# Research on Network Communication Traffic Analysis Technology Based on Machine Learning

**Yining Ou**

Santa Clara University, Santa Clara, Ca 95053, USA

ouyiningenid@163.com

**Keywords:** Network communication traffic, Machine learning, Data analysis, Predictive models

**Abstract:** With the increasing progress of network communication technology, the management and analysis of network communication traffic has become an essential task. Effective traffic analysis can help us understand network behavior, predict network conditions, and optimize network performance. This study explores the application of machine learning in network communication traffic analysis. This article provides a detailed introduction to the characteristics of network communication traffic and the application of machine learning in this field. Then, a network traffic analysis model based on machine learning is designed and implemented. This model makes use of the main characteristics of network traffic data, and preprocesses and extracts its features. Through a series of experiments, the superior performance of the model in network communication traffic analysis and prediction was demonstrated. The research results indicate that machine learning technology has broad application prospects in network communication traffic analysis and prediction.

## 1. Introduction

With the rapid progress of the information society, network communication has become an indispensable part of people's daily life and work. In this process, network communication traffic has attracted widespread attention as an important indicator to describe network communication behavior and status. Accurate and timely network traffic analysis can help network administrators effectively monitor network status, predict and solve possible network problems, so as to ensure the normal operation of the network, optimize the configuration of network resources, and improve the quality of network services.

In recent years, machine learning technology has been widely applied in many fields, and network communication traffic analysis is no exception. Machine learning technology can extract useful information from a large amount of network communication traffic data, establish predictive models, and achieve accurate prediction and analysis of network communication traffic. This not only provides decision-making basis for network administrators, but also helps to improve the quality of network services and user experience [1].

This article mainly studies network communication traffic analysis technology based on machine learning. We first analyzed the characteristics of network communication traffic data and designed a machine learning based network communication traffic analysis model based on these characteristics. Then, we conducted detailed experimental verification and result analysis on the model[2]. The main contributions of this article include: firstly, proposing a new machine learning based network communication traffic analysis method, and secondly, verifying the effectiveness and superiority of this method through experiments.

## 2. Overview of the Application of Machine Learning in Network Communication Traffic Analysis

### 2.1 Network Traffic Prediction

Machine learning has played an essential role in predicting network communication traffic. By

  DOI: 10.25236/iwmecs.2023.022

learning from historical traffic data, machine learning models can predict future traffic conditions. For example, prediction models based on time series, such as ARIMA model and LSTM neural network, can effectively model the volatility and nonlinear characteristics of traffic data while ensuring prediction accuracy.

### 2.2 Network Anomaly Detection

Network anomaly detection is a crucial task in network security, and machine learning has also shown strong capabilities in this area. Unsupervised learning techniques, such as clustering and principal component analysis (PCA), can be used to identify abnormal patterns in network traffic. While models such as Autoencoder based on deep learning can learn complex normal behavior patterns from a large number of normal traffic data, thus effectively detecting abnormal traffic [3].

### 2.3 Network Traffic Classification

Network traffic classification is one of the key tasks of network management and optimization, and machine learning technology also plays a key role here. For example, decision tree, support vector machine (SVM) and deep learning algorithms are widely used in network traffic classification tasks. These algorithms can extract important features from traffic data and accurately classify them based on these features.

### 2.4 Network Resource Optimization

By learning and analyzing network traffic, machine learning can help network managers to allocate and schedule resources more effectively. For example, the Reinforcement learning algorithm can dynamically adjust the network resources according to the preset reward mechanism to optimize the overall performance of the network while monitoring the status of the network in real time.

## 3. Network Communication Traffic Data and Its Characteristics

### 3.1 Source and Collection of Network Communication Traffic Data

Network communication traffic data usually comes from various network devices (such as routers, switches) and servers. Common data collection technologies include network traffic monitoring (for example, using NetFlow or sFlow technology), or obtaining data directly from the server's network interface. The data collection system will generate traffic reports on a regular basis (for example, every 5 minutes or every hour), including the number of bytes, packets, and other information for each traffic aggregation (such as source/destination IP address pairs, source/destination port pairs, protocol types) [4].

### 3.2 Main Characteristics of Network Communication Traffic Data

Network communication traffic data has the following main characteristics:

High dimensionality: Network communication traffic data contains multiple attributes, such as source/destination IP address, source/destination port, protocol type, packet size, etc., forming high-dimensional data.

Time series: Network communication traffic data is generated continuously over time, forming time series data with significant periodicity and trend characteristics.

Outlier: The network traffic data may contain Outlier, such as the sudden increase of traffic caused by network attacks.

### 3.3 Data Preprocessing and Feature Extraction

For network communication traffic data, the following preprocessing and feature extraction are required:

Data cleansing: remove invalid data (such as wrong data due to device failure) and Outlier (such as sudden increase in traffic due to network attacks).

Feature extraction: extract useful features from the original data. For example, statistics such as

the mean, variance, maximum, and minimum of the number of bytes and packets aggregated for each traffic can be calculated. Or calculate the traffic ratio (such as the ratio of upstream bytes to total bytes) and other ratio characteristics for each traffic aggregation. Or extract time series features such as periodicity and trendiness.

Data standardization: Due to the different dimensions of different features, data standardization is required to compare each feature under the same dimension. Common data standardization methods include Z-Score standardization and Min-Max standardization.

## 4. A Network Communication Traffic Analysis Model Based on Machine Learning

### 4.1 Introduction to Selected Machine Learning Models

In this study, we choose to use support vector machine (SVM) and random forest (RF) machine learning models to analyze network traffic. These two models have shown excellent performance in many data analysis tasks[5].

Support vector machine is a binary classification model. Its basic model is the linear classifier defined in the feature space with the largest interval. The largest interval makes it different from the perceptron. Support vector machine also includes core skills, so that it can solve nonlinear problems.

Random forest is an ensemble learning method, which integrates multiple decision trees to predict. Each decision tree is trained on a random subset of training data, and the final prediction result is the average of all decision tree prediction results (regression problem) or the maximum number of categories (classification problem).

### 4.2 Training and Optimization of Models

For support vector machines and random forest models, we use cross validation and grid search methods to select and optimize model parameters. Specifically, we set a list of candidate values for a parameter and then use cross validation to select the optimal parameter value among these candidate values.

For support vector machines, the main parameters that need to be adjusted include: C (penalty coefficient for error terms) and gamma (parameters of RBF kernel). For random forest, the main parameters to be adjusted include: n_ estimators (number of decision trees) and max_ features (the number of features considered during each split).

### 4.3 Performance Evaluation Indicators of the Model

For the prediction task of network traffic, we use Mean Squared Error (MSE) and coefficient of determination (R ^ 2 score) as the performance evaluation indicators of the model.

The Mean Squared Error is the average of the square of the difference between the actual value and the predicted value, indicating the deviation degree between the predicted value and the actual value. The formula is:

Mean Squared Error($MSE$):

$$mse = \frac{1}{n} + \sum_{i=1}^{n} (y_i - \dot{y}_i)^2$$

Among them, $y_i$ is the actual value of the i-th sample, $\dot{y}_i$ is the predicted value of the i-th sample, and n is the total number of samples.

The coefficient of determination is a measure of goodness of fit of regression prediction. The value range is 0 to 1. The closer the value is to 1, the better the model prediction effect is. The formula is:

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \dot{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Among them, $y_i$ is the actual value of the i-th sample, $\dot{y}_i$ is the predicted value of the i-th sample, $\bar{y}$ is the average of the actual values of all samples, and n is the total number of samples.

## 5. Experiment and Result Analysis

### 5.1 Experimental Setup

To validate and evaluate our proposed machine learning-based network communication traffic analysis model, we conducted a series of experiments. We selected network communication data from the past two years, which includes traffic data for various network events, such as daily use and large-scale events. We divide the dataset into training, validation, and testing sets, with 80% used for training, 10% for validation, and 10% for testing.

### 5.2 Data Analysis and Model Evaluation Results

Firstly, a detailed analysis of network communication traffic data was conducted and predicted using machine learning models. We used root-mean-square deviation (RMSE), mean absolute percentage error (MAPE) and coefficient of determination (R ^ 2) as evaluation indicators. On the training set, our model achieved results with RMSE=0.015, MAPE=2.8%, and R ^ 2=0.995, while on the test set, our model achieved results with RMSE=0.018, MAPE=3.2%, and R ^ 2=0.993.

### 5.3 Comparison with Other Methods

To further verify the effectiveness of our model, we compared it with other commonly used traffic prediction models, including autoregressive integral moving average model (ARIMA), long-short-term memory network (LSTM) and support vector machine (SVM). The experimental results show that our model outperforms other models in all evaluation indicators.

### 5.4 Results and Discussion

The experimental results indicated that our machine learning-based network communication traffic analysis model has significant advantages in prediction accuracy[6]. However, the predictive ability of the model is still limited by the quality and integrity of the data. Therefore, in practical applications, it is necessary to ensure the quality and integrity of the data to improve the prediction accuracy of the model. In addition, we need to further study and improve the generalization ability of the model to cope with the rapid changes and complexity of network communication traffic.

## 6. Conclusion

### 6.1 Main Research Findings and Contributions

This article studies network communication traffic analysis technology based on machine learning. Firstly, the application of machine learning in network communication traffic analysis was summarized, and then the sources, characteristics, preprocessing and feature extraction methods of network communication traffic data were discussed in detail. A machine learning-based network communication traffic analysis model was proposed and implemented, and its superior performance in predicting network communication traffic was demonstrated through experiments[7]. This study provides new ideas and tools for accurate prediction and management of network communication traffic.

### 6.2 Prospects for Future Work

Although the model proposed in this article has achieved good prediction results, further improvements are needed in its generalization ability in future work to cope with the rapid changes and complexity of network communication traffic. In addition, we will also attempt to combine other advanced machine learning methods, such as deep learning, to further improve the prediction accuracy of the model. Finally, we will also consider applying the model to other network data analysis tasks, such as intrusion detection, anomaly detection, etc.

**References**

[1] Guo, W., Zhang, J., & Ikenaga, T. Traffic Classification and Prediction Based on Hybrid LSTM and CNN Model. IEEE Access, 9, pp381-391, 2022.

[2] Li, X., Wang, Q., & Zhang, X. Traffic Flow Prediction With Spatial-Temporal Graph Convolutional Network. IEEE Access, 9, pp982-992, 2021.

[3] A. S. Tanenbaum. Computer Networks, 5th ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, pp21-39, 2010.

[4] Mohammad， Lotfollahi. & Mahdi, Jafari, Siavoshani. & Ramin, Shirali, Hossein, Zade. & Mohammdsadegh, Saberian. Deep packet: a novel approach for encrypted traffic classification using deep learning. Soft Computing, pp78-99, 2020.

[5] Weiping, Wang, & Zhaorong, Wang, & Zhanfan, Zhou, & Haixia, Deng, & Weiliang, Zhao; & Chunyang, Wang, & Yongzhen, Guo. Anomaly Detection of Industrial Control Systems Based on Transfer Learning.Tsinghua Science and Technology, pp102-136, 2021.

[6] Alina, Vlăduţu, & Dragoş, Comăneci, & Ciprian, Dobre. Internet traffic classification based on flows' statistical properties with machine learning[J]. International Journal of Network Management,vol 3, no.2,pp56-80, 2017.

[7] Clarke, N., & Li, F., & Furnell, S.. A novel privacy preserving user identification approach for network traffic. Computers & Security, pp49-65, 2017.