Mobile network traffic prediction based on machine learning

Huike Shi

Beijing University of Posts and Telecommunications

shk79631@163.com

Abstract. In order to better cope with the overall network efficiency and energy consumption caused by the tide phenomenon of the network traffic of the base station and the physical capacity expansion caused by the increasing network traffic demand, we need to predict the network traffic of the base station in real time, so as to guide the design of the time-sharing switching program of the base station and provide suggestions for future planning and construction. In this paper, ARIMA model and LSTM model are used to predict the base station traffic respectively, and RMSE is used as the model evaluation index. The experimental results show that the deviation RMSE predicted by ARIMA model is 1.904, and the deviation RMSE predicted by LSTM model is 1.993. Therefore, ARIMA model predicts more accurately, that is, it performs better in predicting the base station traffic data of ARIMA model.

Keywords: base station traffic prediction; Time series prediction; ARIMA model; LSTM model

1 Introduction

With the progress of the times and the development of society, mobile network communication technology is also constantly iterating. Today, when 4G is widely used, the construction of 5G network has also been spread out in a large area, and has affected all aspects of people's life. With the continuous development of mobile Internet technology, the data traffic in China is also growing explosively, followed by the traffic load of the base station. In the situation of increasingly fierce competition among operators, the construction of 5G also brings great cost pressure. At this time, on the premise of ensuring the user's network business experience and service quality, reducing the energy consumption of wireless communication systems becomes the key. Due to the tidal phenomenon of the base station, the number of users will be greatly reduced in some periods. The tide phenomenon makes the BTS traffic show peaks and troughs in different periods. If the BTS is configured according to the number of carrier frequencies in the low capacity period in the high traffic stage, it will bring users a bad Internet experience; If the base station configuration is operated according to the number of carrier frequencies in the high-capacity period in the low traffic stage, a large amount of unnecessary resources will be wasted. Therefore, we need to establish

H. Mallick et al. (Eds.): ICEMCI 2022, AEBMR 231, pp. 1691–1698, 2023.
https://doi.org/10.2991/978-94-6463-098-5_191

an effective prediction model for the network traffic of the base station, predict the network traffic of the base station in real time, set the automatic switching procedure of the base station based on the prediction results, improve the overall efficiency of the network, and guide operators to better plan and design physical expansion.

2 Research on base station network traffic

2.1 Analysis of network traffic data characteristics

The network traffic shows some unique distribution characteristics with the change of time. Previous studies have shown that the distribution characteristics of network traffic data with time are mainly divided into self similarity, burst, non-stationary and periodicity. Self similarity was first clearly proposed by Leland and others in the early 1990s, which refers to the similarity between the whole and parts of a complex system, and the fine structure or properties between this part and that part, or that the part taken from the whole can reflect the basic characteristics of the whole. Burstiness and non-stationary mean that at a certain point in time, the network traffic will suddenly decrease or increase without warning. Periodicity means that the network traffic of the base station will show a periodic change with the change of time. This periodic change also reflects some network behavior data of the user, and the behavior habits of the user always have a certain regularity. At the same time, the collection of network traffic data is also based on the hourly cycle, Therefore, the network traffic of the base station will also exhibit periodic characteristics.

2.2 Research status

Based on the analysis of the characteristics of network traffic data, the network traffic data has a strong correlation with time, which is time series data. Therefore, the time series data prediction method can be used to predict the long-term and short-term development trends of network traffic changes. According to the types of prediction models, prediction methods can be divided into traditional statistical model prediction, neural network model prediction and Support Vector Machine (SVM).

At present, the research on base station network traffic prediction mainly focuses on the time series prediction model. Traditional time series modeling was first used in time series prediction. Literature [1] predicts the short-term traffic behavior trend of LAN based on ARMA model. The linear prediction model can not well meet the characteristics of network traffic burst and non-stationary, and is more suitable for short-term prediction. Therefore, when using the linear prediction model for time series prediction, EMD / EEMD decomposition can be used to balance the characteristics of non-stationary. Reference [2] proposed a real-time network clustering model based on EMD clustering, which has better prediction effect compared with ARIMA model and EMD-ARMA model. With the increasing popularity of machine learning, more and more machine learning models have been applied to the field of time series prediction, which has an irreplaceable position in the research. Literature [3] uses lasso algorithm to compress variables and establishes XGBoost model to predict mobile network traffic. After successful applications in NLP, CV and other fields, deep learning models have also been gradually introduced to solve time series prediction problems, improving the efficiency of solving large-scale data. Literature [4] uses LSTM neural network to predict base station traffic, and compares it with ARIMA model. The model has stronger fitting effect and better prediction performance. According to the characteristics of network traffic, document [5] proposed a TVF-EMD-LSTM network traffic prediction model composed of TVF-EMD algorithm and LSTM network. Reference [6] proposed a wavelet neural network prediction model based on particle swarm optimization to predict mobile network traffic. Reference [7] uses the improved wavelet neural network model to dynamically predict the load flow of the base station. Document [8] proposed a model based on Prophet and EMD / EEMD joint Prophet to predict the network traffic of the base station cell.

In terms of base station network traffic prediction, this paper uses ARIMA model and LSTM model to predict base station traffic respectively, analyzes and compares the two prediction models, and selects the model with higher accuracy and better prediction effect to guide operators to better plan and build base stations, improve the overall network efficiency and reduce network energy consumption.

3 Base station network traffic prediction

3.1 Data set introduction

The data used in this paper comes from the hourly traffic data of the base station cell from March 1 to April 19, 2018 provided by MathorCup university mathematical modeling challenge. The traffic data is collected in hours, and the unit is GB. In this paper, 1152 traffic data of a cell are randomly selected as the data for model training and testing, of which 80% of the data is the training set and 20% of the data is the test set.

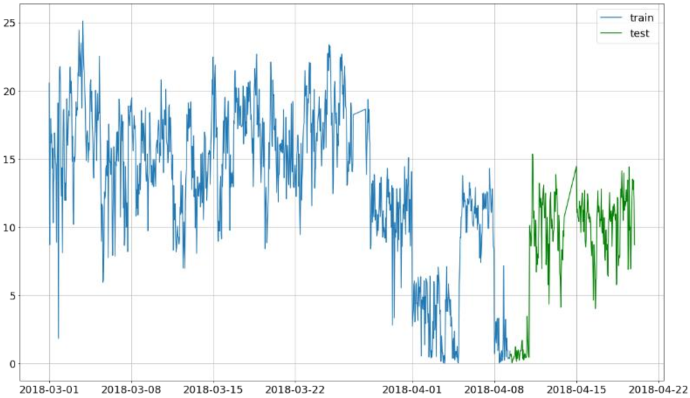

Fig. 1. BTS network traffic data set (Photo credit: Original)

3.2 Data preprocessing

Data stationarity analysis.

Before the ARIMA model is used to predict the traffic of the base station cell, the stationarity of the time series must be verified first. In this paper, ADF test is used to test the stationarity of flow data.

ADF test, also known as unit root test, is to determine whether there is a unit root in a sequence. If the sequence is stationary, there is no unit root; Otherwise, there will be unit roots.

It is assumed that H0 has a unit root, that is, the time series is a non-stationary series. If the obtained significance test statistics are less than three confidence levels (10%, 5%, 1%), then there is (90%, 95%, 99%) confidence to reject the original hypothesis.

According to the ADF test, the p value is 0.3119, which was significantly greater than the significance level α (0.05), accept the original assumption, and explain that the time series data of the network traffic of the base station is a non-stationary series.

Normalization.

In this paper, the value range of each feature is normalized to [0,1] through the method of maximum and minimum value normalization. For each one-dimensional feature $x_i, i = 1,2,3 \ldots n$, the feature value x_{ij} of the jth sample, $j = 1,2,3 \ldots m$, after normalization:

$$x_{ij}^{*} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{1}$$

The results are mapped between [0,1], where x_j^{max} and x_j^{min} are the maximum and minimum values of feature x_i on all samples, respectively.

3.3 Model evaluation index

In this paper, the root mean square error (RMSE) is used as an index to evaluate the prediction accuracy of the model, and the formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2} \tag{2}$$

Where y_i is the actual value of the ith sample and $f(x_i)$ is the predicted value of the ith sample. The smaller the value of $RMSE$, the better the prediction effect of the model.

3.4 Network traffic prediction based on ARIMA model

Model introduction.

ARIMA model (Auto Regressive Integrated Moving Average model), that is, differential autoregressive moving average model, also known as autoregressive moving average model, combines Autoregressive model (AR), Moving Average model (MA) and Autoregressive Integrated Moving Average model ARIMA (p, d, q), which is one of the time series prediction and analysis methods. In ARIMA (p, d, q), AR is "auto-regressive", p is the number of autoregressive terms, MA is the "moving average", q is the number of moving average terms, and d is the number of differences (orders) made to make it a stationary sequence. When the training data is a non-stationary sequence, it is necessary to use the difference method to turn the original sequence into a stationary sequence. "Difference" does not appear in the English name of ARIMA, but it is a key step in the realization of ARIMA model.

AR (Auto Regression) model, that is, autoregressive model, describes the relationship between the current value and the historical value, and uses the historical time data of the variable itself as an independent variable to predict the data at its future time point. General p-order autoregressive model AR:

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_p x_{n-p} + u_n \tag{3}$$

MA (Moving Average) model in AR model, if u_n is not a white noise, it is usually considered as a q-order moving average. Namely

$$u_n = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q} \tag{4}$$

Among them, ε_n represents a white noise sequence. When $x_n = u_n$, that is, the current value of the time series has no relationship with the historical value, but only depends on the linear combination of the historical white noise, and the MA model is obtained:

$$x_n = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q} \tag{5}$$

Analysis of experimental results.

In this section, ARIMA model is used to predict the traffic of the BTS cell. The first 80% of the data is used as the training set, and the last 20% of the data is used as the test set.

According to the previous analysis, this time series is a non-stationary series, so it is necessary to convert the data into stable data through difference, and then regress the dependent variable to its lag value and the present value and lag value of the random error term.

According to the ADF test of the difference sequence, the p value is 1.984e-17, which is less than the significance level α (0.05), thus rejecting the original assumption. It indicates that the time series data of the base station network traffic is a stationary series after the first-order difference. Therefore, the value of d in the ARIMA (p, d, q) model is 1.

After d is determined, the values of p and q are determined to be 4 and 3 respectively according to the autocorrelation diagram and partial autocorrelation diagram of the first-order difference sequence.

After the ARIMA (p, d, q) model is built, the model is trained. Fig. 2 is a comparison diagram of the real value and the predicted value of the test set data, in which the RMSE value of the test set data is 1.904.

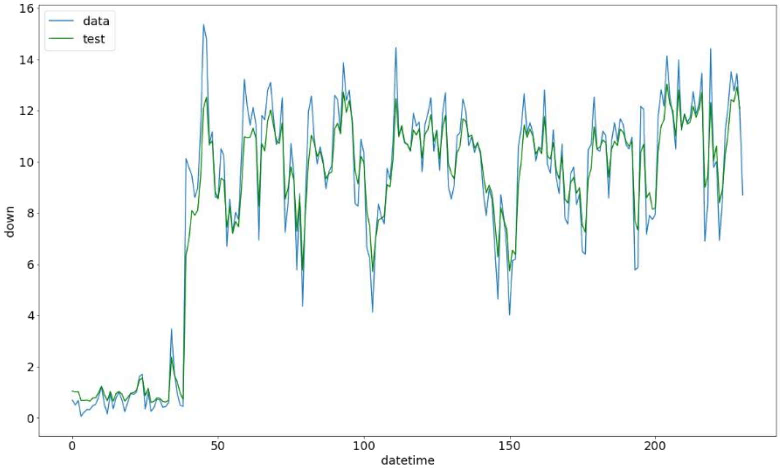

Fig. 2. Comparison between real value and predicted value (Photo credit: Original)

3.5 Network traffic prediction based on LSTM model

Model introduction.

LSTM network is a special RNN model, which is proposed to solve the gradient dispersion problem of RNN model.

The core of LSTM is "cell state". In LSTM, specially designed "Gates" are used to introduce or remove information in cell state. LSTM mainly includes three different gate structures: forgetting gate, memory gate and output gate. These three gates are used to control the infomation retention and transfer of the LSTM, and are finally reflected in the cell state and the output signal h_t.

The forgetting gate is composed of a $sigmoid$ neural network layer and a bitwise multiplication operation. The output signal of the forgetting gate is:

$$f_t=\sigma\left(W_f * [h_{t-1}, x_t] + b_f\right) \tag{6}$$

W_f and b_f are neural network parameters, x_t is the input signal at time t, h_{t-1} is the last output signal of LSTM.

The memory gate includes $sigmoid$ neural network layer and $tanh$ neural network layer. The output of $sigmoid$ neural network layer is:

$$i_t=\sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{7}$$

The output of $tanh$ neural network layer is:

$$C_t^*=tanh(W_c * [h_{t-1}, x_t] + b_c) \tag{8}$$

Update the cell state at time t according to the forgetting gate and the memory gate:

$$C_t = f_t * C_{t-1} + i_t * C_{t-1}^* \tag{9}$$

The output gate integrates the cell state at time $t-1$ with the output signal h_{t-1} at time $t-1$ and the input signal x_t at time t as the output signal at the current time:

$$o_t=\sigma(W_o*[h_{t-1},x_t]+b_o) \tag{10}$$

$$h_t=o_t*tanh(C_t) \tag{11}$$

Analysis of experimental results.

In this section, the LSTM model is used to predict the traffic of the base station cell, and the normalized data is divided into a training set and a test set. The first 80% of the data is used as the training set, and the last 20% of the data is used as the test set. The LSTM model used in this paper includes three layers of LSTM hidden layer and one layer of fully connected output layer.

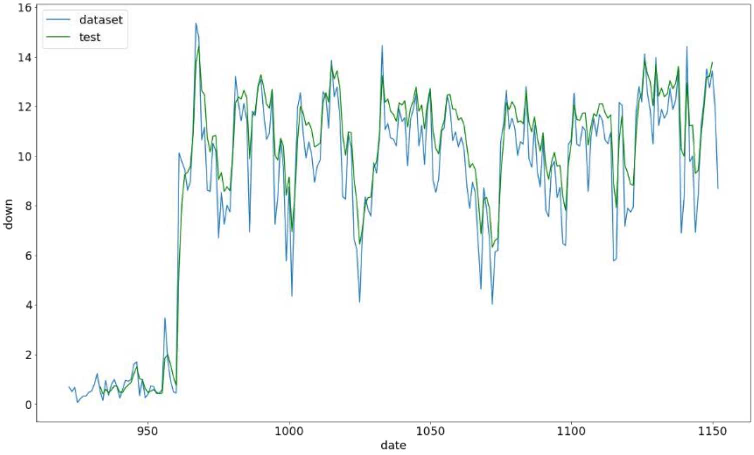

Fig. 3. Comparison between real value and predicted value (Photo credit: Original)

After the LSTM model was built, the model was trained. the network traffic in the 13th hour is predicted according to the traffic data in the previous 12 hours. Fig.3 shows the fitting effect of the model on the test set, where the RMSE value of the test set data is 1.993.

4 Conclusion

In this paper, we use ARIMA model and LSTM model to predict the traffic data of the base station. The results of the experiments show that the RMSE value of the ARIMA model test set is smaller. It indicates that the ARIMA model has a smaller deviation and performs better in the prediction of the traffic data of the base station. It provides a prediction method to guide the time-sharing switching program of the base station in the short term and to better plan and design the physical expansion in the long term.

References

1. Yiyong Lin. Application of ARMA model in LAN short-term traffic prediction [J]. Computer engineering and applications, 2017,53 (S2):88-91.
2. Lishuang Yao, Real-time Network Traffic prediction model based on EMD clustering [J]. Computer science, 2020,47 (11A):316-320.
3. Shaojie Liu. Research on Mobile Network Traffic Prediction based on Big data [D]. North China University of Technology, 2021.
4. Jianwei Hu. Traffic prediction of mobile communication base Station based on LSTM [J]. Information Technology and Informatization, 2021(5):84-86.
5. Jiapeng Ren. Research on Traffic Prediction and Base Station Sleep Based on Machine Learning [D]. Jilin University, 2020.
6. Chenyu Pan. Prediction of mobile network traffic and 5G users [D]. Beijing Jiaotong University, 2021.
7. Xiaoshuang Sun. Base Station Dormancy in Heterogeneous Networks: A Method Based on Traffic Prediction [J]. Electronic Technology and Software Engineering, 2017(10):27-29.
8. Jiachen Zhang. Design and Implementation of wireless Network Traffic Prediction method for base Station cell [D]. Beijing University of Posts and Telecommunications, 2021.