RLCS: A Classification System Based on Random Forest and Logistic Regression for Hybrid zero-day traffic

Yulong Liang, Fei Wang, Shuhui Chen
the National University of Defense Technology
Changsha, China
{liangyulong20,wangfei09a,shchen}@nudt.edu.cn

ABSTRACT

Traffic classification has attracted public attention for a long time because of its essential role in network management. However, the presence of zero-day traffic, network traffic generated by previously unknown applications, leads to a significant reduction in the practicability and effectiveness of conventional traffic classification schemes. This poster innovatively proposes a traffic classification scheme named RLCS to accomplish the high accurate traffic classification task in hybrid zero-day traffic. The evaluations with real-world traffic verify the effectiveness and broad applicability of RLCS.

KEYWORDS

traffic classification, zero-day applications, random forest, logistic regression

ACM Reference Format:
Yulong Liang, Fei Wang, Shuhui Chen. 2022. RLCS: A Classification System Based on Random Forest and Logistic Regression for Hybrid zero-day traffic. In *Asia-Pacific Workshop on Networking (APNet '22), July 1–2, 2022, Fuzhou, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3542637.3542644

1 INTRODUCTION

A high-precision traffic classification system is the basis of network management and monitoring. However, the increasing popularity of traffic encryption and dynamic ports puts methods based on DPI and port mapping in crisis. Consequently, the ML-based traffic classification method has quickly become one of the hottest points in the related field.

However, ML-based methods usually have poor classification performance in the social network. The most crucial reason is that the considerable number of zero-day traffic leads to a decrease in the accuracy of conventional classification models [4]. More specifically, conventional classification models misclassify zero-day traffic into known classes without perceiving the changes in the environment.

This poster proposes a novel traffic classification system, RLCS, to tackle the problem, aiming to accomplish the traffic classification task in hybrid zero-day traffic. The proposed RLCS is capable of accurate zero-day traffic detection and efficient application traffic classification. Experiments were conducted on a large-scale, real-world traffic datasets. Preliminary results prove the effectiveness and broad applicability of RLCS.

2 SYSTEM DESIGN

Our scheme is an ensemble system with a double-layer structure. The lower layer comprises two random forests, with the upper layer composed of two logistic regression models.

The ability of our solution to identify zero-day traffic is based on two key points. Firstly, the probability vector outputted by the random forest provides an essential basis for determining whether an input flow belongs to known applications. In this poster, we define the probability vector ρ of each flow as $\rho = \{p_1, ..., p_N\}$, where ρ_i denotes the probability that the flow belongs to the $i-th$ application. Random Forest is an ensemble model in which a flow's probability vector computed from the model is the average of all probability vectors output from each sub-classifier. Secondly, logistic regression models are constructed to distinguish between known and zero-day traffic by probability vectors. Logistic regression models perform the binary classification task, trained by simulated zero-day traffic and partial known applications' traffic.

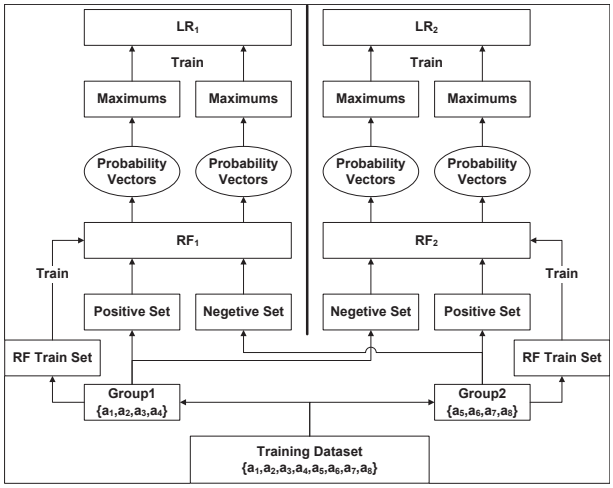


Figure 1: RLCS construction.

APNet '22, July 1–2, 2022, Fuzhou, China

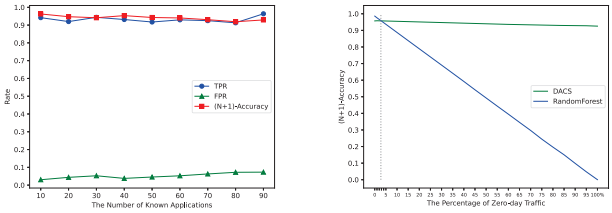
ACM ISBN 978-1-4503-9748-3/22/07…$15.00
https://doi.org/10.1145/3542637.3542644

System Construction. Fig. 1 shows the construction of RLCS. We divide the training dataset into two groups, Group1 and Group2, each containing traffic of half applications. Firstly, we randomly select a portion of the traffic in Group1 to train the first random forest (RF1). Similarly, partial traffic in Group2 is used by training the second random forest (RF2). Secondly, we construct the positive and negative sets for training logistic regression model. For brevity, the following describes the construction of the first logistic regression model (LR1). Samples in the positive set come from Group1. To simulate the zero-day traffic, samples in the negative set come from Group2, which has different applications from the training set of RF1. Afterward, we input all samples from the positive and negative sets into RF1 and obtain the corresponding probability vectors. Then, we replace each sample in the positive and negative sets using the maximum of all elements in the corresponding probability vector. Finally, we construct a binary classification model LR1 using the changed positive and negative sets. To promote the performance, the training set of each sub-model should not use the same sample. Additionally, RLCS can be updated by retraining one of the model pairs, such as RF1 and LR1.

Traffic Classification. Zero-day traffic detection and application traffic classification for a test flow are executed sequentially. Zero-day traffic detection aims to map the test flow into two generic classes, "Known" and "Unknown". Application traffic classification classifies traffic judged to be known into specific applications at a fine-grained level. For brevity, we assume that there is now a flow ξ to be tested. We input ξ into RF1 and RF2 and get two probability vectors, respectively. Then, the maximum of each probability vector is inputted into the corresponding logistic regression model. Afterward, we obtain two labels, denoting the positive label as *positive* with the negative label as *negative*. If LR1 outputs a label of *positive*, ξ is judged to belong to an application in the Group1. Conversely, a label of *negative* means ξ is a sample from applications outside Group1. Therefore, if each label is *negative*, ξ is judged as a zero-day flow. Conversely, ξ is assigned to a known generic class because there is at least one *positive*. If there is only one *positive* from the $i-th$ logistic regression model, the final result is equivalent to the classification result of the $i-th$ random forest. If each label is *positive*, we choose to follow the classification result of the particular random forest, which outputs the maximum probability.

3 PRELIMINARY EVALUATION

Experiments are conducted to evaluate the performance of RLCS on a public mobile traffic dataset, NUDT_MobileTraffic [1], taken from the work of Chen's team at the National University of Defense Technology. The traffic of 105 randomly selected applications is used for subsequent experiments, and the last 15 applications are considered zero-day applications. We automatically extracted features through CICFlowMeter [2, 3], which is a network flow generator and analyzer developed by Canadian Institute for Cybersecurity. Before experiments, we removed flows with less than five packets, deleted samples containing INF and NAN, and normalized the features. No traffic and information from zero-day applications will be allowed to participate in any training phase. As an evaluation of the comprehensive performance of RLCS, we focus on the False Positive Rate (TPR) and False Positive Rate (FPR) of zero-day traffic detection. Furthermore, we defined the *(N+1)-Accuracy* as the accuracy of the (N+1)-class classification task, which maps the test flow into one unknown generic class and all the known classes.



(a) (b)

Figure 2: Classification results in varying scenarios. (a) Varying data scales. (b) Varying percentages of zero-day traffic.

Evaluation with Varying Data Scales. The classification performance of RLCS is evaluated with varying known application numbers. We maintain a 4:1 ratio of known flows to zero-day flows and change the number of known applications from 10 to 90. As we can see in Fig. 2(a), no matter how much the application volume increases, TPR maintains more than 0.9, and the FPR maintains less than 0.1. Moreover, (N+1)-Accuracy maintains more than 0.9, and we can always keep more than 0.9 precision in determining whether a flow belongs to known applications.

Evaluation with Varying Percentages of Zero-day Traffic. We consider scenarios with varying percentages of zero-day traffic to test the adaptability of RLCS to the environment. The random forest, which has outstanding performance in traditional traffic classification, is employed as a comparison. We keep the number of known applications at 40 and gradually change the percentage of zero-day traffic in the test from 0% to 100%. As we can see in Fig. 2(b), regardless of the change in the environment, our scheme can be applied very well and always maintain a high level of accuracy because of its significant advantage of the ability to identify zero-day flows in hybrid traffic. In contrast, the random forest has extremely poor applicability when the rate of zero-day traffic is too large.

4 CONCLUSION

This poster proposed a traffic classification system, performing classification tasks accurately in hybrid zero-day traffic. Preliminary experiments prove the effectiveness and broad applicability of RLCS.

REFERENCES

[1] Abby-ZS. 2022. NUDT_MobileTraffic. https://github.com/Abby-ZS/NUDT_MobileTraffic.
[2] Gerard Draper-Gil, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. 2016. Characterization of encrypted and vpn traffic using time-related. In *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*. sn, 407–414.
[3] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. 2017. Characterization of tor traffic using time based features.. In *ICISSp*. 253–262.
[4] Jun Zhang, Xiao Chen, Yang Xiang, and Wanlei Zhou. 2013. Zero-day traffic identification. In *International Symposium on Cyberspace Safety and Security*. Springer, 213–227.