Comparison of Machine Learning Algorithms for Sequential Dataset Prediction

Zhuang Ma[1], Tao Shen[1(✉)], Zhichao Sun[2], Kaining Xu[3], and Xingsheng Guo[3]

[1] School of EE, University of Jinan, Jinan 250022, China
cse_st@ujn.edu.cn
[2] Inspur Electronic Information Industry Co., Ltd., Langchao Rd. 1036, Jinan 250101, China
[3] Jinan Lingsheng Info Tech. Co., Ltd., Huaiyin District, Jinan 250000, China

Abstract. Accurate traffic flow prediction can provide basis for traffic control and travel planning. Accurate prediction is very important to the control and management of traffic flow in large cities. However, on the one hand, the traffic flow data information has the complicated interior space design relevance of discrete systems, that is, different kinds of connection relevance between different pavement nodes. On the other hand, it has dynamic duration correlation, that is, the spatial correlation of road nodes will change with time. At the same time, traffic flow data information generally shows a certain periodicity, but there are also some anomalies and specificity. According to XGBoost algorithm of artificial intelligence algorithm, LSTM optimization calculation method is created and compared. Based on the practical exploration of data information, it can be concluded that LSTM digital model can clearly predict traffic flow, and LSTM is better than traditional equipment learning model.

Keywords: Deep learning · Prediction · Time series · LSTM

1 Introduction

With the steady improvement of the development level of the national economy, the data of automobiles continues to grow, while the phenomenon of traffic congestion is inevitable. Now traffic prediction is a new idea, which is of great help to improve the development of regional economy and society. Now the proposal of intelligent transportation provides strong support for the prediction of traffic flow. The traditional machine learning [1] algorithm can predict the traffic flow, but its accuracy and accuracy have been improved. This paper improves the prediction based on the traditional machine learning algorithm XGBoost, and compares the previous deep learning algorithm LSTM. The main comparative data include MSE, RMSE, MAE and R-Squared to evaluate the model. The other part of the paper is to screen and preprocess the experimental data of traffic flow, train the two algorithms and adjust the hyperparameters to select the most

This work is supported by Shandong Key Technology R&D Program 2019JZZY021005 and Natural Science Foundation of Shandong ZR2020MF067.

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 535, pp. 380–385, 2024.
https://doi.org/10.1007/978-3-031-50580-5_33

appropriate algorithm model to predict the best results. For LSTM algorithm, the number of hidden layers, the number of hidden layer units, learning rate and exit rate are proposed to optimize four super parameters. The last part of the paper is the prediction result and summary of the algorithm [2].

2 Background

Most predictions of traffic flow are supported by the Internet of Things. The Internet of Things allows real-time traffic data to be collected in the cloud through smart sensors and other hardware devices and analyzed and processed via the Internet. The data can help traffic authorities understand real-time traffic flows and take effective measures [3]. The current Internet of Things is all over the place, the nonlinear load generated during the operation of the power system will affect the stability. Aiming at this, a nonlinear prediction method based on the distribution Internet of Things is proposed to realize the prediction of the nonlinear load with high precision and controllability through the model [4]. The prediction of wild vegetation based on the Internet of Things can also be applied to ecology [5]. The Internet of Things has been applied in various fields and has been developed accordingly.

The development of IoT technology also benefits from the attention and support of the government and enterprises, as well as the recognition of its potential value in improving production efficiency, reducing costs and improving quality of life [6]. The development of IoT technology [7] also benefits from the attention and support of the government and enterprises, as well as the recognition of its potential value in improving production efficiency, reducing costs and improving quality of life.

3 Related Work

At this stage, in the traffic flow forecasting industry, relevant research staff have already made a detailed exploration. The prediction of the total traffic volume at the initial stage is based on the entity model of statistical methods. For example, Ahmed et al. Clearly put forward the common method of solving the differential autoregressive moving average entity model (ARIMA), which is also the first time that this optimization algorithm was used in the transportation industry in ancient history. ARIMA is a method of integrating autoregressive (AR) and mobile assessment (MA) models. The optimization algorithm model can predict the traffic flow and traffic flow rate, providing a strong application for traffic control. On the premise of the classical model, non parametric statistical models, such as Kalman filter model and wavelet basic theory model, as well as neural network models, such as SVM algorithm (SVM) and K-NN (K-NN), are given, and they are used for traffic trip prediction. Compared with this, the neural network model is more available and accurate than other traditional machine learning algorithms. LSTM models can effectively deal with the problem that the gradient direction is too fast and disappears in the information link of long-term prediction of long-term coded traffic flow data, thus improving the accuracy and precision of prediction.

Many relevant research workers have also joined this trend, and many research practices have shown that this is a new research method compared with BP neural

network model. Compared with the speed and accuracy of traditional machine learning algorithms, LSTM algorithm is more outstanding.

Although traditional machine learning algorithms can basically be used in tasks such as sorting, classification and regression, some deep neural network algorithms are particularly obvious in dealing with problems with higher accuracy and relatively fast speed. For some problems in the above data prediction, this paper compares XGBoost algorithm with LSTM algorithm to improve the traditional machine learning algorithm. This work starts with the data processing method, the missing value filling and filtering. Then the model fitting is practiced to obtain the best solid model, and then the two algorithms are compared to further verify the advantages of one algorithm.

4 Make Plans and Design Experiments

First, prepare the obtained data, and select the cleaned data for feature analysis and selection, and select the features that have a great impact on the predicted experimental conclusions to generate a sample of this experiment. Select the cleaned data to carry out feature selection, and select the features that have certain influence on the prediction experiment conclusion [8], so as to form a sample of this experiment. Later, XGBoost solid model is used for practice. After parameter adjustment of the algorithm model, the best model is selected, and the features such as weather and temperature are input. The algorithm model is trained according to these features to predict the traffic flow data. Different algorithms are imported into the same data set to obtain different prediction results. After the introduction of LSTM algorithm, training and prediction can also be carried out, and the optimal value is selected by comparing the results.

Data set of Metro-Interstate-Traffic-Volume-Encoded was selected for this experiment. The data set has 11 features, including: holiday, temp, rain_1h, snow_1h, clouds_all, weather_main, weather_description, date_time, traffic_volume. At first, the data is not directly available, but must be encoded relative to its data. In this experiment, the data is coded and the top ten features are input to predict the characteristics of traffic flow.

5 Conduct Experiment

The experimental process is shown in Fig. 1. The general process of the experiment is to collect data and then process the data and then select the features in the data. After obtaining the appropriate characteristics of the sample, the data set is divided into the training set and the test set. This part is divided into two steps: one step is to train the model and predict the model; the other step is to do grid search and cross validation to get the optimal combination of parameters and then finally to predict the traffic flow.

In this experiment, a total of 48,205 data were selected, and finally, 9,641 data were allocated as test samples and 38,564 data were allocated as training samples according to the proportion of the model and imported into the algorithm model respectively.

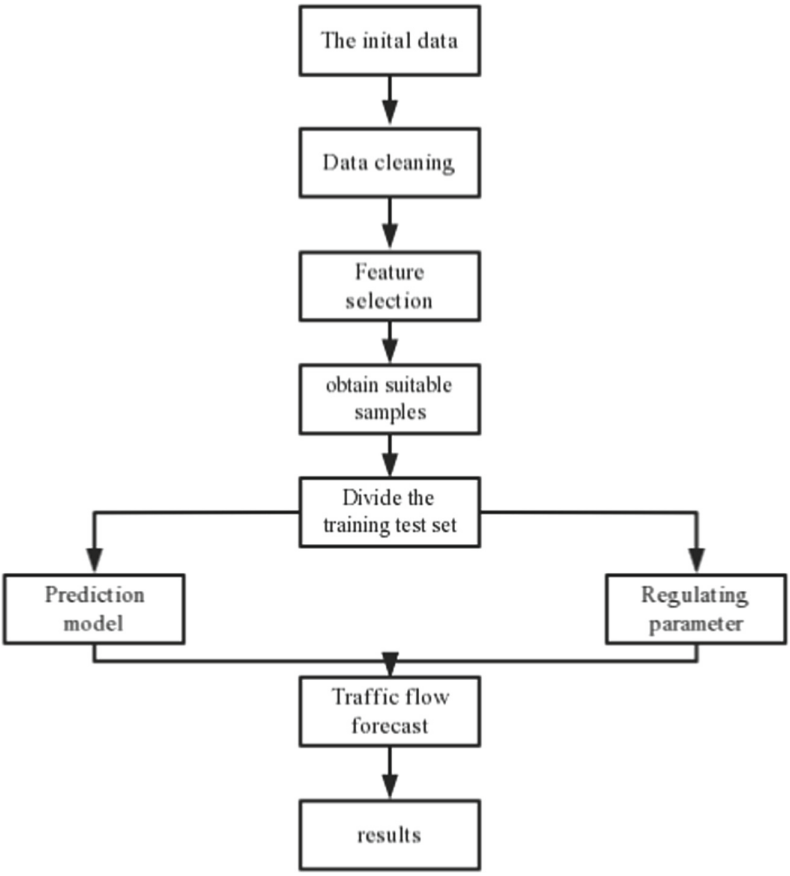

Fig. 1. Experimental flow chart

6 Experimental Results and Analysis

Figure 2 shows the comparison between the predicted value and the real value of traffic flow mentioned in this paper. In the figure, the red curve represents the real traffic flow data, while the blue curve represents the predicted traffic flow value. We randomly select 100 pieces of data for empirical prediction drawing. We can see from the figure that the predicted results are similar in all cases, which are of certain reference value.

Shown in Table 1: For the successful prediction of the data set of traffic flow, for the XGBoost to divide the training set and the test set, the prediction result can be directly obtained and compared with the real value.

For the LSTM algorithm model, the hidden layer is an important structure of the model. Increasing the number of hidden layers can reduce the prediction error and improve the accuracy of the prediction [9]. In theory, the more layers of the model, the better the training result will be, and even the prediction accuracy can reach 100%. However, when the prediction model is applied to the actual data, the model will appear

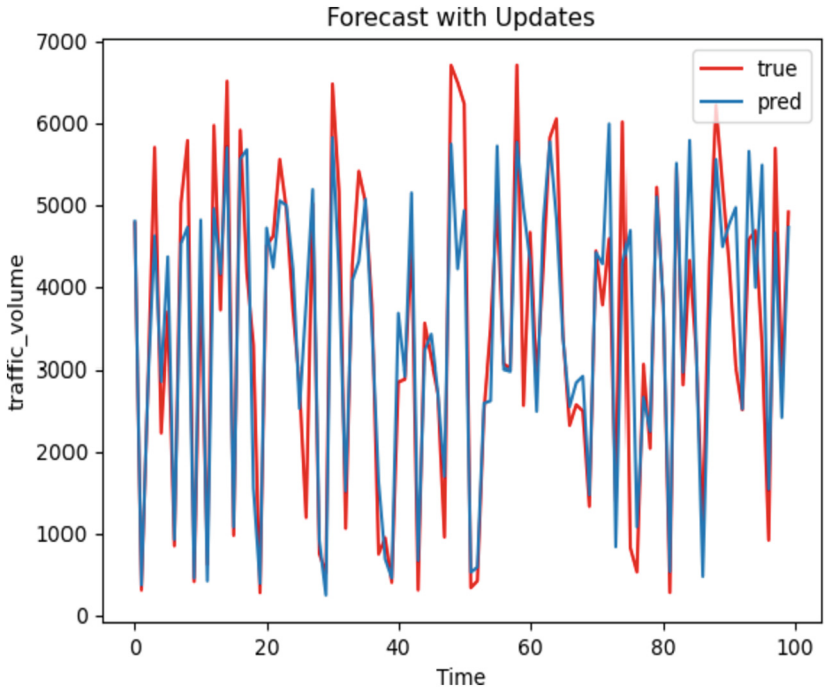

Fig. 2. True and predicted values (Color figure online)

over-fitting, which is also a point to be improved, and the prediction result will also be seriously reduced. Therefore, the training model will find the appropriate parameter model from the training, and then conduct the training and predict the most total results [10].

Table 1. Comparison of algorithm results

	XGBoost	LSTM
MSE	681823.16	335289.82
RMSE	825.72	595.76
R-Squared	0.825	0.936

7 Conclusion

Traffic flow prediction is always an important problem in intelligent manufacturing systems. In this paper, XGBoost algorithm and LSTM algorithm are applied to predict the traffic flow data set. The results of MSE, RMSE, and R-Squared are better than those

of XGBoost. In addition, we also compared the prediction performance of this method with the traffic flow information, and verified the advantages of this method.

The experimental results show that both XGBoost and LSTM can predict traffic flow data. The result of LSTM's high degree of data fitting is better than that of XGBoost algorithm. Through the comparison of table data, it can also be seen that LSTM algorithm also needs to be improved. The optimization of the four super parameters proposed by LSTM algorithm, namely the number of hidden layers, the number of hidden layer units, the learning rate and the exit rate, although both algorithms can predict the data, the input features are relatively few. The actual situation needs to be improved. Complete more scientific and perfect traffic flow prediction.

References

1. Sun, B., Geng, R., Zhang, L., Li, S., Shen, T., Ma, L.: Securing 6G-enabled IoT/IoV networks by machine learning and data fusion. EURASIP J. Wirel. Commun. Netw. **2022**(113), 1–17 (2022)
2. Sun, B., Geng, R., Xu, Y., Shen, T.: Prediction of emergency mobility under diverse IoT availability. EAI Endorsed Trans. Pervasive Health Technol., 1–9 (2022)
3. Kim, J.-G., Park, J.: A Study on the security technology of the IoT (Internet of Things). J. Korea Soc. Inf. Technol. Policy Manag. **9**(6), 571–575 (2017)
4. Dong, Y.L., et al.: Nonlinear load harmonic prediction method based on power distribution Internet of Things. Sci. Progr. **2021 (**2021)
5. Li, X., Pak, C., Bi, K.X.: Analysis of the development trends and innovation characteristics of Internet of Things technology - based on patentometrics and bibliometrics. Technol. Anal. Strateg. Manag. **32**(1), 104–118 (2020)
6. Sun, B., Geng, R., Shen, T., Xu, Y., Bi, S.: Dynamic emergency transit forecasting with IoT sequential data. Mob. Netw. Appl. (MoNet), 1–15 (2022)
7. Sun, B., Ma, L., Shen, T., Geng, R., Zhou, Y., Tian, Y.: A robust data-driven method for mutiseasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**(6692390), 1–11 (2021)
8. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Tech. Vjesn. **24**(5), 1597–1607 (2017)
9. Shao, H., Li, W., Cai, B., Wan, J., Xiao, Y., Yan, S.: Dual-threshold attention-guided GAN and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation. IEEE Trans. Ind. Inform. (2022)
10. Yan, S., Shao, H., Xiao, Y., Liu, B., Wan, J.: Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises. Robot. Comput. Integr. Manuf. **79**, 102441 (2023)