Towards Time-varying Classification Based on Traffic Pattern

Yiyang Shao*‡, Luoshi Zhang†, Xiaoxian Chen†, Yibo Xue‡§
* Department of Automation, Tsinghua University, Beijing, China
† Computer Science & Technology College, Harbin Univ. of Sci. & Tech., Harbin, China
‡ Research Institute of Information Technology, Tsinghua University, Beijing, China
§ Tsinghua National Lab for Information Science and Technology, Beijing, China
shaoyy11@mails.tsinghua.edu.cn, {zhangluoshi, chenxiaoxian, yiboxue}@tsinghua.edu.cn

***Abstract*—Many important network security areas, such as Intrusion Detection System and Next-Generation Firewall, leverage Traffic Classification techniques to reveal application-level protocols. Machine Learning algorithms give us the ability to identify encrypted or complicated traffic. However, classification accuracies of Machine Learning algorithms are always facing challenges and doubts in practical usage. In this paper, we propose a time-varying Logistic Regression model embedded with traffic pattern. The comparison between original Logistic Regression model and time-varying one shows an effective improvement in accuracy. We hope to exploit a new way to implement Machine Learning algorithms in network traffic analysis areas by considering the characteristics of traffic changes in time domain.**

***Index Terms*—Traffic Classification; Traffic Pattern; Time-varying Model; Logistic Regression**

I. Introduction

With the Internet rapidly evolving, there is a growing demand to identify Internet application types in network security instruments, *e.g.,* Network Intrusion Detection System (NIDS) and Next Generation Firewall (NGFW). Internet Service Providers (ISPs) also rely on traffic classification to analyze and optimize their networks. Traditionally, Deep Inspection (DI) techniques provide the most effective way to classify network traffic, and have been widely deployed in middle boxes that provide network security services. Traffic classification based on deep inspection can only solve those unencrypted traffic, however, more and more network applications transmit their data with encrypted protocols nowadays, bringing a severe challenge to traffic classification.

Machine learning algorithms ignore packets' payload and utilize statistical characteristics of traffic, thus providing us the ability to classify those encrypted or complex protocols. Researchers have applied many novel algorithms, including Naive Bayes, Support Vector Machine, Logistic Regression, and so on, and identified a variety of protocols successfully [1]. But the fact is that accuracies of these algorithms are good enough in experiments, commonly over 90 percentage, but in practice these algorithms couldn't work well, achieving quite few usage in industrial deployment.

We study the decrease of accuracy from experiment to practice, and find that this is mainly because most machine learning algorithms require that the samples obey independent and identically distribution (i.i.d.). However, because network system is time-varying, the probability distributions of application's statistical characteristics also change with time. This sharply influences the decision of classifiers and makes machine learning algorithms impractical online.

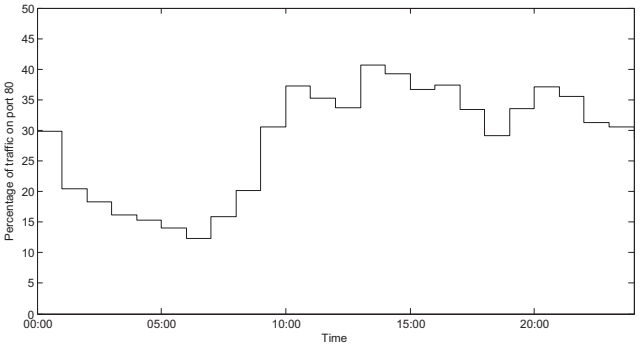

Fig. 1. An example of traffic pattern on port 80 during a day

In this paper we propose a time-varying Logistic Regression model embedded with traffic pattern. Unlike the common ways that train and classify traffic using a static model, we consider the influence of traffic changes in time domain and embed traffic pattern into classification model. The comparison between original Logistic Regression model and time-varying one shows an effective improvement in accuracy. We hope our work can exploit a new way to implement Machine Learning algorithms in network traffic analysis area.

II. Traffic Pattern

Network traffic can be seen as the physical reflector of humans' online behavior. Obviously, specific human behavior patterns will conduct that network traffic obeys some regularities of distribution [2]. For example the traffic volume is low in midnight and high in daytime. Another example can be that the web search traffic is larger in working time than it in non-working time, and the video traffic acts oppositely. We define the distribution regularity of traffic in a time window as the *traffic pattern*. Fig.1 shows a typical traffic pattern on port 80 of a day in an office building network in campus.

We use destination port to separate different patterns in this paper. Noted that we have no priori knowledge of any port

978-1-4799-5890-0/14/$31.00 ©2014 IEEE

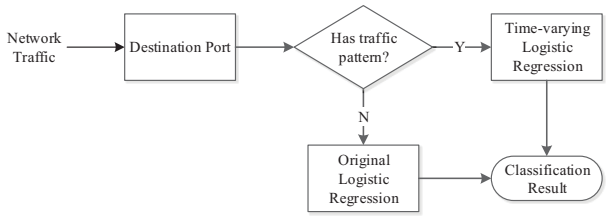

Fig. 2. Classification Procedure

number, which means we cannot assume that HTTP protocol commonly works on port 80 or BitTorrent protocol works on port 6881-6889. Here we only use port number to separate different types of traffic, just like using a raster to separate spectrum. So traffic pattern differs on ports, and we find that on many ports the traffic patterns are stable in different days.

To formalize the traffic pattern, we define the probability of a specific port. For a specific port n, if it has a stable daily pattern, its probability is defined as follow:

$$P(n,t) = \{P(n) \mid P(n) \in (0,1), t \in [0,24)\}$$

in which $P(n)$ represents the percentage of traffic volume on port n among all ports at the time t.

In practice, we calculate the average traffic volume percentage of port n in each hour, and use this average as the representative of this hour. So, for each port n, $P(n,t)$ is a segmented step function, as Fig.1 shows.

III. Time-varying Logistic Regression

Once we get traffic patterns, we can embed them into machine learning models to enhance the classification accuracy. In this paper, we choose Logistic Regression as an example to show how we do this. However, not only Logistic Regression can be embedded with traffic pattern, almost all machine learning algorithms can fit our deduction.

We consider a 2-class case in the logical deduction, in which $c \in \{c_i,\ i = 1,2\}$ are the labels of each class, $x \in R$ are the features of samples, $\omega \in R$ are the parameters in logistic function. So the probability that x belongs to class c is:

$$P_o(c \mid x) = \frac{e^{\omega^T \cdot x}}{1 + e^{\omega^T \cdot x}}$$

in which $P_o(c \mid x)$ is the original conditional probability, and the original decision function of Logistic Regression is:

$$P_o(c \mid x) \geq 0.5, x \in c_1$$

$$P_o(c \mid x) < 0.5, x \in c_2$$

Now we consider the time-varying model, and use $P_t(c \mid x)$ to represent its conditional probability, so we have:

$$P(c \mid x) = \frac{P(c,x)}{P(x)} = \frac{P(x \mid c) \cdot P(c)}{P(x)}$$

$$\frac{P_o(c \mid x)}{P_t(c \mid x)} = \frac{P_o(x \mid c) \cdot P_o(c)}{P_t(x \mid c) \cdot P_t(c)} \cdot \frac{P_t(x)}{P_o(x)}$$

according to the Bayes' theorem.

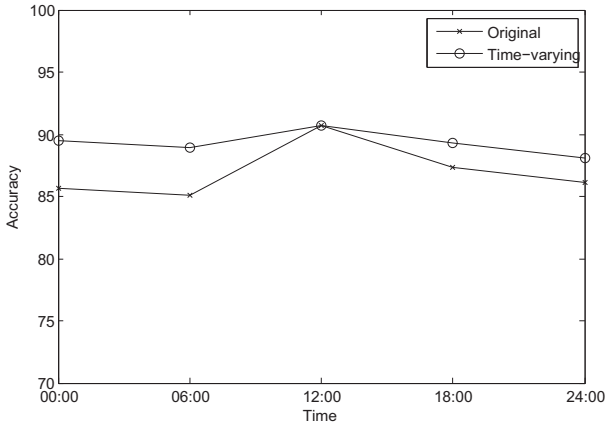

Fig. 3. Classification Accuracy of BitTorrent

Because the density function $P(x \mid c)$ keeps constant with time changing, which means $P_o(x \mid c) = P_t(x \mid c)$, so the decision function of time-varying model is:

$$arcmax\ P_t(c \mid x) = argmax\ P_o(c \mid x) \cdot \frac{P_t(c)}{P_o(c)}$$

Specific to our traffic classification area, $P_o(c)$ can be calculated using the training data, and $P_t(c)$ is just the concept of traffic pattern in former section.

In practice, not all of the ports have stable traffic patterns, we do our classification as Fig.2 shows. When the traffic to be classified comes, first it need to judge whether this port has traffic pattern. If it has, then the time-varying model is used to enhance classification accuracy, and the original model will be used when the port doesn't have stable pattern.

IV. Evaluation

Due to the limitation of space, here we only demonstrate a simple two-class classification of BitTorrent and non-BitTorrent. The traffic traces used in our experiments are from campus network and cover a whole day to generate the pattern. The traffic pattern in this case is generated on port 6881. The classification model is trained using part of the data between 11:30 a.m. to 00:30 p.m. Results are shown in Fig.3. We make 5 test points at different time intervals. In the original model, classification accuracy is remarkably influenced by time, while the time-varying model shows obvious progress and stays stable as time changes. So we conclude that the time-varying Logistic Regression model we proposed in this paper achieves an effective improvement in accuracy.

V. Acknowledgments

This work was supported by the National Key Technology R&D Program of China under Grant No.2012BAH46B04.

References

[1] T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[2] A. Hassidim, D. Raz, M. Segalov, and A. Shaqed, "Network utilization: the flow view," in *Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM)*, 2013.