Journal of Information Processing Systems, Vol.6, No.3, September 2010 DOI : 10.3745/JIPS.2010.6.3.307

A Classifiable Sub-Flow Selection Method for Traffic Classification in Mobile IP Networks

Akihiro Satoh*, Toshiaki Osada**, Toru Abe***, Gen Kitagata**,
Norio Shiratori** and Tetsuo Kinoshita**

Abstract—Traffic classification is an essential task for network management. Many researchers have paid attention to initial sub-flow features based classifiers for traffic classification. However, the existing classifiers cannot classify traffic effectively in mobile IP networks. The classifiers depend on initial sub-flows, but they cannot always capture the sub-flows at a point of attachment for a variety of elements because of seamless mobility. Thus the ideal classifier should be capable of traffic classification based on not only initial sub-flows but also various types of sub-flows. In this paper, we propose a classifiable sub-flow selection method to realize the ideal classifier. The experimental results are so far promising for this research direction, even though they are derived from a reduced set of general applications and under relatively simplifying assumptions. Altogether, the significant contribution is indicating the feasibility of the ideal classifier by selecting not only initial sub-flows but also transition sub-flows.

Keywords—Mobile IP Network, Traffic Classification, Network Management, Traffic Engineering, Machine Learning

1. INTRODUCTION

Rapid progress of wireless communication technologies has opened new possibilities to achieve mobile IP networks, and they are one of the most sought-after networks of the next generation. The mobile IP networks are composed of a variety of elements (e.g. nodes and networks), and they are capable of seamless mobility. The key benefit of seamless mobility is that existing connections are maintained even though the elements change their points of attachment to the Internet. Thus users can be seamlessly provided with services irrespective of their movement.

To manage the point of attachment for a variety of elements in mobile IP networks, traffic classification is an essential task. The aim of traffic classification is to associate observed traffic with a specific application, and the classification results are used for profiling network usage and controlling the traffic under institutional policies (e.g. filtering, shaping, and priority routing).

Most of the classical classifiers involve direct inspection of each packet's header [1] and/or payload [2, 3]. Thus effectiveness of these classifiers has diminished. Header-based classifiers

Manuscript received April 30, 2010; accepted August 23, 2010.
Corresponding Author: Akihiro Satoh
* Graduate School of Information Sciences, Tohoku University, Japan (satoh@ka.riec.tohoku.ac.jp)
** Research Institute of Electrical Communication, Tohoku University, Japan
*** Cyberscience Center, Tohoku University, Japan

rely on well-known ports defined by IANA [4], but they cannot classify traffic of new applications that use either non-native port numbers or encapsulated packets. Payload-based classifiers rely on application specific signatures in the payloads, but they have the following problems: (1) Heavy operational costs, such as regular updates for tracking minor changes in an application's packet payload formats, are imposed upon administrators; (2) High computational costs, which are necessary for decoding every packet's payload traversing networks; (3) Ability of third parties to lawfully inspect packet's payloads has been constrained by government privacy regulations.

The above-mentioned limitations in the classical classifiers have motivated many researchers to use initial sub-flow features for the traffic classification [5, 6]. An initial sub-flow means initial *N* consecutive packets taken from a communication between two elements, and the features are statistical patterns, such as packet size, packet inter-arrival time, and packet order, in externally observable packets composing the flow. Initial sub-flow features based classifiers associate observed traffic per flow with a specific application by Machine Learning (ML) algorithms, and consequently traffic classification is achieved without direct inspection of each packet's header and/or payload.

However, these classifiers cannot classify traffic effectively in mobile IP networks. The classifiers depend on initial sub-flows, but they cannot always capture the sub-flows at the point of attachment for a variety of elements to the Internet because of seamless mobility. Although a few classifiers have been used to solve the problems, they have reached only cursory traffic classification (i.e. bulk traffic, high-bitrate and low-bitrate real-time traffic) [7] and extremely limited traffic classification (i.e. UDP-based game traffic and others) [8, 9] without initial sub-flows. Thus the ideal classifier should be capable of traffic classification based on not only initial sub-flows but also various types of sub-flows.

In this paper, we mainly focus on attainments of the ideal classifier to manage the point of attachment in mobile IP networks. We define application behaviors, and analyze the sub-flow features in consideration of the application behaviors. On the basis of the analytical results, we propose a classifiable sub-flow selection method, and the method selects transition sub-flows by the generation of a data packet sequence. We evaluate effectiveness of the proposed method through experiments on real traffic traces collected at an edge network in the Internet, and the experimental results are so far promising for this research direction, even though they are derived from a reduced set of general applications and under relatively simplifying assumptions. Altogether, the significant contribution is indicating the feasibility of the ideal classifier by selecting not only initial sub-flows but also transition sub-flows.

The remainder of this paper is organized as follows: In Section 2, we discuss related works and their limitations in mobile IP networks. In Section 3, we describe datasets of traffic traces used in analyses and evaluations. We propose a sub-flow selection method based on the analytical results in Section 4, and design our classifier in Section 5. Then we evaluate the accuracy of our classifier in Section 6. Finally, we conclude this paper in Section 7 with a summary of both contributions and future works.

2. RELATED WORKS

In this section, we describe concepts of mobile IP network and flow features based classifiers.

According to the background, we discuss limitations of the existing classifiers.

2.1 Mobile IP Network

In traditional IP networks, when a node moved from one network to another network, all the connections would have to be restarted. To overcome this problem, concepts of mobile IP networks have been proposed. The mobile IP network is defined as a network composed of a variety of elements, which are capable of seamless mobility, meaning, communications without interruption caused by their movements. The elements are not only nodes (e.g. computers and mobile phones) but also networks (e.g. personal mobile networks and vehicle networks).

Mobile IPv6 (MIP) [10] and Network Mobility (NEMO) [11] have been considered fundamental technologies to achieve mobile IP networks. The key benefit of these technologies is that the existing connections are maintained even though the element changes the point of attachment to the Internet. MIP and NEMO have similar mechanisms except that MIP provides nodes mobility whereas NEMO provides network mobility. For this reason, we describe MIP in detail here.

Figure 1 shows a Mobile Node (MN) which is moving from Network A to Network B and communicating with the Correspondent Node (CN) in MIP. The MN has two types of addresses: one is a permanent fixed address, called the “Home Address” (HoA), and the other is a temporary address called the “Care of Address” (CoA). HoAs are assigned to the MN from the Home Agent (HA) in the Home Network (HN), that is, the network to which the MN is originally attached.

In the communication between the MN and CN, all the messages must pass through the HA. When the CN communicates with the MN, it sends the packets to the MN’s HoA. In the HN, the HA detects those packets. If the MN doesn’t exist in the HN, the HA encapsulates the packets and sends them through a tunnel (i.e. IP-in-IP tunnel) to the MN’s CoA. When the MN communicates with the CN, the process is reversed: the MN sends the encapsulated packets through a tunnel to the HA, which de-encapsulates them and sends them to the CN.

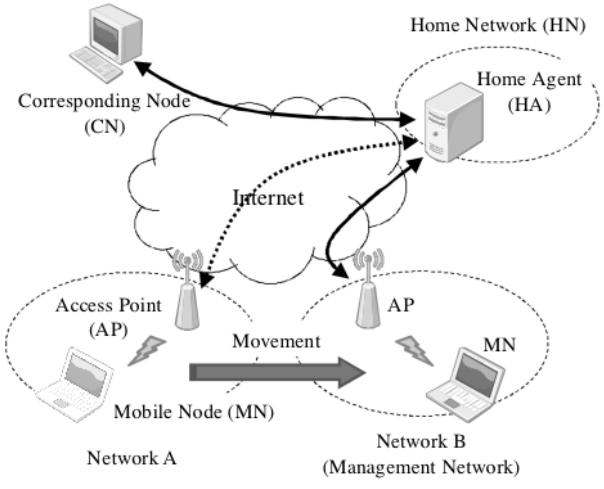


Fig. 1. Communication between two nodes in Mobile IPv6 (MIP)

2.2 Flow Features based Traffic Classifiers

In the explanation of flow features based classifiers, we should define the following three terms relating to flow:

- **full-flow** — a bi-directional flow captured over its entire lifetime, from the establishment to the finish of the communication between two elements with the same five-tuple, i.e. the source and destination IP addresses, port numbers and protocol number;
- **sub-flow** — *N* consecutive packets taken from a full-flow;
- **initial sub-flow** — initial *N* consecutive packets from the point where communication was established.

Figure 2 illustrates an overview of a general flow features based classifier, and the classifier consists of two phases: a training phase and a classification phase.

The training phase includes sub-flow selection function, flow features calculation function, and training function. First, the sub-flow selection function extracts all initial sub-flows from a training dataset which contains traffic traces of only target applications. Next, the flow features calculation function calculates features (e.g. packet size, packet inter-arrival time, and packet order) from each initial sub-flow, and represents them numerically. Finally, the training function derives the classifier model from the features of all initial sub-flows by ML algorithms.

In the classification phase, the flow features calculation function calculates the features of an initial sub-flow from new traffic, and feeds the classification function with them. The classification function compares the flow features with the classifier model, and outputs the classification results which define the new traffic to be deemed a member of a specific target application.

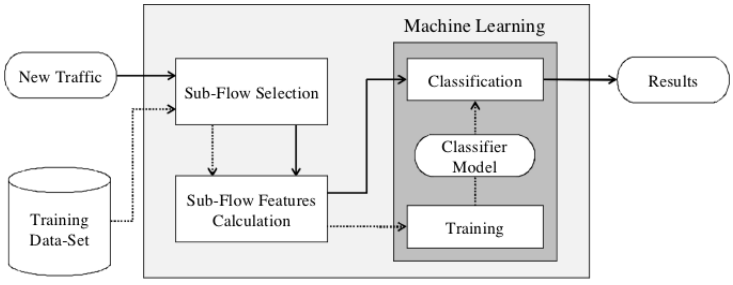


Fig. 2. Overview of a general flow features based classifier

2.3 Related Works and Their Limitations

Recently there have been many different works in the field of traffic classification [12], and the majority of them present classifiers based on initial sub-flow features [5, 6]. The motivation for using the initial sub-flow features depends on two observations: (1) Different applications typically have distinct features of the sub-flow which contains packets of a negotiation process; (2) The sub-flow can be captured as soon as a communication begins between two nodes. For these reasons, traffic classification is achieved without direct inspection of each packet's header and/or payload under the certain time limitations.

Most published works assume that the initial sub-flow is captured and available for traffic classification. However, this assumption is not accepted in mobile IP networks because of seamless mobility. In cases involving a changing point of attachment caused by movement of an element within a communication, the classifiers begin to capture the flow at a point in time when the flow is already in progress. As a result, they cannot capture the initial sub-flow. In a few works, classifiers have been used to solve these problems, and they have reached cursory traffic classification (i.e. bulk traffic, high-bitrate and low-bitrate real-time traffic) [7] and extremely limited traffic classification (i.e. a UDP-based game traffic and others) [8, 9] without initial sub-flows. Although they attain some positive results, the results cannot be precise enough to allow both fine-grained traffic control and detailed network usage profiles. Thus the ideal classifier should be capable of traffic classification based on not only initial sub-flows but also various types of sub-flows.

3. DESCRIPTION OF TRAFFIC TRACES

In this section, we describe datasets of traffic traces used in the analyses of Section 4 and evaluations in Section 6.

We collected traffic traces at a traditional IP network. However, any full-flow included in the traces was configured to be treated as a flow in mobile IP networks by removing initial *M* packets from it. We used these traffic traces instead of those from an actual practical mobile IP network for the following two reasons: (1) There were no practical mobile IP networks available for our measurements at the time of experimentation; (2) Although it is possible to construct a pint-sized mobile IP network, it would be impractical in capturing an adequate amount of traffic in the network.

Table 1 presents a summary of the two datasets (i.e. T1 and T2) both of which consist of traffic traces collected at the same observation point. Specifically, we installed a high speed monitoring box [13] to capture traffic through the observation point, and the point was connected to the Internet with a full-duplex Gigabit Ethernet. The datasets T1 and T2 were traffic traces observed over one month from April 1 UTC 2007 and one month from June 1 UTC 2007, respectively.

To establish a reference point in analyses and evaluations, we firstly extracted all full-flows with the same five-tuple from each dataset. Secondly, the full-flows were investigated by both well-known port number and deep packet inspection tool, and then they were given corresponding application labels (i.e. HTTP, SMTP, POP3, IMAP, and OTHERS) according to our decision about target applications. Finally, we removed all TCP control packets (i.e. SYN, FIN, or

Table 1. Summary of two datasets

	T1		T2	
Application	Flows	Volume	Flow	Volume
HTTP	78635	1.5GB	88207	1.7GB
SMTP	102349	150MB	90802	120MB
POP3	34454	95MB	20740	80MB
IMAP	20560	70MB	17004	85MB
OTHERS	216680	2.5GB	310080	3.0GB

ACK with no data) from the datasets because exchanges of these packets were application-independent.

4. ANALYSIS AND PROPOSAL

In this section, we define transition sub-flows in consideration of application behaviors, and analyze the effectiveness of their sub-flow features for traffic classification. On the basis of the analytical results, we propose a sub-flow selection method, and describe details of the method.

4.1 Definition of Application Behaviors and Transition Sub-Flows

An application can be expressible as a finite state machine driven by an event. A transition is a state change triggered by a particular input event. Each application starts from an initial state, and accepts various events as instructions of transition. Whenever the application accepts the event, it changes to the next state and executes actions associated with the transition. We define these states as application behaviors, and the relationships among applications and their behaviors are depicted in Table 2.

To realize the ideal classifier, we assume the effectiveness of a transition sub-flow, which is defined as; N consecutive packets taken from a transition point of application behavior. The rationale behind the possibility of the ideal classifier is the following. An application behavior consists of two or more packets, and the packets are categorized into two types: "control" and "data". The former packet is pre-defined as a specific message in each application, and the latter packet is a fragment of user-requested data such as a web page and/or mail body. The features of transition sub-flows have the ability to distinguish each application, because the sub-flow contains sequence exchanging control packets.

Table 2. Relationships among applications and their behaviors

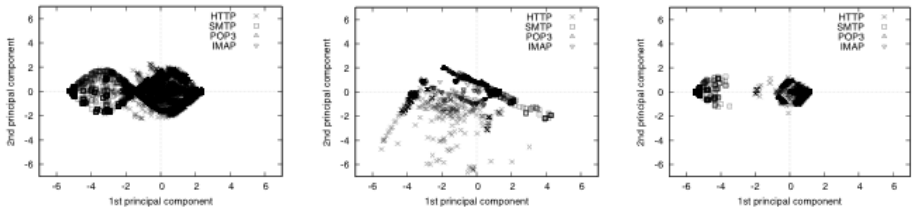
Application	Application Behaviors
HTTP	get, post, Java Applet/Active X
SMTP	authentication, send e-mail
IMAP	handle e-mail and directory

4.2 Analysis of Sub-Flow Features

To verify our assumptions described in Section 4.1, we consider the following three types of sub-flows and investigate their features:

- **all sub-flows** — all sub-flows taken from a flow;
- **transition sub-flows** — each sub-flow including the transition point of application behavior;
- **data sub-flows** — each sub-flow composed of data packets only.

To employ features, i.e. packet order, packet size, and packet direction, a sub-flow x composed of N packets is represented as a N-dimensional vector: $\boldsymbol{x} = (s_1(x), s_2(x), \ldots, s_n(x), \ldots, s_N(x))$. The absolute value of $s_n(x)$ denotes the payload size of the n-th packet in the sub-flow x and the sign of $s_n(x)$ represents the direction of the n-th packet (i.e. $s_n(x)$ is positive for upload packet and negative for download packet).



(a) all sub-flows (b) transition sub-flows (c) data sub-flows

Fig. 3. Three types of sub-flow features

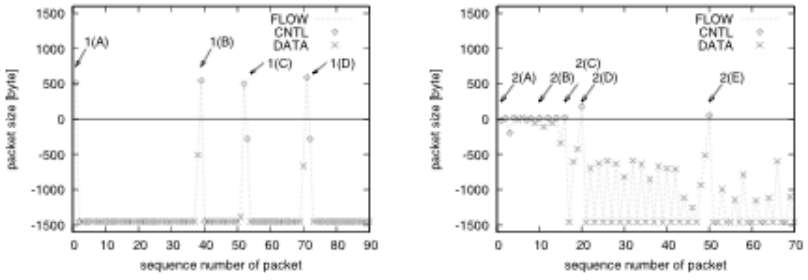
To investigate the features of the three types of sub-flows, we took sub-flows (N = 5) from ten full-flows of each application in T1, and made a set of sub-flows for each type. Figure 3 shows the features of the three types of sub-flows. In Figure 3, for each type of sub-flow set, the two most significant components were determined by PCA (Principal Component Analysis) [14], and their values were plotted on a scatter chart.

Figure 3(a) shows the scatter chart of all sub-flows. In this chart, the sub-flows of each application form a cluster; however, these clusters mostly overlap each other. These results tell us that a classifier trained on features of all sub-flows may have trouble to classify the observed traffic in a specific application, thus an appropriate sub-flow selection is indispensable for making effective classifiers. Figure 3(b) shows the scatter chart of transition sub-flows. In this chart, the sub-flows of each application also form a cluster; furthermore, their overlaps are smaller as compared to Figure 3(a). It illustrates that appropriate selection of the sub-flow can remove unclassifiable sub-flows from all sub-flows. These results motivate us to select the transition sub-flow for traffic classification. Figure 3(c) shows the scatter chart of data sub-flows. In this chart, the sub-flows form only two clusters caused by the sub-flow directions. These results indicate that the features of data sub-flows do not depend on the type of application.

4.3 Analysis of Application Behaviors and These Features

We analyzed the full-flows of several applications, and clarified the relation between features and behaviors in consideration of sequence number. In order to analyze them, we adopted visualization techniques described in literature [15].

Figure 4 shows full-flow features with application behaviors in HTTP and IMAP (due to space limitations, we omit the results of other applications). Each packet is represented by a point in two dimensions, where the X-coordinate corresponds to the sequence number of a packet, and the Y-coordinate represents the packet size and the direction. Positive Y values indicate packets of upload direction, and negative values of Y indicate packets of download direction. In either case, the magnitude of the Y-coordinate gives the packet's size in bytes. For example a point at (10, -500) means that a packet of download direction and 500 bytes in length, arrived 10-th after the start of the flow. Furthermore the arrows indicate the transition points of application behavior in Figure 4. Their labels denote the following behaviors: 1(A), 1(B), 1(C), and 1(D) are "get web-data"; 2(A) is "authentication", 2(B) is "display mailbox's list", 2(C) is "select mailbox", and 2(D) and 2(E) mean to "fetch e-mail".



(a) HTTP (b) IMAP

Fig. 4. Flow features with application behaviors in each application

The figures illustrate that the transition point of an application behavior appears before and after the sequence transmitting data packets and the sequence is composed of transmitting uni-directional consecutive packets. Furthermore, the transition point appears many times during the lifetime of a full-flow. Consequently, the transition sub-flows can be captured, even when the full-flow is already in progress.

4.4 Sub-Flow Selection Method

We propose a classifiable sub-flow selection method, and the method solves the critical problem of how to select transition sub-flows for traffic classification in mobile IP networks. The idea being brought forward in this research is to select the transition sub-flows, which are fragments of user-requested data in each application, by the generation of a data packet sequence. For this idea, we rely upon the following solid foundational understandings obtained from the analytical results in Section 4.2 and Section 4.3: (1) Transition sub-flow appears before and after the sequence transmitting data packets; (2) Features of data sub-flow have similarity among different applications because these features depend on the directions only; (3) Transition point of application behavior appears many times during the lifetime of a full-flow.

The method detects the start point and end point (i.e. two transition points) of a data packet sequence which is composed of transmitting uni-directional consecutive packets. Then transition sub-flows are selected by extracting *N* consecutive packets from the detected points.

5. DESIGN OF OUR CLASSIFIER

We build the transition sub-flow selection function based on the proposed method, and use this function as the sub-flow selection function in Figure 2. One of the advantages of the proposed method is that the functions of the existing classifiers can be reused, thus other functions (i.e. flow features calculation, training, and classification) are implemented by reference to a literature [16]. In this section, we describe the details of these functions.

5.1 Sub-Flow Selection

On the basis of the proposed method, the sub-flow selection function detects transition points,

Algorithm 1. Transition Points Detection

```
1:   input : x_full, pt_now
2:
3:   i ← pt_now
4:   j ← pt_now
5:   while d_i(x_full) = d_{j+1}(x_full) do
6:       j ← j+1
7:   end while
8:
9:   if j − i ≥ T_num then
10:      pt_start ← i
11:      pt_end ← j
12:      output : pt_start, pt_end
13:  end if
```

and selects transition sub-flows.

First, the function detects the start and end points of a data packet sequence and the points are considered as the transition points. According to the analytical results in Section 4.3, we simply define a data packet sequence as transmitting uni-directional T_{num} consecutive packets. The detection process is described in Algorithm 1, where $d_i(x)$ is the direction of the i-th packet in a flow x. The algorithm investigates packets through a full-flow x_{full}, and detects the transition points (i.e. the start point pt_{start} and end point pt_{end}).

Next, for each detected point, the function selects a sub-flow composed of N consecutive packets: for each start point pt_{start}, from the (pt_{start} - N + 1)-th to pt_{start}-th packets are extracted as a sub-flow, and for each end point pt_{end}, from the pt_{end}-th to (pt_{end} + N - 1)-th packets are extracted as a sub-flow. The selection process is described in Algorithm 2, where $p_i(x)$ is the i-th packet in a flow x.

Finally, the function outputs the selected sub-flows as the transition sub-flows, and the results are fed into the next function.

5.2 Flow Features Calculation

The flow features calculation function decides and calculates features of the sub-flow for numerical representation. Literature [17] introduces 249 available flow features to indicate fine-grained differences; however using the entirety of features is not always necessary. Since most of the features don't influence the accuracy of traffic classification, we should decide on some efficient features by abandoning the irrelevant and redundant ones. In Section 4.2, the analytical results show the possibility for traffic classification with a high degree of accuracy. Thus our classifier utilizes three features (i.e. packet order, packet size, and packet direction) and their numerical representations.

5.3 Training in Machine Learning

Depending on the representation, the training function derives classifier model for classification in machine learning. To measure a similarity between two sub-flows x_i and x_j, we use the classical metric, that is, euclidean distance: $dist.\ (x_i, x_j) = \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2$, where $\| \boldsymbol{x} \|_2$ means the 2-norm of vector $\boldsymbol{x}$. In order to extract common application behaviors in the N-dimentional space, we relied on the hierarchical clustering algorithm [18], because it is difficult to know the number of application behaviors beforehand. The algorithm aggregates sub-flows when similarity between sub-flow and cluster is less than threshold T_{class}. As a clustering result, an individual cluster corresponds to each application behavior. Ultimately, the function outputs a classifier model composed of the centroid of the cluster and the application label corresponding to it.

Algorithm 2. Sub-Flow Selection

```
input : x_full, pt

i ← pt
if d_i(x_full) ≠ d_{i+1}(x_full) then
    for j = 0 to N − 1 do
        if d_{i+1}(x_full) ≠ d_{i+j+1}(x_full) then
            p_{j+1}(x_sub) ← p_{i+j}(x_full)
        else
            break
        end if
    end for
    output : x_sub
else if d_i(x_full) ≠ d_{i−1}(x_full) then
    for j = 0 to N − 1 do
        if d_{i−j}(x_full) ≠ d_{i−j−1}(x_full) then
            p_{N−j}(x_sub) ← p_{i−j}(x_full)
        else
            break
        end if
    end for
    output : x_sub
end if
```

5.4 Classification in Machine Learning

The classification function associates sub-flow of new traffic with a specific application label by comparing classifier model, and outputs the classification results which define the new traffic to be deemed a member of a specific target application.

First of all, the function searches the classifier model to find the best fit for the sub-flow, and then selects which application is the most likely for the sub-flow given the set of target applications. At last, the function associates new traffic with a specific application. The classification results are used for profiling network usage and controlling the traffic under institutional policies (e.g. filtering, shaping, and priority routing), automatically.

6. EVALUATIONS

In this section, we evaluate the effectiveness of our proposed method through the experiments on the two datasets depicted in Section 3.

6.1 Accuracy Metrics

Three types of flows were obtained as the experimental results: correctly classified flows, misclassified flows, and unclassified flows. An unclassified flow means that it could not be classified by our classifier because no transition sub-flow could be selected from it. To evaluate the accuracy of our classifier, we employed three metrics:

true ratio — the ratio of correctly classified flows to all flows in each application;
false ratio — the ratio of misclassified flows to all flows in each application;
unclassified ratio — the ratio of unclassified flows to all flows in each application.

6.2 Methodology

We employed two datasets, i.e. T1 and T2, depicted in Table 1. The dataset T1 was used as a "training dataset" for creating classifier models, and the dataset T2 was used as a "testing dataset" for assessing the validity of our classifier. Our classifier relied on the Ward Clustering Algorithm [19]. This algorithm requires several conditional parameters: The number of packets in a sub-flow was $N = 5$; the two thresholds were $T_{num} = 2$ and $T_{class} = 1000$.

To evaluate the effectiveness of the proposed method for traffic classification in mobile IP networks, we carried out the following three steps: (1) We trained our classifier with 10000 samples which were initial and transition sub-flows of each target application (i.e. HTTP, SMTP, POP3, and IMAP) randomly chosen from a training dataset; (2) Our classifier classified all flows taken from the testing dataset into each target application, and then output the classification results; (3) The accuracy metrics were calculated from the classification results in consideration of missed packets caused by seamless mobility.

6.3 Discussion

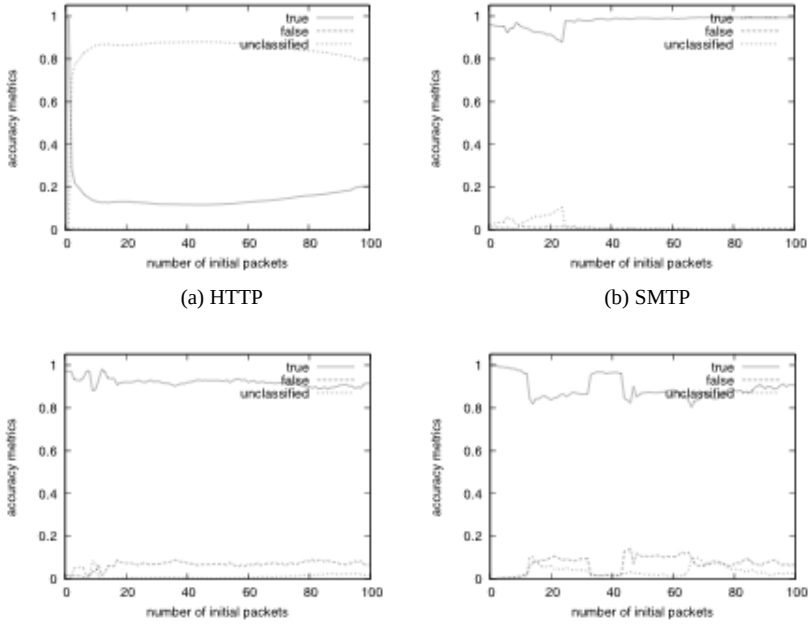
The experimental results are shown graphically in Figure 5, where the X-coordinate corresponds to the number of initial packets removed from each flow in the testing dataset, and the Y-coordinate represents the accuracy metrics (i.e. true ratio, false ratio, and unclassified ratio). The

number of removed initial packets was varied from 0 to 100 in all flows for each chart.

The true ratio of HTTP depicted in Figure 5(a) shows excellent results when the flow is captured from the beginning, but the true ratio rapidly drops below 0.2 if our classifier misses more than the first few packets. This decrease in the true ratio was caused by the increase in the unclassified flows. Our classifier selected only a transition sub-flow taken from the beginning point of a flow because many flows consist of only one application behavior. However, note that the results indicated a high degree of true ratio when our classifier could select the transition sub-flows taken from a flow in progress. Thus the proposed method is effective even though it is necessary to consider flows other than the transition sub-flows.

On the other hand, true ratios of the other target applications (i.e. SMTP, POP3, and IMAP) stay more than 0.8 (i.e. Figure 5(b), 5(c), and 5(d)) until the number of removed packets has increased beyond the early period of each flow. The results imply that our classifier does not have adequate ability enough to achieve practicable traffic classification. These errors were attributed to a shortage in training for rare application behaviors, that is, a mixture of dominant application behaviors and rare ones. Thus these errors can be reduced two ways: (1) We calibrate sub-flows chosen from the training dataset for the achievement of the uniform distribution among these application behaviors; (2) Our classifier adopts more intellectual ML algorithms for the function of training and classification.

It is worth noting that we decided on several conditional parameters (i.e. $N = 5$, $T_{num} = 2$, and $T_{class} = 1000$) through trial and error. Depending on the particular application, we were trying to



(a) HTTP (b) SMTP

(c) POP3 (d) IMAP

Fig. 5. Accuracy metrics of our classifier

classify, there would be a trade-off between keeping *N* low for timely classification and reduced memory consumption and keeping *N* high for acceptable accuracy. Other parameters similarly influenced many functions in our classifier. Accordingly, characterizing parameters is a subject for our future works.

The above-mentioned results are so far promising for our research direction, even though they were derived from a reduced set of applications and under relatively simplifying assumptions. Altogether, the significant contribution is indicating the feasibility of the ideal classifier by selecting not only initial sub-flows but also transition sub-flows. Consequently, we believe that the proposed method could have a profound effect on traffic classification in mobile IP networks.

7. CONCLUSIONS

In this paper, we have proposed a classifiable sub-flow selection method for the realization of the ideal classifier. Our experimental results so far have been promising for this research direction, even when they were derived from a reduced set of general applications and under relatively simplifying assumptions.

In future work, we will experiment in classifying traffic of various types of applications. On the basis of those experimental results, we intend to improve the sub-flow selection method repeatedly with the ultimate goal of being able to classify traffic of all types of applications in mobile IP networks.

REFERENCES

[1] D.E. Taylor, "Survey and Taxonomy of Packet Classification Techniques," ACM Computing Surveys, Vol.37, No.3, pp.238-275, Sep., 2005.

[2] V. Paxson, "Bro: A System for Detecting Network Intruders in Real-Time," Proceedings of the 7th USENIX Security Symposium, pp.31-51, Jan., 1998.

[3] H. Dreger, A. Feldmann, M. Mai, V. Paxson, and R. Sommer, "Dynamic Application-Layer Protocol Analysis for Network Intrusion Detection," Proceedings of the 15th USENIX Security Symposium, pp.257-272, Aug., 2006.

[4] "Internet Assigned Numbers Authority (IANA)," http://www.iana.org/assignments/port-numbers/.

[5] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic Classification through Simple Statistical Fingerprinting," ACM SIGCOMM Computer Communication Review, Vol.37, No.1, pp.5-16, Jan., 2007.

[6] L. Bernaille, R. Teixeira, and K. Salamation, "Early Application Identification," Proceedings of the 2006 ACM CoNEXT Conference, 2006.

[7] M. Tai, S. Ata, and I. Oka, "Environment-Independent Online Real-Time Traffic Identification," Proceedings of the 4th International Conference on Networking and Services, pp.230-235, 2008.

[8] T.T. Nguyen and G. Armitage, "Training on Multiple Sub-Flows to Optimise the Use of Machine Learning Classifiers in Real-World IP Networks," Proceedings of the IEEE 31th Conference on Local Computer Networks, pp.369-376, Nov., 2006.

[9] T.T. Nguyen and G. Armitage, "Clustering to Assist Supervised Machine Learning for Real-Time IP Traffic Classification," Proceedings of the IEEE International Conference on Communications, pp.5857-5862, May, 2008.

[10] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," Internet RFC 3775, Jun., 2004.

[11] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "Network Mobility (NEMO) Basic Support Protocol," Internet RFC 3963, Jan., 2005.
[12] T.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, 4th Quarter 2008.
[13] "Cpmonitor," http://www.cysols.com/.
[14] J. Shlens, "A Tutorial on Principal Component Analysis," Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 2005.
[15] C.V. Wright, F. Monrose, and G.M. Masson, "Using Visual Motifs to Classify Encrypted Traffic," Proceedings of the International Workshop on Visualization for Computer Security, pp.41-50, Nov., 2006.
[16] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamation, "Traffic Classification on the Fly," ACM SIGCOMM Computer Communication Review, Vol.36, No.2, pp.23-26, Apr., 2006.
[17] A.W. Moore, D. Zuev, and M.L. Crogan, "Discriminators for Use in Flow-Based Classification," Technical Report, Aug., 2005.
[18] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, Vol.31, No.3, pp.264-323, Sep., 1999.
[19] J.H. Ward, "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, Vol.58, No.301, pp.236-244, 1963.

Akihiro Satoh
Received his Master's degree in Information Sciences from Tohoku University, Japan in 2008. Now he is in the Ph.D. program in Graduate School of Information Sciences (GSIS), Tohoku University. His research interests include network management mechanisms, traffic analysis, etc. He is a Student Member of IPSJ.

Toshiaki Osada
Received his doctoral degree from the Graduate School of Information Sciences, Tohoku University, Japan in 2009. Presently he is a postdoctoral fellow of the Research Institute of Electrical Communication, Tohoku University. His research interests include routing technologies for mobile ad hoc networks and overlay networks, and network management. He is a member of IPSJ.

Toru Abe

Received his ME and PhD degrees in information engineering from Tohoku University in 1987 and 1990, respectively. From 1990 to 1993, he was a research associate at the Education Center for Information Processing of Tohoku University. From 1993 through 2001, he was an associate professor at Japan Advanced Institute of Science and Technology (JAIST). Currently, he is an associate professor at the Information Synergy Center of Tohoku University. His research interests include knowledge engineering and pattern recognition.

Gen Kitagata

Is an associate professor of the Research Institute of Electrical Communication of Tohoku University, Japan. He received a doctoral degree from the Graduate School of Information Sciences, Tohoku University in 2002. His research interests include agent-based computing, network middleware design, and symbiotic computing. He is a member of IEICE, IPSJ.

Norio Shiratori

Is currently a Professor at the Research Institute of Electrical Communication (RIEC), Tohoku University, Japan. Before moving to RIEC in 1993, he was the Professor of Information Engineering at Tohoku University from 1990 to 1993. Prior to that, he served as an Associate Professor and Research Associate at RIEC, Tohoku University, after receiving his Doctoral degree from Tohoku University in 1977. He also served as the vice Director of RIEC, and IFIP representative of Japan. Now he is the President of IPSJ (Information Processing Society of Japan). He is a fellow of IEEE, IPSJ and IEICE. He also contributes through serving in various capacities, such as: General Chair of the 9th IEEE ICOIN-9 (1994), IFIP Joint International conference FORTE/PSTV'97, and 12th IEEE ICOIN-12 (1997); Program Chair of ICPADS'96 (1996) and ICPP-99 (1999). He was one of the leaders in Japan Gigabit Network (JGN) national project and headed several national projects such as, SCOPE—funded by Ministry of Internal Affairs and Communications and Dynamic Networking project—sponsored by JSPS. He is currently leading two other national projects. Besides that, in 2006, the proposed idea of his research group on Mobile IPv6 was approved and standardized by IETF. He has proposed a new concept of Flexible Computing and is still working in this direction. His recent research interest is in Ubiquitous and Symbiosis computing. He has published more than 15 books and over 400 refereed papers in computer science and related fields. He was the recipient of IPSJ Memorial Prize Wining paper award in 1985, Telecommunication Advancement Foundation Incorporation Award in 1991, Best Paper Award of ICOIN-9 in 1994, IPSJ Best Paper Award in 1997, and many others including the most recent Outstanding Paper Award of UIC-07 in 2007.

Tetsuo Kinoshita

Is a professor at the Research Institute of Electrical Communication of Tohoku University. He received Dr.Eng. degrees in information engineering from Tohoku University, Japan, in 1993. His research interests include knowledge engineering, agent engineering, and knowledge-based/agent-based systems. He received the IPSJ Research Award, the IPSJ Best Paper Award and the IEICE Achievement Award in 1989, 1997 and 2001, respectively. Dr. Kinoshita is a member of IEEE, ACM, AAAI, IEICE, IPSJ, JSAI, and Society for Cognitive Science of Japan.