UDC 004.75

STUDY OF THE EFFICIENCY OF MACHINE LEARNING ALGORITHMS FOR TRAFFIC CLASSIFICATION IN MOBILE NETWORKS

Sushko O.V., Astrakhantsev A.A.
Educational and Scientific Institute of Telecommunication Systems, Igor Sikorsky Kyiv Polytechnic Institute, Ukraine
E-mail: aastrakhantsev@its.kpi.ua, sushko.oleksandra@gmail.com

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ТРАФІКА В МОБІЛЬНИХ МЕРЕЖАХ

У статті ставиться актуальне завдання аналізу ефективності алгоритмів машинного навчання для вирішення завдання класифікації трафіка в мобільних мережах у режимі реального часу. Для досягнення мети аналізується точність класифікації та швидкодія для найпоширеніших алгоритмів машинного навчання та визначається оптимальний алгоритм за критерієм точності класифікації. Окрім цього, у статті виконано оцінку важливості полів датасету для класифікації трафіка.

The article presents the actual task of analyzing the effectiveness of machine learning algorithms to solve the task of traffic classification in mobile networks in real time. To achieve the goal, the classification accuracy and speed of the most common machine learning algorithms are analyzed and the optimal algorithm is determined based on the criterion of classification accuracy. In addition, the article evaluates the importance of dataset fields for classification.

Classification of network traffic can be carried out based on the use of information from different levels of the OSI (Open Systems Interconnection) model. At the physical layer, analysis and classification can be performed based on bit sequences and volume traffic [1]. At higher levels, port numbers, packet contents, flow identifiers, and packet headers can be used for this. At the same time, the characteristics of network traffic at each of the levels differ. For example, at the packet flow level, network traffic is characterized by packet size and the time interval between packets. Analysis at the bit sequence level mainly concerns such characteristics as transmission intensity and channel bandwidth. At the packet flow level, the procedure for the arrival of IP packets, i.e. their delay and loss, is also considered.

In many works, such algorithms as random forest (RF), KNN, ANN and SVM are proposed for use. Therefore, these algorithms were chosen for research in this work. In addition to machine learning algorithms for traffic classification, Deep Packet Inspection (DPI) methods also can be used. Deep Packet Inspection is the most advanced technology for traffic classification because it is the most accurate technique [2]. Therefore, often the most popular products, both commercial and open source, rely on DPI when classifying traffic. However, the

actual effectiveness of DPI is still uncertain, as the limited number of public datasets limits the comparison and reproducibility of results [3].

Now let's move on to the analysis of features for traffic classification. Consider the packet fields given in the dataset that can be used as features for model training and traffic classification. A short list of fields is given in the table 1.

Table 1. List of fields for traffic classification.

Flow ID	Flow Duration	Flow bytes/s	Average Packet Size
Source IP	Total Fwd Packets	Flow packets/s	Avg Fwd Segm Size
Source Port	Total Backward Packets	Average Packet Size	Avg Bwd Segm Size
Destination IP	Total Length of Fwd Pck	Flags (x 8)	Fwd Header Length
Destination Port	Total Length of Bwd Pck	Packet Length Mean	Down Up Ratio
Protocol	Fwd Packet Length Max	Bwd Packet Length Max	Label
Timestamp	Fwd Packet Length Min	Bwd Packet Length Min	App name

As can be seen from the table. 1, there are a sufficient number of fields that are essential and necessary to perform a classification with high accuracy, but there are also a large number of fields whose influence is difficult to assess at this stage. It should be emphasized that, on the one hand, an increase in the number of fields (features) for classification allows to increase the accuracy of classification, on the other hand, an increase in the number of features increases its complexity, therefore, after determining the best classification algorithms, it will be necessary to optimize the features of classification to form a minimally sufficient set of fields (features), which will ensure the specified classification accuracy.

Various metrics [3] will be used to evaluate the effectiveness of machine learning algorithms: accuracy, precision, recall, and F1 metric.

Accuracy means the ratio of correctly classified samples (packets) of the traffic flow to the total number of samples:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \ , \quad (1)$$

Precision is a measure of the ratio of positive, correctly predicted packets in traffic to the total number of positive classification predictions:

$$Precision = \frac{TP}{TP+FP} . \quad (2)$$

Recall measures the ratio of actual positive, correctly predicted packets in traffic:

$$Recall = \frac{TP}{TP+FN} \ . \quad (3)$$

The F1 metric represents the average of clarity and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} . \quad (4)$$

At the first stage of research, a list of applications that could not be successfully classified based on the data of the studied dataset was filtered. The number of available packets was used as a filtering criteria. The 25 applications

with the fewest packets (fewer than 500 packets in the dataset) were discarded.

Using a balanced dataset for the artificial neural network (ANN) algorithm, an assessment of the maximum achievable accuracy and its dependence on the speed code was performed.

Also, for the balanced dataset, an assessment of the best model parameters and classification accuracy was performed in the case of using the "random forest" (RF) algorithm.

The classification accuracy in this case depends on which packet fields (traffic information) are present in the dataset and on the importance of these fields for recognition. Therefore, in order to compare the importance of the fields, we conducted a study where we removed fields from the dataset that did not affect the accuracy of the classification (Fig. 1).

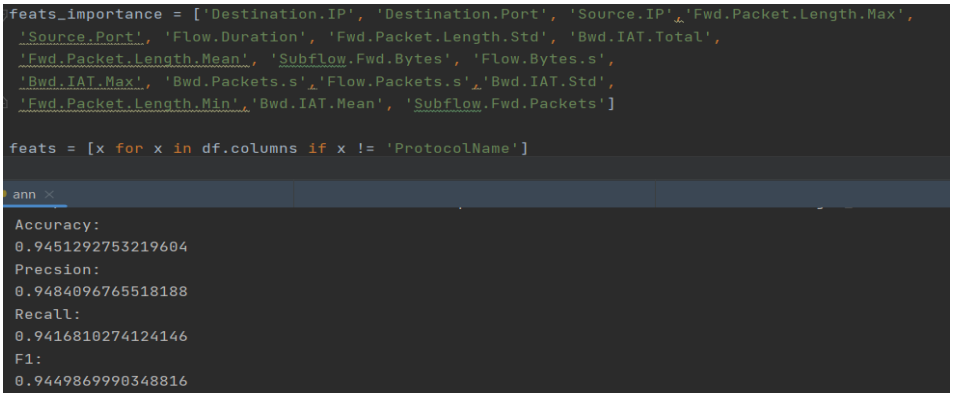

Fig. 1. Important of dataset fields for classification, which affect accuracy.

Conclusion. The scientific novelty consists in determining the parameters of machine learning algorithms that are optimal according to the criterion of accuracy for solving the problem of traffic classification in mobile communication networks of the 5th and 6th generations. Also, importance of the parameters (fields) of the dataset for classification are defined. The number of used field was decreased from 54 to 18 with decrease accuracy on 0.02. The proposed algorithms and parameters are the first stage of multi-step processing of packets in the network, which, together with clustering, slicing and distributed processing, will improve the efficiency of the mobile communication system in general.

References

1. Zaborovsky, V.S. (2010), Traffic analysis in packet switching networks, St. Petersburg: SPbGPU, 90p.
2. Bujlow, T., Carela-Español, V., Barlet-Ros, P. (2015), "Independent comparison of popular DPI tools for traffic classification", Computer Networks, No. 76, P.75-89. DOI: https://doi.org/10.1016/j.comnet.2014.11.001
3. AlZoman, R.M.; Alenazi, M.J.F. (2020), “A Comparative Study of Traffic Classification Techniques for Smart City Networks”, Sensors, No. 21, 4677, P. 1-17. DOI: https://doi.org/10.3390/s21144677