, we can find potential laws, so as to grasp the changes of traffic flow more accurately [3]. This provides a strong support for the construction of intelligent transportation system.

Research on highway traffic flow prediction has achieved certain results both domestically and internationally. At present, the mainstream research methods include time series analysis, machine learning model and deep learning algorithm [4]. These methods all show good prediction results under certain conditions. However, with the complexity of traffic system and the sharp increase of data, the existing forecasting technology still needs to face many technical challenges [5]. For example, how to deal with high-dimensional and heterogeneous traffic data, how to enhance the real-time and stability of the forecasting model, and how to accurately capture the nonlinear dynamic characteristics of traffic flow [6]. Although the research results are rich, there is still room for research in multi-source data fusion, improvement of prediction accuracy and extension of prediction duration, which requires further research and innovation.

Based on this background, this article is devoted to deeply discussing the prediction of highway traffic flow. The research goal is to develop a more accurate and efficient traffic flow forecasting method by combining big data analysis technology. The research content involves the key steps of traffic flow data collection, preprocessing, feature extraction and prediction model construction. The research results of this article are expected to bring new perspectives and methods to the field of highway traffic flow prediction and promote the development and application of intelligent

transportation systems.

2. Theoretical basis of big data analysis and highway traffic flow forecast

In today's wave of informatization, data has become an indispensable wealth. As a core tool to tap the deep value of this resource, big data analysis technology is increasingly penetrating into various industries [7]. By collecting, storing, processing and analyzing a large number of fast and diversified data, it reveals the mystery and trend behind the data. In the field of road traffic, the importance of big data analysis technology is self-evident. Using this technology, we can deeply mine huge data such as traffic flow, speed and occupancy rate, which provides strong data support for traffic flow prediction [8]. Highway traffic flow shows multi-dimensional characteristics such as time variability, spatial interaction and unpredictability. Time variability is manifested in the fluctuation of traffic flow with time, for example, the traffic flow at peak hours is far more than that at ordinary times. Spatial interaction is manifested in the interaction of traffic flows between different road sections and intersections, and congestion often leads to chain effects. Unpredictability comes from many uncertain factors in the traffic system, such as weather and accidents, which makes traffic flow prediction full of challenges. A deep understanding of these characteristics will help us to grasp the changing law of traffic flow more accurately and lay the foundation for the design of forecasting model.

At present, many technical means have been widely used in the research of traffic flow forecasting methods [9]. The deep learning method uses the powerful learning ability of neural network to model and predict complex traffic flow data. Each method has its own advantages and is suitable for different scenarios, and it needs to be chosen according to the specific situation [10]. When selecting the traffic flow forecasting algorithm model, we must consider many factors comprehensively. The first is the characteristics and quality of data, and different data adapt to different algorithm models. Secondly, the accuracy and timeliness of prediction are required, and different scenarios have different expectations for the accuracy and timeliness of prediction results. Futhermore, we need to consider the complexity and interpretation of the model, as well as the stability and robustness of the algorithm. Based on these factors, we can select the algorithm model that best meets the actual needs and provide an accurate and stable solution for highway traffic flow forecasting.

3. Highway traffic flow forecasting algorithm model based on big data analysis

In the prediction of highway traffic flow based on big data analysis, data preprocessing and feature engineering are crucial steps. There are some problems in the original traffic data, such as missing, abnormal and noise, which will affect the accuracy of the prediction model. Therefore, this article first cleans the data, removes or fills in missing values, corrects abnormal data, and smoothes the noise. Then, feature engineering is carried out to extract useful features from the original data for traffic flow prediction. These characteristics may include time characteristics (hours, weeks, months, etc.), spatial characteristics (road section location, intersection type, etc.) and traffic state characteristics (speed, occupancy rate, etc.). Data preprocessing and feature engineering can provide high-quality data input for the subsequent algorithm model construction.

In the aspect of algorithm model construction, LSTM is selected as the prediction model. In order to improve the accuracy of prediction, the input data set contains historical traffic flow data, and also integrates time characteristics (such as hours, weeks, months, etc.) and weather characteristics (such as temperature and precipitation, etc.) in the hidden layer. The LSTM unit controls the flow of information through forget gate, input gate and output gate, so as to realize the memory of historical information and the prediction of future state. The formulas are as follows:

Forget gate:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \quad (1)$$

Input gate:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \quad (2)$$

Output gate:

$$O_t = \sigma\left(W_O \cdot \left[h_{t-1}, x_t\right] + b_O\right) \quad (3)$$

Where: f_t, i_t, o_t is the output of forgetting gate, input gate and output gate respectively. σ is the activation function. W_f, W_i, W_O is the weight matrix. b_f, b_i, b_O is an offset. h_{t-1} is the hidden state of the last moment. x_t is the input of the current moment.

In the key link of parameter setting, this paper closely combines the characteristics of data itself and the specific requirements of forecasting work, and makes detailed adjustments to several core parameters of LSTM. The number of LSTM units, learning rate, batch size and other parameters are set as shown in Table 1:

Table 1: LSTM Model Core Parameter Settings Table

Parameter name	Parameter value
Number of LSTM units	128
Learning rate	0.001
Batch size	64
Number of hidden layers	2
Offset term	True
Dropout rate	0.2
Bidirectional LSTM	True
Sequence length	10
Batch processing first dimension	True
Initial hidden state	Random initialization
Initial cell state	Random initialization

The training and optimization process of the model is the core to improve the prediction accuracy. At this stage, the pretreated normative data set is fully used to train the LSTM model. Through the back propagation algorithm, the model will adjust its internal weight and deviation according to the prediction error in each iteration, and gradually approach the real traffic flow change law. The training process is expressed as:

$$\theta^* = \mathrm{Train}\left(\mathrm{D}_{\mathrm{train}}, \mathrm{Model}, \theta\right) \quad (4)$$

Where $-\theta^*$ represents the optimized network parameters obtained after training. θ represents the initial network parameters. Train stands for training process.

In order to prevent the model from over-fitting in the training process, this paper adopts two powerful weapons: cross-validation and regularization technology. Cross validation can evaluate the performance of the model on different data subsets and ensure the generalization ability of the model. Regularization technology limits the complexity of model parameters by introducing additional constraints, so as to effectively prevent the model from over-fitting the training data. After numerous training and optimization iterations, an accurate and stable LSTM prediction model is finally obtained. This model provides strong support for the prediction of highway traffic flow.

4. Experimental verification and result analysis

In order to verify whether the algorithm model of big data analysis in highway traffic flow forecast is effective, this section designs a set of experimental schemes. The main purpose of the experiment is to measure the prediction performance of LSTM model on actual traffic data. Therefore, a typical road traffic flow data set from an open traffic database is selected. This data set

covers the traffic signals at different time points and different road sections.

Before the model training, the original data are strictly preprocessed and feature engineering is carried out. Data preprocessing includes steps such as missing value filling, abnormal value processing and noise smoothing to improve data quality. In the aspect of feature engineering, I extracted a variety of features from my original data that are helpful to improve the prediction accuracy. The data after these steps are converted into a standard format suitable for the input of LSTM model.

Next, the preprocessed data set is divided into two parts: training set and test set, with a ratio of 7:3, so as to ensure the generalization ability of the model to unknown data. Then start model training according to the LSTM model architecture and parameter configuration. In the training, Adam optimizer is used to accelerate the convergence of the model, and the hyperparameters such as learning rate are adjusted by monitoring the change of loss function.

After the model training is completed, the test set is used to evaluate its performance. MSE (Mean Square Error) and MAE (Mean Absolute Error) are selected as evaluation indexes. By predicting the data in the test set and comparing it with the real value, the calculated MSE and MAE values are used to quantify the performance of the model (see Table 2).

Table 2: Overall Performance Evaluation Metrics of the LSTM Model in Traffic Flow Prediction

Evaluation Metric	Value
MSE	150.23
MAE	10.45

The LSTM model has high accuracy in predicting traffic flow. The MSE value is 150.23, relatively low, indicating good predictive stability of the model. The MAE value is 10.45, indicating that the model has a relatively small error in predicting traffic flow.

Table 3: Comparison of Predicted and Actual Traffic Volumes

Time Period	Road Section Description	Actual Traffic Volume (vehicles/hour)	Predicted Traffic Volume (vehicles/hour)	Difference (vehicles/hour)
06:00-07:00	Urban Main Road	235	240	-5
07:00-08:00	Urban Main Road	310	305	+5
08:00-09:00	Urban Main Road	350	355	-5
17:00-18:00	Peak Hours	450	445	+5
18:00-19:00	Peak Hours	475	470	+5
19:00-20:00	Complex Section	290	295	-5
20:00-21:00	Simple Section	180	185	-5
21:00-22:00	Simple Section	150	155	-5

It is not difficult to find from Table 3 that there is little difference between the predicted results of LSTM model and the actual traffic at different time periods and road sections. The model can accurately capture the changing trend of traffic flow in urban main roads, rush hours and complex and simple sections. During the rush hour (such as 17:00-19:00), although the traffic conditions are complex and changing rapidly, the model can still maintain a good prediction accuracy with a difference of only 5 vehicles/hour. For complex sections (such as 19:00-20:00) and simple sections (such as 20:00-22:00), the model also performs well. Even when the traffic flow is relatively low, the prediction error remains at a low level.

Table 4: Prediction Performance of the LSTM Model in Different Time Periods and Road Segments

Time Period/Road Segment	MSE	MAE
Peak Hours	120.34	9.12
Off-Peak Hours	180.12	11.78
Complex Road Segments	130.56	10.23
Simple Road Segments	160.89	10.90

This article also analyzes the prediction effect of the model in different time periods and road

sections, and finds that the prediction accuracy of the model is particularly prominent in peak periods and complex road sections, as shown in Table 4. The prediction accuracy of the model is particularly prominent in peak hours and complex road sections. In the peak period, the MSE value is 120.34, and the MAE value is 9.12, both of which are lower than the corresponding values in the off-peak period, which shows that the model can still maintain high prediction accuracy in the period with large traffic flow. In complex road sections, the MSE value is 130.56 and the MAE value is 10.23, which are relatively low, indicating that the model can handle complex road conditions and accurately predict traffic flow. This further confirms the advantages of LSTM model in the field of highway traffic flow prediction.

Table 2 and Table 3 respectively show the overall performance of LSTM model in traffic flow forecasting and the forecasting effect in different time periods and road sections, and quantify the forecasting accuracy and advantages of the model through specific numerical values. Through the experimental verification and result analysis, we fully proved the effectiveness and practical application value of the highway traffic flow forecasting algorithm model based on big data analysis.

5. Conclusions

Through experimental verification, this article fully demonstrates the excellent forecasting ability of highway traffic flow forecasting algorithm model based on big data analysis, especially the LSTM model in actual combat. With its powerful time series data processing ability, the model can accurately capture the dynamic changes of traffic flow and provide timely and accurate forecasting information for traffic management departments. This information plays a vital role in optimizing the allocation of road network resources, planning traffic diversion schemes in advance, and effectively alleviating traffic congestion.

Looking ahead, the field of highway traffic flow forecasting is still full of challenges and opportunities. We will continue to dig deep into this field and actively explore and try more advanced algorithms and technologies. Futhermore, we will pay close attention to the latest development of big data technology in the transportation field, especially the application of multi-source data fusion technology and real-time data processing technology. By integrating more dimensional traffic data and realizing real-time processing and analysis of the data, we can further improve the real-time and robustness of the model and make it better adapt to the complex and changeable traffic environment. We believe that with the continuous progress of technology and the continuous accumulation of data, highway traffic flow forecast will become more accurate and efficient. This will inject new vitality into the development of intelligent transportation system and push the traffic management to a more intelligent and refined direction.

References

[1] Feng Xinyi, Liu Yiting, Xiao Zhixiong, et al. Research on the Driving Forces of Urban Traffic Congestion Supported by Multi-source Big Data [J]. Geospatial Information, 2024, 22(7): 36-40.

[2] Wang Jiusheng, Dai Xuhai, Miao Zhongyan, et al. A Time Series Prediction Model for Highway Traffic Flow Based on Multivariate Spatio-temporal Relationships [J]. Journal of Highway and Transportation Research and Development, 2023, 40(10): 175-182.

[3] Cui Nana, Xia Haishan, Zhang Chun, et al. Research on the Spatio-temporal Premium Effect of Urban Rail Transit Based on Network Big Data [J]. Geography and Geo-Information Science, 2022, 38(1): 133-137.

[4] Liao Kai, Zhang Runtao, Yang Zian, et al. Architecture and Application of a Big Data Platform for the Integration of Transportation and Energy [J]. Automation of Electric Power Systems, 2022, 46(12): 20-35.

[5] Wang Xuefei, Ding Weilong. A Short-term Traffic Flow Prediction Method for Expressway Big

Data [J]. Journal of Computer Applications, 2019, 39(1): 87-92.

[6] Gu Yanyan. Research on the Interaction between Urban Public Transportation and Land Use Supported by Big Data [J]. Acta Geodaetica et Cartographica Sinica, 2023, 52(10): 1799-1799.

[7] Yang Yanni, Xi Yukun, Shen Yuanfei, et al. Research on the Characteristics of Travel Mode Choice in Public Transportation Systems Driven by Big Data [J]. Journal of Transportation Systems Engineering and Information Technology, 2019, 19(1): 69-75.

[8] Feng Zhiming, You Shiqing, You Zhen. Research on the Spatial Identification and Structural Characteristics of Chinese Urban Agglomerations Based on Traffic Big Data Networks [J]. Geographical Research, 2023, 42(7): 1729-1742.

[9] Pan Ying, Xu Wenjie, Yan Yanwen, et al. Exploration on the Construction of a Big Data Resource Catalog for Urban Rail Transit [J]. Urban Rail Transit Research, 2021, 24(6): 141-145.

[10] Xiang Yun, Xu Chengcheng, Yu Weijie, et al. Research on the Dominant Travel Distances of Urban External Transportation Modes Based on Population Migration Big Data [J]. Journal of Transportation Systems Engineering and Information Technology, 2020, 20(1): 241-246.