A Traffic Classification Approach based on Characteristics of Subflows and Ensemble Learning

Changyu Wang
MOE Key Lab for Intelligent Networks and Network Security
Xi'an Jiaotong University Xi'an, China
Email: cywang@sei.xjtu.edu.cn

Xiaohong Guan, *IEEE Fellow*
MOE Key Lab for Intelligent Networks and Network Security
Xi'an Jiaotong University, Xi'an, China
Email: xhguan@sei.xjtu.edu.cn

Tao Qin
MOE Key Lab for Intelligent Networks and Network Security
Xi'an Jiaotong University, Xi'an, China
Email: tqin@sei.xjtu.edu.cn

***Abstract*—Recently, network traffic classification has attracted a great deal of attention among researchers. In this paper, we proposed a traffic classification approach based on characteristics of subflows and ensemble learning. Aiming at neutralization of unstable network environment as well as taking advantage of ensemble learning, we divided the traffic flows into different subflows in order to reduce the affection of time. Moreover, we develop truncation method on flows for real-time processing and an aggregation machine learning method based on accuracy of each classifier to different applications. Finally, the experimental results based on actual traffic traces collected from the campus network of Xian Jiaotong University verify the effectiveness of our methods.**

I. Introduction

NETWORK traffic classification has attracted tremendous attention and efforts from global scientists over recent years. A great deal of fields ranged from network security to network management, such as Quality of Service(QoS) control, network traffic charging and intrusion detection, benefit from the achievement of network traffic classification [1]. However, the classification result is easily affected by external network environment. To achieve high performance of network traffic classification as well as neutralize the influence by network environment, this paper proposes a traffic classification approach based on characteristics of subflows and ensemble learning to settle the problem from another exploratory perspective.

Inspired by previous research achievements as well as to efficiently utilize time relevant features, we propose a traffic classification approach based on characteristics of subflows and ensemble learning. The process could be elaborated as follows: Firstly, flows are truncated from beginning with first few packets. This measure aims at identifying applications before the termination of a connection, which are suitable for online traffic monitoring. After that, flows are segmented into several small fragmentss. The segmentation intends to handle with time-related features via a relatively coarse approach which avoids to get trapped into unstable time relevant features. Afterwards, we extract the features which are relevant to the corresponding application from subflows. PSD(packet size distribution) and packet arrival intervals statistically related features are selected as inputs to classifiers based on their differentiation abilities for the three representative applications: Http, BitTorrent and Skype. Subsequently, the classifiers output their results with the help of relatively mature machine learning techniques. Then our aggregation method is exploited to combine results. The combination is based on the classification performance of different classifiers to different result labels. The aggregated result elevates the final classification performance. This approach is proved efficiently boosting the overall classification performance on the test dataset.

TABLE I
Basic Statistical Feature of the Dataset

Dataset Size	Packets Count	Flow Count	
		TCP Connection	UDP Connection
83.3GB	54218796	957881	599315

II. Data Collection and Basic Analysis

A. Data Collection

Based on the traffic monitoring platform set in the campus network of Xi'an Jiaotong University, we collect some actual traffic traces. The user within the collection network region includes regular students and network administrators. Some basic statistical features are described in the Table I.

B. Pre-Classification by NDPI

In order to establish the benchmark for training and testing dataset, state-of-art flow classification tool-NDPI is applied to construct ground truth for subsequent procedures for flow classification. The NDPI library is an open-source tool which is built under the framework of the prevalent OpenDPI library [2]. Due to high accuracy achieved by nDPI, the classification result obtained from NDPI is chosen as the ground truth for training and testing. Based on nDPI's classification result, we decide to utilize the three most representative applications: http, bittorrent and skype, to verify our proposed method,

978-3-901882-89-0 @2017 IFIP

III. Framework of the Proposed Method

A. System Model

Figure. 1 illustrates the classification process of our proposed scheme, which concentrates on flow-level traffic classification. It is hypothesized that several applications' statistical attributes are only distributed among some specific intervals in a flow [3]. For instance, a http flow could be divided into three states: connection construction, connection duration, connection release. Http could be identified by 3 handshakes in the state of connection construction. According to the instability of time-relevant features for traffic classification as well as the preceding assumption, a heuristic approach is applied by segmenting a flow into subflows. Statistical features are extracted from each subflow and then each of them is input into the corresponding classifier for embryo traffic classification. All the embryo classification results are combined based on the ensemble machine learning model for final classification results.

The explicit procedures could be elaborated as following steps:

Step 1:Flow interception. All the flows are truncated into the same length for the sake of quick detections.

Step 2:Flow segmentation. Flows are separated into subflows to compensate for the instability of time-relevant features. At the same time, the assumption of unique attributes among different intervals is satisfied.

Step 3:Feature extraction and machine learning processing. Features are distilled from each subflow according to the following feature analysis and then go through corresponding classifiers which exert machine learning techniques.

Step 4:Results Aggregation. Results from individual classifiers are combined to generate a final predicting result. The voting rule depends on the performance of each classifier to its predicting result.

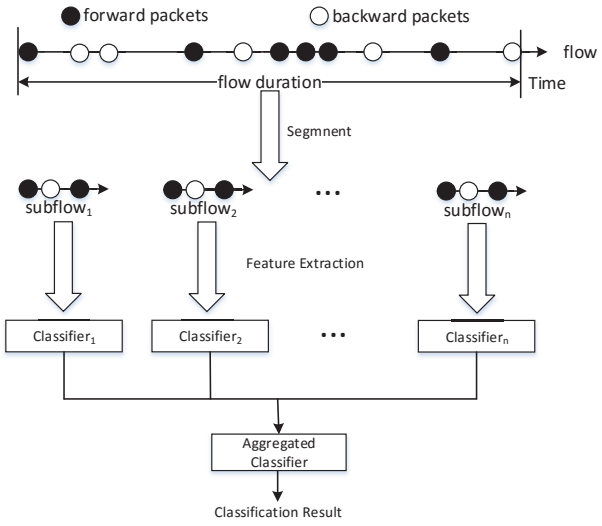

Fig. 1. System Model

B. Parameter Designation for Flow Interception

For the sake of predicting the category of a flow, it is necessary to monitor first few packets rather than the entire flow transmission. We employ a method called 'Category Distance' to select the appropriate number of the first few packets. The method could be elaborated as equation (1)

$$d(\alpha,\beta) = \sum_{target(flow(i))=\beta} d(flow(i),\mu(\alpha)) + \sum_{target(flow(j))=\alpha} d(flow(j),\mu(\beta)) \tag{1}$$

where α and β are two categories like HTTP and BitTorrent. For every flow belonged to a category β, we calculate the distance between its feature vector and the center of the corresponding vectors of category α(which denotes as $\mu(\alpha)$). Feature vector is a vector that includes selected features extracted from a flow and center vector is the calculated center of a bunch of feature vectors. Then sum them up as the distance from α to β. Similarly, the distance from β to α could be obtained. The aggregated result is the distance between α to β, which could indicate the discrimination degree between two categories.

The distance between two items is measured by means of Mahalanobis distance as equation (2) [4].

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})} \tag{2}$$

where $\vec{x}$ and $\vec{y}$ denote feature measurements of two individual flows and S^{-1} represents the inverse matrix of the corresponding covariance matrix. This measurement takes correlation of the dataset into consideration and thus could be taken as unitless and scale-free for elimination the interruptions by different scales of features.

By contrast to distance between diverse categories changing along with different packets number, the truncation position of the first few packets is determined.

C. Results Aggregation

Here we would like to present a method for integration of classification results generated from individual classifiers. Consider a set of results from N classifiers, $Result = \{R_1, R_2, ..., R_n\}$. Meanwhile, there are k predefined applications, which are denoted as $ApplicationSet = \{\omega_1, \omega_2, ..., \omega_k\}$. For each $R_i (0 \leq i \leq n) \in Result$, $R_i \in ApplicationSet$.

As for different classifiers,they exhibit different responses to various applications. For instance, skype has some specific features which could distinguish it from others in the data transmission phase. Therefore, the intuitional means to deal with results of aggregation is voting based on the respective response to different applications of individual classifiers. The voting weight is chosen based on each predicatior's classification accuracy rate to the corresponding classification result. The weight is calculated by the function (3) [5]

$$weight_{\omega i} = -\log(\frac{1-accuracy_{\omega i}}{1+accuracy_{\omega i}}) \tag{3}$$

where $weight_{\omega i}$ represents the weight of the i-th classifier when its classification result is ω and $accuracy_{\omega i}$ denotes corresponding training accuracy rate. Instead of simple summarization of results of classifiers, this nonlinear function aims at amplifying the influence of classifiers which have good classification performance on some specific applications and trimming the effect of classifiers which have relatively poor performance.

The final aggregated result could be described as

$$\omega^* = \underset{\omega \in ApplicationSet}{\arg\max} - \sum_{i=1}^{n} \triangle_{\omega i} \log(\frac{1 - accuracy_{\omega i}}{1 + accuracy_{\omega i}}) \quad (4)$$

where $\triangle_{\omega i}$ is a binary value function as

$$\triangle_{\omega i} = \begin{cases} 1 & \text{if } R_i = \omega \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

D. Method Justification

Here is a proof of why the integrated result achieves a better result than that of a single classifier. Consider P(F,L) is a result of a single classifier, L is the training set and F is an intact flow to be classified. The aggregated result can be described as:

$$P_a(F, L) = E_{f \in F} P(f, L) \quad (6)$$

where f represents segmented subflows, $P_a(F, L)$ is the aggregated result and E is on behalf of expectation. Set y as the class label of respective subflows. The average classification error by respective single classifier is written as:

$$e_s = E_{f \in F}(y - P(f, L))^2 \quad (7)$$

The comparitive classification error calculated from aggregated classifier is

$$e_a = E_{f \in F}(y - P(F, L))^2 \quad (8)$$

By Cauchy-Schwarz inequality, we can obtain

$$(E(X))^2 \leq E(X^2) \quad (9)$$

where X is a random variable, thus we get

$$(E_{f \in F} P(f, L))^2 \leq E_{f \in F} P^2(f, L) \quad (10)$$

From equation (7), (8), (10), we can reach our conclusion

$$\begin{aligned} e_s &= E_{f \in F}(y - P(f, L))^2 \\ &= E_{f \in F} y^2 - 2E_{f \in F}(yP(f, L)) + E_{f \in F} P^2(f, L) \\ &\geq E_{f \in F} y^2 - 2E_{f \in F}(y) E_{f \in F} P(f, L)) \\ &+ (E_{f \in F} P(f, L))^2 \\ &= E_{f \in F}(y - P_a(F, L))^2 \\ &= e_a \end{aligned} \quad (11)$$

The deduction is implemented under an assumption that the mean performance of all the individual classifiers which train on subflows is similar to a classifier which trains on the entire flows. We hypothesize this assumption holds in our experiment. From the inequation (11), the aggregated classifier achieve a lower classification error rate compared with that of several singular classifiers exerted on subflows.

IV. Experiment Result

A. Parameter Designation for Flow Interception

The Mahalonobis distance between three main applications changing along with different packets counts is illustrated in Fig. 2. In the figure, the greater distance between two applications implies that these applications are more distinguishable from one another on the selected feature degree. As the analysis result tells, the trend of the distance of the applications goes up as the packets count arises in the early stage and it barely changes when packets count is greater than 1500. To obtain a stable feature representation and a goal for real-time classification, we choose 1200 as a threshold of packets counts in the generation of subflows.

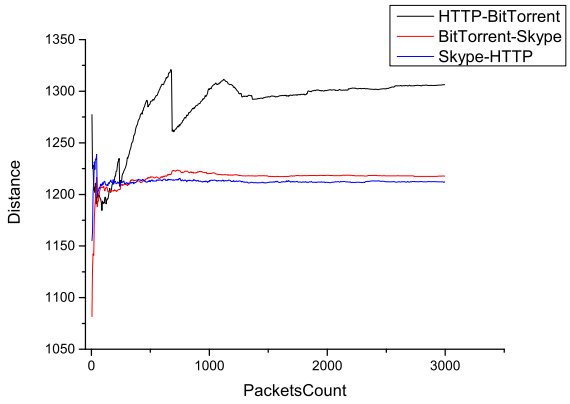

Fig. 2. Mahalanobis Distance Measurement

B. Performance Evaluation

Based on the number of the first few packets that we choose, we separate 1200 packets uniformly into three phrases which stand for beginning of the connection, data transmission and transmission closing, respectively. The former research implies that naive bayes classifier could achieve a satisfying result on traffic classification [6]. Thus naive bayes is exploited in our experiment.

We employ precision, recall and F_1 to assess the performance.[7]. The PSD and time interval features are distilled as input to each of three classifiers which belongs to the corresponding phase of a flow. The performance of three main applications is shown in Table III via precision, recall and F_1 measurement.

In Table III, the phase of BoC, DT and TC represent beginning of connection, data transmission and transmission closing individually. Due to the available precision and recall rate, there is an apparent trend that the classification performance behaves weaker as the phrase goes by. The BoC gives out a remarkable performance and TC provides an undesirable result on classification. The most two reasonable explanations after the scrutiny of dataset and result are: i) the most part of potential patterns of the three main applications lie in the front of connections. This explanation verifies the justification

TABLE II
AGGREGATED CLASSIFICATION RESULT

Method		Proposed Method	C4.5	Naive Bayes	SVM
HTTP	Precision(%)	92.74	67.56	73.64	71.87
	Recall(%)	98.00	75.34	95.00	90.76
	F Measure(%)	95.30	71.24	82.97	80.22
BitTorrent	Precision(%)	82.14	70.54	76.77	80.24
	Recall(%)	84.89	58.67	76.00	79.68
	F Measure(%)	83.49	64.06	76.38	79.96
Skype	Precision(%)	90.33	74.65	90.36	87.94
	Recall(%)	81.00	68.65	72.00	73.76
	F Measure(%)	85.41	71.52	80.14	80.23
Overall	Precision Rate(%)	89.53	67.56	81.53	81.33

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT PHASE

Phase		BoC	DT	TC
HTTP	Precision(%)	90.12	78.47	60.95
	Recall(%)	97.33	78.60	64.00
	F Measure(%)	93.59	78.54	62.44
BitTorrent	Precision(%)	82.14	67.09	55.99
	Recall(%)	84.33	70.67	57.67
	F Measure(%)	83.22	68.33	56.81
Skype	Precision(%)	88.43	69.50	58.33
	Recall(%)	79.00	65.33	53.67
	F Measure(%)	83.45	67.35	55.90
Overall	Precision Rate(%)	86.89	71.67	58.44

of the choice of the first few packets. ii) We find that nearly 20% of the flows are the connections that possess less than 1200 packets and 13% flows are less than 800 packets. The mentioned flows are truncated by the classifiers. The features of those flows could not represent themselves for the applications to which they belong on a statistical perspective.

F_1 is the harmonic mean of precision and recall. Thus we choose it as the accuracy in equation (4). We choose a dataset from 20:00 to 22:00 as our test set. In order to verify the performance of our proposed method, we import some state-of-the-art network traffic classification methods from previous literature as comparison. The final aggregated result is displayed in Table II.

In Table II, the C4.5, Naive Bayes and SVM are implemented on a single classifier as comparison of the proposed aggregated method. The features which are available for these classifiers are extracted from entire flows. When turn to vertical comparison, our proposed method, the aggregated method has a classification performance promotion about 3 percent up over the original training performance of BoC in Table III. This finding origins from the join force of three classifiers with multi-feature sources. When we shift the view to horizontal comparison, we can find the aggregated method has lifted the precision rate about 8 percent up. The result could verify the effectiveness of our time-scale separation and aggregated method.

V. CONCLUSION

In this paper, with an aim at network traffic classification, we develop an effective flow fragments and aggregation machine learning method. For the sake of avoiding interruption posed by packet time interval, we separate flows into several fragments as independent classifiers and combine the results from the classifiers based on their traffic classification performances. The final aggregated result proves that the proposed method gains a classification advance about 8 percent up over single classifier by classical traffic predicting methods as well as a classfication promotion about 3 percent up over the best one of its separated training classifiers.

VI. ACKNOWLEDGEMENT

The research presented in this paper is supported in part by the National Natural Science Foundation of China(61502438,61672026), Natural Science Foundation of Shaanxi Province (2016JM6040) and Fundamental Research Funds for the central university

REFERENCES

[1] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[2] L. Deri, M. Martinelli, T. Bujlow, and A. Cardigliano, "ndpi: Open-source high-speed deep packet inspection," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 2014, pp. 617–622.

[3] T. T. Nguyen, G. Armitage, P. Branch, and S. Zander, "Timely and continuous machine-learning-based classification for interactive ip traffic," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 6, pp. 1880–1894, 2012.

[4] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, vol. 50, no. 1, pp. 1–18, 2000.

[5] S.-T. Wu, F.-H. Hou, and F. Dai, "Linear approximating method in the transacting process of nonlinear standardization of data." *Xinxi Gongcheng Daxue Xuebao/ Journal of Information Engineering University*, vol. 8, no. 2, pp. 250–253, 2007.

[6] J. Zhang, C. Chen, Y. Xiang, W. Zhou, and Y. Xiang, "Internet traffic classification by aggregating correlated naive bayes predictions," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 5–15, 2013.

[7] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*. Springer, 2005, pp. 345–359.