Traffic Arrival Prediction for WiFi Network: A Machine Learning Approach

Ning Wang, Bo Li, Mao Yang(✉), Zhongjiang Yan, and Ding Wang

School of Electronics and Information,
Northwestern Polytechnical University, Xi'an, China
ningwang22668800@mail.nwpu.edu.cn, {libo.npu,yangmao,zhjyan}@nwpu.edu.cn

Abstract. At present, Wi-Fi plays a very important role in the fields of online media, daily life, industry, military and etc.

Exactly predicting the traffic arrival time is quite useful for WiFi since the access point (AP) could efficiently schedule uplink transmission. Thus, this paper proposes a machine learning-based traffic arrival prediction method by using random forest regression algorithm. The results show that the prediction accuracy of this model is about 95%, significantly outperforming the linear prediction flow. Through prediction, resources can be reserved in advance for the arrival of data traffic, and the channel can be optimally configured, thereby achieving better fluency of the device and smoothness of the network.

Keywords: Wi-Fi Network · Artificial intelligence · Big data · Machine learning · Random forest · Regression

1 Introduction

1.1 Current Status of WiFi and Artificial Intelligence

Wireless communication and network development are fast, and WiFi is one of the most important data service bearers. As one of the most important carrying methods of wireless communication services, WiFi has the following characteristics:

The main advantage of WiFi is that it is wireless, so it's not constrained by the wired environment. Therefore, it is very suitable for mobile office users and has broad application prospects.

The transmission power specified by IEEE802.11 cannot exceed 100 mW, and the actual transmit power is about 60–70 mW, which means that WiFi power is very low, it is healthy and safe to use.

To set up a wireless network, we just need a wireless network card and an AP, so that it can be combined with the existing wired architecture in a wireless mode. Sharing network resources, erection costs and complex procedures are far lower than traditional wired networks.

B. Li et al. (Eds.): IoTaaS 2019, LNICST 316, pp. 480–488, 2020.
https://doi.org/10.1007/978-3-030-44751-9_40

WIFI technology as a supplement to high-speed wired access technology, can transmit very fast and it is cheep, WIFI technology is broadly used in wired access wireless Extended field.

At present, WiFi has been widely developed and applied in various fields such as daily life, industrial development, and military development. It brings convenience to people's lives, contributes to the development of science and technology, and provides assistance for military development and industrial progress. The business volume is bound to increase gradually. In order to meet the growing needs of network development, it is imperative to update and develop the WiFi technology. The academic and industrial circles are paying attention to the key technology research and standardization promotion of the next generation WiFi.

1.2 One of the Existing Problems

Due to the characteristics of distributed random access, WiFi networks may cause collisions between data packets and interference between cells and cells. With the development of WiFi technology, the collision and interference will become more serious in the high-density deployment scenarios of the next-generation WiFi, which seriously suppresses the performance of the WiFi system (Quality of Systems, QoS) and user experience quality (Quality of Experiences, QoE).

If we can accurately estimate the arrival time of the next packet of the service based on the known arrival characteristics of the service, such as the packet length, transmission time, packet interval, and packet arrival time of the service data packet, Targeted QoS guarantees can significantly improve QoE.

1.3 Existing Literature Research

Intelligentization is a research hotspot in recent years. In particular, machine learning and deep learning have rapidly penetrated into various fields and achieved extremely beneficial effects. They have brought powerful driving forces to various industries including communications and networking. The existing research on machine learning in business arrival prediction is as follows:

Wang et al. [1] focus on the application of MLN and summarize the basic workflow for explaining how to apply machine learning technology in the network field.

Jiang et al. [2] use data-driven video quality prediction to make the best decisions to improve Internet Video Quality of Experience (QoE) through Key Feature Analysis (CFA) design and implementation.

Fadlullah et al. [3] solves the application of deep learning in network traffic control system, and points out the necessity of investigating the decentralized work of deep learning applications in various network traffic control.

Kato et al. [4] presents the appropriate input and output characteristics of heterogeneous network traffic, and describe how the modified system works and its difference from traditional neural networks.

Mao et al. [5] build systems that learn to manage resources directly from experience. They offer DeepRM, a sample solution that turns packaged task issues with multiple resource requirements into learning problems.

The rest of the articles providing ideas in the creation process of this paper are listed in the references.

However, unfortunately, the existing literature has less predictions on network traffic, and a small number of predictions of network traffic do not capture data in specific applications, and use machine learning to predict the arrival of data packets.

1.4 Work Done in This Article

In order to study and improve the QoS and QoE of WiFi in different scenarios, this paper is based on different scenarios (including scenarios with better service quality such as school library, service quality, such as school dormitory, poor service quality such as school).

The business class obtained by the canteen knows the arrival characteristics of the data packet, extracts the relevant time information, uses the machine learning method to learn and model, and predicts the arrival time of the next data packet. Found verified by simulation and real data acquired data sets, the verification test of the accuracy of the output, provided linear prediction model are compared, the results show that: under different scenarios, the established model to study the time of arrival of the data packet are better.

1.5 Article Chapter Structure

Section 1 introduces the clarification of the problem and outlines the work done in this paper.

Section 2 introduces the related techniques and resources used in this experiment.

Section 3 introduces the prediction model and accuracy of the packet arrival time based on the random forest algorithm.

Section 4 is the analysis of the experimental content, simulation prediction and results, and finally the discussion, summary and outlook.

2 Traffic Arrival Prediction Scheme

2.1 Related Introduction

In the process of obtaining data, this paper uses Huawei's packet capture interface based on test sample, and uses adb program to access the mobile phone to obtain data. In the initial information processing, Notepad++ and Wireshark software were used to extract the quintuple information and capture the stream.

In this paper, the Anaconda platform is used in the process of data processing and model building. The function library greatly reduce the lack of library support in the Python development process. It is a very rich and powerful Python development platform.

Scikit-learn is a Python library, known as Sklearn, which is widely used to solve regression problems.

2.2 Introduction of Core Ideas

The idea of the implementation is to apply the corresponding software to capture and extract the data stream, obtain the time data, use the random forest regression algorithm to model the data, and then use the established model to predict the time interval of the next packet arrival.

The random forest modeling process is Fig. 1 as follows:

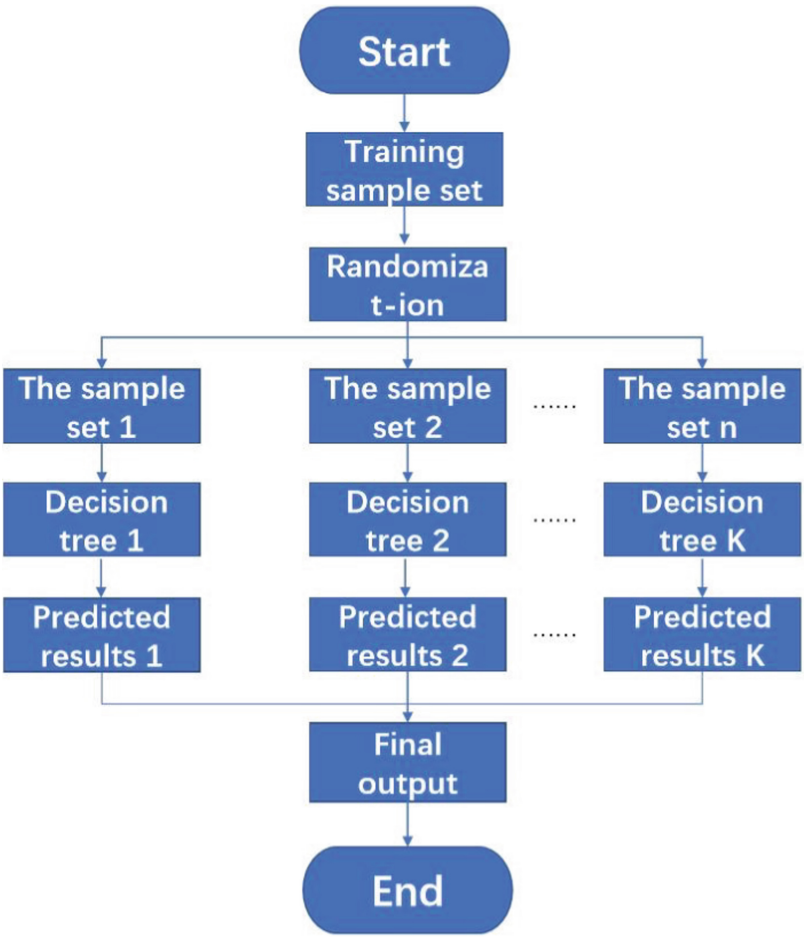

Fig. 1. The structure of the RF model

2.3 Experimental Planning and Data Collection

Under different scenarios, such as library, canteen, dormitory, the software was measured respectively, each case to ensure sufficient number of experiments, in order to get plenty of experimental data, using the control variable method, in order to better control network signal coherent condition, can use mobile phone hot manner, collection and network transmission in the process of the log information, and obtain the corresponding pcap file, get the corresponding files, through the "five yuan group" flow analysis as characteristics were caught, then for each packet received time information, and use data to establish model, to forecast the next packet arrival time.

2.4 Introduction to Data Sets

For example, in the data obtained during Mobile Legends, each game records a time (game start time) when opening the interface, records a time (opening time) after the official opening, and records a time after the game ends. Through the obtained log information and pcap file, corresponding to the start time and end time, the data is extracted, the time data of the received packet is obtained, and the data is processed to obtain a data set required for modeling.

3 Machine Learning Process

3.1 Random Forest Regression Algorithm

In this paper, we use the random forest regression algorithm to model the data set and simulate the test processing. This paper chooses the random forest regression algorithm to do the experiment based on the following reasons:

(1) There may be a potential correlation between the reception times of different packets of the same stream, but these correlations are difficult to measure correctly, so algorithms that are sensitive to multicollinearity between features are not applicable. The random forest algorithm is not sensitive to the correlation between features, nor does it need to select features. It is very suitable for this regression experiment.
(2) The random forest algorithm is very robust and relatively insensitive to discrete data points. Due to the certain interference of the received time information, it is inevitable that there will be noise data. The random forest algorithm can effectively avoid the influence of these data on the final model.
(3) The random forest algorithm aggregates a large number of classification trees, which can improve the model's prediction accuracy, and the random forest algorithm has a fast calculation speed, so the speed performance is excellent when the amount of data is large.

Random forest regression algorithm: Depend on the function of the random trees, random forests can be applied to classification and regression problems. For the

regression algorithm: The cart tree is a regression tree, and the principle adopted is to gain the minimum mean square error. That is, for the arbitrary division feature F, the data sets N1 and N2 are divided into two parts according to any partition point v, to obtain the feature that makes the mean square error of each group of data N1 and N2 reach the minimum, and the sum of the mean square error of N1 and N2.

The expression is:

$$\min_{F,v}\left[\min_{m_1}\sum_{x_i\in N_1(F,v)}(y_i-m_1)^2+\min_{m_2}\sum_{x_i\in N_2(F,v)}(y_i-m_2)^2\right] \tag{1}$$

Where m1 is the sample from the N1 data set and m2 is the sample from the N2 data set.

The cart tree is predicted based on the mean of the leaf nodes, so the prediction of the forest is the average of all the predicted results of trees.

3.2 Data Preprocessing

The log information of the obtained network stream is captured, and then the log information is captured by using Notepad++ and Wireshark, and the packet length and the receiving time information of each data packet are obtained, and then the time data is normalized by programming. Make the size of the data in the range of [0, 1]. In order to satisfy the requirement that the model predicts the arrival time of the next packet according to the time of every four packets, the data is processed into a txt file of five data per line. The data set is output in txt for use. The time information is extracted by programming, and then the time information is extracted according to the obtained data set file lenandtime.txt, a total of 8406, and then the data is cleaned, in order to facilitate observation and processing, balance the influence of each input feature value, need The raw data is normalized to become the value of the [0, 1] interval, and the normalization formula is as follows:

$$X=\frac{x_i-min(x)}{max(x)-min(x)} \tag{2}$$

Then, through processing, a data set of five time data per line required for model training is obtained. The data set thus obtained meets the experimental requirements, and then the model training can be performed.

3.3 Model Training

Regression algorithm to model, put the data set into the model, scramble the data, use 70% of the data after the disruption as training data, and test the remaining 30%. The data is tested on the model. After training the model, you can test the data set, test the model, and visualize the test results, select the appropriate amount to evaluate the model, and ensure the readability of the results.

4 Performance Evaluation

4.1 Comparison and Analysis of Training Effects

Random forest regression model validation. Validation models need to pass appropriate evaluation indicators. The mean square error (MSE), that is, the (true value - predicted value) and then squared and then summed and then averaged, obviously, in the process of regression model prediction, the smaller the value of MSE, the higher the accuracy of the model. The formula is as follows:

$$MSE = C1m \sum_{i=1}^{m} (y_i - \hat{y})^2 \tag{3}$$

The goodness of fit R-Squared can test the fitting degree of the regression model to the sample data, the value is between 0 and 1. The higher the goodness of fit R-Squared, the higher the interpretability of the representative model:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{4}$$

Where: y is the actual observation, i.e. the time in the measured data; $\hat{y}$ is the predicted time value of the model; $\overline{y}$ is the average of the actual time data; m, n is the number of the samples.

Molecules are all the errors predicted by the models we train. The denominator does not consider anything else. The result we predict is the average of y. If the result is 0, then our model, the accuracy is quite poor, similar to the guess. If the result is 1, it means that our model has no errors. If the result is a number between 0–1, which is the degree of our model, the closer the value is to 1, the better the model. If the result is negative, it means that our model is not as good as guessing.

In the simulation prediction of the model, 70% of the data is used for modeling, and the remaining 30% of the data is used to train and test the model.

The predicted results were evaluated using MSE (mean square error) and R-Squared: The final model predicts that the MSE tends to be stable when the decision tree is above 31, and the MSE is on the e-6 scale. The result is very small. R-Squared is 0.98 and above, which is very close to 1. Explain that the model works well.

The result is shown in Fig. 2 as follows:
The results of multivariate linear fitting through matlab are:

$$y = 0.3119 - 0.2000 * x_1 - 0.4000 * x_2 - 0.6000 * x_3 + 0.8000 * x_4 \tag{5}$$

Comparison of the multivariate linear fit with the R-Squared results of the model in this paper is shown in Table 1.

Table 1. Model comparison

	Model of this paper	Linear fitting model
The amount of data	8401	8401
R-Squared	0.98	0.40

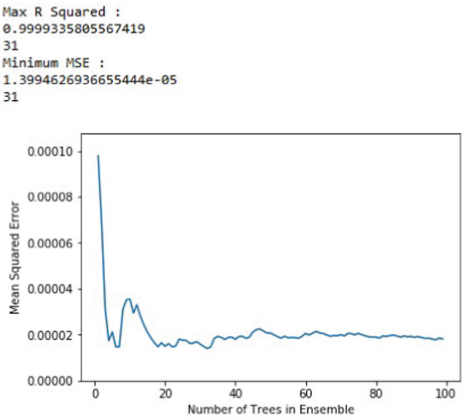

Fig. 2. Result

5 Discussion

According to the experimental data in the table, the modeling results in this paper are significantly better than the linear fitting results of Matlab, so the model built in this paper has practical significance in application.

6 Conclusions and Future Works

With the development of science and technology and the growing demand of people, WiFi will certainly develop in a more advanced, faster and more general direction in the future, and artificial intelligence will also play an important role in promoting the development of WiFi. In the future study and research life, I will pay more attention to the combination of the two, in order to make my own contribution to the development of WiFi.

Acknowledgement. This work was supported in part by the National Natural Science Foundations of CHINA (Grant No. 61771390, No. 61871322, No. 61771392, No. 61271279, and No. 61501373), the National Science and Technology Major Project (Grant No. 2016ZX03001018-004), and Science and Technology on Avionics Integration Laboratory (20185553035).

References

1. Wang, M., et al.: Machine learning for networking: workflow, advances and opportunities. IEEE Netw. **32**(2), 92–99 (2018)
2. Jiang, J., et al.: CFA: a practical prediction system for video QoE optimization. In: Usenix Conference on Networked Systems Design and Implementation USENIX Association (2016)
3. Fadlullah, Z., et al.: State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun. Surv. Tutorials **19**, 2432–2455 (2017)
4. Kato, N., et al.: The deep learning vision for heterogeneous network traffic control: proposal, challenges, and future perspective. IEEE Wireless Commun. **24**, 146–153 (2017)
5. Mao, H., et al.: Resource management with deep reinforcement learning. In: The 15th ACM Workshop. ACM (2016)