Traffic Flows Forecasting Based on Machine Learning

Vladimir Deart, Moscow Technical University of Communications and Informatics, Russia

Vladimir Mankov, Training Center Nokia, Russia*

Irina Krasnova, Moscow Technical University of Communications and Informatics, Russia

ABSTRACT

The article aims to develop a model for forecasting the characteristics of traffic flows in real-time based on the classification of applications using machine learning methods to ensure the quality of service. It is shown that the model can forecast the mean rate and frequency of packet arrival for the entire flow of each class separately. The prediction is based on information about the previous flows of this class and the first 15 packets of the active flow. Thus, the random forest regression method reduces the prediction error by approximately 1.5 times compared to the standard mean estimate for transmitted packets issued at the switch interface.

KEYWORDS

Classification, Lasso, Linear, Network Traffic, Preprocessing, Random Forest, Regression, Robust, Voting, XG Boost

1. INTRODUCTION

Every year, the traffic of telecommunications networks is growing rapidly, and the number of users and services is increasing. There are new types of applications, each of which requires ensuring the quality of service at the proper level. There are several main categories of services, such as voice, video, data, and many of their extensions: audio and video conferencing, IPTV, etc., for which appropriate traffic management policies are provided. However, the network operator does not always know the nature and category of incoming traffic, which complicates the task of managing network resources and dynamically allocating bandwidth. Unidentified and non-allocated traffic flows from the network's point of view are processed according to the Best Effort principle.

The first approaches to defining traffic classes were based on a list of well-known TCP and UDP ports, but with the advent of dynamically changing ports, the use of this method became impossible.

The well-known DPI (Deep Packet Inspection) technology allows for "deep" analysis of packet headers at the upper levels of the OSI model. Nevertheless, with the help of the DPI system, it is also not always possible to identify the nature of the data flow, for example, in cases of encrypted or tunneled traffic. In addition, DPI technology does not allow to predict the future characteristics of traffic flows, such as the rate or frequency of arrival of packets.

DOI: 10.4018/IJERTCS.289198

*Corresponding Author

The flow characteristics are often evaluated as packets arrive at the network interface. Flow Table as an example allows estimating the current mean rate of each identified flow. But the current flow rate does not always correspond to the mean rate of the entire flow, so forecasts using the mean rate remain quite inaccurate.

Some approaches involve predicting the bandwidth used based on time series, such as the ARIMA (AutoregRessive Integrated Moving Average) model. But in this case, the forecast is given short-term and is made based on all flows, which does not allow to evaluate the characteristics of each of the flows.

Recently, data mining methods have been used more and more effectively in telecommunications, especially Machine Learning (ML) approaches, to solve a wide range of tasks, including traffic classification and determination of its characteristics.

Supervised Learning and Unsupervised Learning stand out among the methods of machine learning. Supervised Learning methods imply the presence of a database that consists of a certain number of different samples, each of which is characterized by its own set of features and the corresponding class. This database is divided into a training and test sequence. The training sequence is used to build a classifier or regressor model, and the test sequence is used to evaluate it. During testing, the algorithm's efficiency is checked by comparing the predicted values and the true classes. Supervised Learning methods are fast and accurate but can only predict classes known to the model initially. For Unsupervised Learning methods, class values are not defined, which complicates the task and reduces prediction accuracy, but it is possible to detect new classes.

The authors suggest using Supervised Learning methods to speed up classify traffic flows based on their characteristics. The parameters of the first fifteen packets of the flow are analyzed, such as the length of each packet and the inter-interval arrival time between two consecutive incoming packets and the parameters calculated on their basis. The model uses them to determine the class of the traffic flow. A class refers to a specific application. Using Unsupervised Learning methods, the model expands, adds new classes to the model, and refines existing classes (Deart, Mankov, & Krasnova, 2021; Deart, Mankov, & Krasnova, 2020; Mankov, & Krasnova, 2019; Mankov, Deart, & Krasnova, 2021; Mankov, & Krasnova, 2017).

The purpose of this paper is to extend the current model by developing blocks for predicting flow properties. The methods of Supervised Learning are used to estimate the mean rate and interarrival time of packet arrival. The forecast is made for each class separately-thus, the individual properties of the application, the characteristics of the current flow based on the first fifteen packets, and information about previous flows are taken into account. To build the model, the authors analyze five regression methods: Boost, Random Forest regression, Linear, Lasso, and Voting regression. The regression result is compared with the forecast based on the mean rate and the mean value of the interarrival time of packet arrival. The estimation of forecasting results depending on different methods of data preprocessing is also given.

The article is structured as follows: Section 2 provides an overview of the work on the topic of the current article, highlights the main directions of research development, and considers the problems of related fields. Section 3 provides a brief theoretical overview of the methods used to predict traffic characteristics, and section 4 – methods of data preprocessing. Section 5 details the dynamic traffic classification model, which includes real-time traffic classification, adding new applications, and predicting traffic flow characteristics individually for each class. Section 6 describes the experimental studies carried out in the work, in particular, the methodology of database formation, preliminary studies of the data set, the impact of data preprocessing methods on the results of traffic forecasting, and a comparison of the predictive abilities of various models. Section 7 is devoted to the general conclusions of the article and the definition of a plan for future research.