Journal of Theoretical and Applied Information Technology
31st July 2013. Vol. 53 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SSPC ALGORITHM BASED ON THREE DIFFERENT METHODS FOR ONLINE SKYPE TRAFFIC CLASSIFICATION

[1]HAMZA AWAD HAMZA IBRAHIM, [2] SULAIMAN MOHD NOR, [3] IZZELDIN IBRAHIM MOHAMED ABDELAZIZ, [4]ALI AHMED ALFAKI ABDALLA

[1,2,3] Universiti Teknologi Malaysia, Faculty of Electrical Engineering
[4] Universiti Teknologi Malaysia, Faculty of Computing
E-mail: [1]hamysra76@hotmail.com , [2]sulaiman@fke.utm.my , [3]izzeldin@fke.utm.my , [4]aliahmmed2003@yahoo.com

ABSTRACT

Classification of real time applications such as Skype and online games has gained more attention in the last few years. Most of the current Skype classification methods were only valid for offline classification. Each of the three common classification methods (port, payload, statistics based) has some limitations. To increase the quality of Internet traffic classifier, this paper combined the three methods to products a new classification algorithm (SSPC). In the proposed algorithm, each traffic flow was classified parallel three times by one of the three method classifiers. Based on some priority rules, SSPC makes classification decision for each flow. The SSPC algorithm was used to classify Skype traffic in two stages offline and online. The results of both cases shown, SSPC is the higher accuracy when compared with other classifiers. Also, the results indicate that the SSPC algorithm was suitable for online classification decision which is taken within capturing time.

Keywords: *Internet Traffic classification, Skype classification, Online classification, Machine Learning, classification Algorithm*

1. INTRODUCTION

Internet Server Provider (ISP) and network operators are usually interested to know the traffic carried in their networks for the purposes of optimizing network performance and security issues. Therefore, Internet traffic classification is something powerful, particularly interactive traffic applications such as Skype and online games.

Simple classification assumes that, most applications used well-known Port, and the classifier used this port number to determine the application type. However, most Internet applications used unknown port number or more than one application used the same port number, which indicates to failure of port base classification[1]. Another classification method is payload base (deep packet inspection), which is individual packet inspection, looking for specific signatures. However, using of this technique faced by two problems; first, it difficult to detect non-standard port by using packet inspection, because these packets were encrypted. Second, deep packet inspection touches users' privacy.

In order to solve the problem of past classification methods (base port and payload inspection), Machine Learning (ML) techniques were appeared. ML [2] [3] used artificial intelligence to classify IP traffic, which provide a good solution by extracting right information from application features [4]. Moreover, some of the ML algorithms are suitable for Internet traffic flow classification at a high speed.[5]. Because of the rapid sort of real time applications, the main issue when classifying interactive applications is the time of collecting the statistical values (build rules), which assumed to be extremely short.

Most of the proposed ML classification methods was limited for offline traffic classification and cannot support online classification [6]. Online classification means the decision of what is flow/packet belong to; assume to be on the time of capturing. Such like any hardware classifier (PacketShaper, SANGFOR) installed on network router, which is classifying with the passage of the traffic.

Over the last few years, Skype has gained significant attention and has become one of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

most common forms of VoIP software. According to the Skype website [7], Skype users in the last year spent 1.8 billion hours making video calls. Also, at certain times, more than 22 million users were logged onto Skype at the same time.

The main problem meets the Skype online classification decision it is the high speed of Internet traffic. It is difficult to get an online classification decision with huge of Internet traffic. So how to: divide the Internet traffic into flows, calculate flows patterns, and make classification decision online with high Internet traffic speed. Most previous literatures [6] [8] [9] [10] [11] [12] [13] [14] provide a classifiers work with real time traffic, put few of them [15] provided classifiers can get an online decision.

This paper aims to develop online Signature Statistic Port Classifier (SSPC) algorithm, which can identify Skype traffic shortly after capturing time. Our classifier differs from other works since it take the classification decision based on three parallel different methods.

Section 2 describes and analyses the related works. General concepts of classification mechanisms, the three partial algorithms, and SSPC algorithm are discussed in Section 3. In section4, the experiments and results analyzing were illustrate. Finally, the conclusion was provided in section 5..

2. RELATED WORK

In this section, we will address the related works from two points of views; the articles of online classification methods and articles of Skype classification methods.

In [16], the authors claimed that Skype traffic can be identified by observing five seconds of a Skype traffic flow. The classifier achieved more than 98% accuracy and succeeded in identifying suitable traffic features to classify Skype. However, the method and datasets are used only for offline classification. The offline detection has several shortcomings when compared with real online classification.

The authors of [8] focus on traffic measurement in high speed network. The paper analyzes Internet applications to see the traffic measurement characteristics. Then design flow measurement system of high speed network based on Linux kernel. The system was built over some methods; firstly, from a perspective of Network Interface Card (NIC); the system designed a Hash function (group of rules) to classify packet processed by 32-bit systems instead of interrupt to communicate with OS. Third, the system identify the new P2P service by calculate key hash value if there are no existing matching rules. The system was tested in ReadHat9.2 operating system. Some shortness was observed such as The author does not detail what features are used to builds hash rules. As well, the system defines any new traffic flow as new P2P. Moreover, the paper relies majorly on port numbers to identify traditional applications.

The paper [9] proposed a dynamic online method to classify Internet traffic. The method used the concept of two levels: overall traffic level and application level. Data stream mining algorithms are used to continue updated considered datasets. The proposed method has three parts: i) Traffic model; which is prepared datasets, select features, and update model in case of new application. ii) Traffic classification; to classify traffic based on gained features. iii) Change detection; which is run periodically to check if there is a new application. While the paper title includes the words “online traffic classification”, but there is no online classification. Something else, no details about traffic features used for classification.

The study [10] proposed approach for online classification for TCP traffic based on the first n packets. The approach used information from first n packets, and Bayesian network method to decide which kind of application the flow was belongs to. The authors used correlation-based feature selection (CFS) [17] to select optimum features. However, it is something untrusted to classify flow include thousands packets based on the first few packets. This because of, the first packets in many flows can differ from the rest packets statistically. Moreover, the paper did not detail how the online decision was taken.

In [15], the authors proposed a network processors (NPs) classifier, which is based on online hybrid traffic to identify P2P traffic. The classifier is based on two stages: hardware static characteristics and software Flexible Neural Tree (FNT) [18]. In the first stage, the hardware classifier (based on payload and port) filters P2P traffic. In the second stage, the software classifier (based on ML statistical features) is used as statistical diction maker. While the authors theoretically disagree with port and payload classification, they depend so much on both to classify P2P traffic.

In [11], traffic classifier based on Support Vector Machine (SVM)was presented. The dataset include

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

three traces, which collected from three different places. Based on statistical features, the classifier used the first ten packets to identify the flow. Like previous; while the paper title content the words "online classification", but there is no online decision. Also, how to classify flow includes more than 4,600,000 packets based on only ten packets.

The researchers of [12] propose a wireless mesh network traffic classification using C4.5. Sub-flow and application behaviors were applied to results represent a sub-flow. Based on the statistical features of the first n packets, the classifier clusters the flow to one of the defined applications. Similar to the previous, the datasets were captured at real time; however, there are no online classification (capture and classify at the same time)

[14] is a flattering work which proposes a method which suitable for identifying the application association with a TCP flow. This based on total data length sent by client (ACK-Len ab) or server (ACK-Len ba) before it received ACK packets. The work analysis TCP flows to get the characteristics of the two adopted features (ACK-Len ab and ACK-Len ba). The proposed method was verified by using ML classifier (C4.5) to classify four types of Internet application (WWW, FTP, EMAIL and P2P). With the same manner, no online classification was approved.

[13] develop a classifier which quickly identifies an application at any point of a flow's lifetime. Thus, the ML classifier was trained by using sets of features calculated from multiple sub-flows at different points. The classifier recognizes the flow either way (forward or backward) by features swapped called Synthetic Sub-flow Pairs (SSP). Assistance of Clustering Techniques (ACT) as unsupervised clustering ML technique was used to automate the selection process. The problem is different datasets from different dates (may be different network) was considered for ML. This is not consistent with the rule of similarity of training and testing datasets network environment.

[19] is recent work, which is proposed multistage classifier. Binary Particle Swarm Optimization (BPSO) is a method applied by this work to select the best flow features. Three methods (port, payload, and statistical based) integrated into multistage classifier. The idea of this work is extremely good; however, it was not tested as online classification which identifies traffic with capture speed. Another shortness is that the classifier can make his decision only based on the first stage (port based method).

3. ONLINE INTERNET TRAFFIC CLASSIFICATION MECHANISM

3.1 General concepts

Definition1 (Flow) is a group of packets share the same 5-tuples (source address, destination address, source port, destination port, and transport protocol). Flow can represent TCP or UDP packets. We consider unidirectional flows, which is defines client server traffic as different from server client traffic. **Definition2 (real time traffic)** it is Internet traffic captured from our campus network during the period of experiments. **Definition3 (offline decision)** it is classifier decisions about the flows identification, which is taken offline after capturing time. **Definition4 (online decision)** is classifier decisions about the flows identification, which is taken online within capturing time. With existence of continuous development of Internet applications, It is difficult to classify the traffic by using only one classification method [19]. This paper develops online Signature Statistical Port Classifier (SSPC), which is making classification decision in time very near to the capturing time. The classifier makes his final decision based on three parallel partial decisions (port classifier, signature classifier, and statistic classifier)

Port classifier

As mentioned in section 1, port based classification cannot achieve a high accuracy all the time. In this paper port classification was used as a part of our classification system and it represent low priority of SSPC classification decision. In most cases, SSPC classification decisions not making based alone on port classifier, but it shares the decision with the other two classifiers. We develop port classifier algorithm as a part of SSPC algorithm. Port classifier makes his own decision based on port DataBase. Easley, the port classifier algorithm compares the port number of the flow with the ports DataBase. If found then the flow will classify based on port classifier rules.

Signature classifier

Payload classification can achieve high accuracy, but it cannot work with encrypted traffic. As before, SSPC did not fully depend on payload, but it does only represent a part of the final decision. We develop signature classifier algorithm which is the second part of SSPC algorithm; this algorithm take classification decision based on some saved signatures. We add some general signatures (such as DNS query and http host) for the considered applications, which are extracting from the application layer. If the flow has a signature from

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the signature data base, it will classify based on what signature belong to.

Statistic classifier

The main problem meats ML classification is the high false positive. To reduce this problem we consider two issues; firstly, the offline training datasets was continuously updated and collected manually from the same network we need to classify. Secondly, Statistic classifier was supported by the other two classifiers. Statistic classifier algorithm is the third part of SSPC algorithm. Also as before, ML classifier represents a part of the system decision.

SSPC

In the purpose of increasing the classification efficiency, SSPC is proposed. SSPC is a result of the three previous classifiers decisions. Differing from the previous works [15] and [19], SSPC did not base on hardware part; also he did not take his decision based only on one method. The online flow classification was occurred after comparison of three stage classification decisions. Moreover, SSPC was tested for online classification decision.

3.2 SSPC Architecture

Figure 1illsturate the classifier stages, which started by fully packets capturing using traffic mirror. Before delivered to the three classifiers, the traffic was divided into flows based on the 5-tuples. Each flow will classify three times by each of the three classifiers. The port classifier compares the captured flow ports with a list of saved port numbers. If the captured flow is belonged to any group of saved ports, it will identify as its group as. The second classifier (statistical) works parallel with the first classifier. Based on offline training and testing datasets, some classification rules were building. Based on these rules, the statistical classifier (algorithm) makes his online decisions to identify the captured flows. On the other hand, the signature classifier will classify the same traffic at the same time of the previous two classifiers; the classifier will compare a part of captured flow with signature data base. If the signature matches any of saved signatures, the classifier will take his online decisions to identify the captured flows. SSPC is an algorithm which compares between the three classifiers result and makes his online classification decision based on some priorities rules.

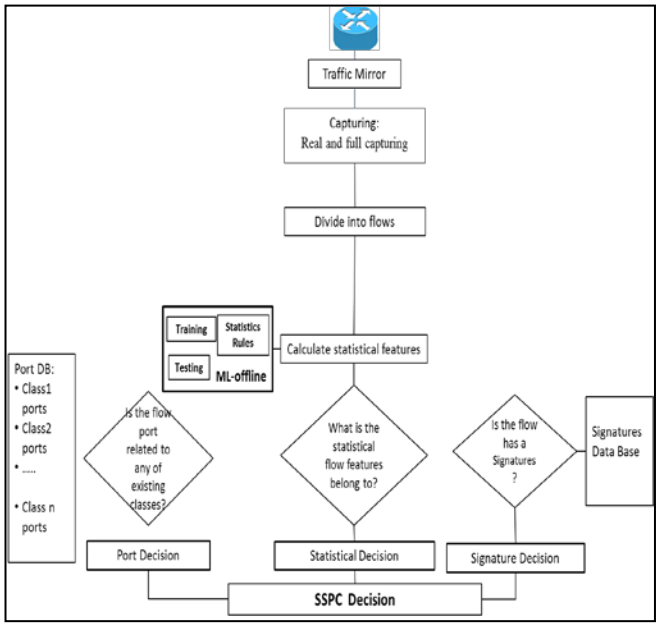

Figure 1 SSPC Architecture

3.3 SSPC Algorithm

The SSPC algorithm was shown below; Matlab version 7.5.0.342 (R2007b) was used to develop the algorithm. The SSPC algorithm consists of three partial classifier algorithms (described in section 4.2); each classifier has own classification decision. Because of accurate of signature classifier, the first priority of SSPC decision goes to signature classifier. If signature classifier makes any decision about this flow, then SSPC decision will equal to signature classifier. The second priority of SSPC happen in case of all partial classifiers has no decision about this flow. In this case, SSPC will classify the flow as unknown. The third priority of SSPC occurs in case of statistics and port classifiers have the same decision and signature classifier have no decision (unknown). In this case, SSPC identifies this flow based on Statistic and port classifiers. When statistic and port classifiers have different opinions about the flow, SSPC will classify this flow as statistic classifier as. The SSPC decision was built based on port classifier in only one case. This occurs when port classifier has a decision and both statistic and signature classifiers have no decision about the flow.

```
1    // Define variables
2    Array port_DataBase;
3    Array signatures_DataBase;
4    string statistical_rules;
5    start packets capturing;
6    //divide captured packets into flows
7    if the packet belongs to an existing flow
8    then adds this packet to the existing flow
9    else
10   initializes a new flow;
```

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
11  end if
12  //run the three algorithm classifiers
13  calculate flow statistic_values;
14  update statistic_rules;
15  for (flow=0; flow=-1;flow++)
16  {
17  // check signature
18  inspects n packets in the flow;
19  if class_signatures found
20  decision1=classify this flow according to
        signatures_DataBase;
21  else
22  decision1=classify this flow as unknown;
23  end if
24  //check statistical
25  if statistical_of_the_flow achieved any statistic_rules
26  decision2=classify this flow according to
        statistic_rules;
27  else
28  decision2=classify this flow as unknown;
29  end if
30  // Check port
31  if flow_port in port_DataBase
32  decision3=classify this flow according to
        port_DataBase;
33  else
34  decision3=classify this flow as unknown;
35  end if
36  end if
37  // calculate SSPC decision
38  if decision1 != "unknown"
39  SSPC_decision= decision1;
40  else if decision1 = decision2= decision3="unknown"
41  SSPC_decision= ="unknown"
42  else if decision2 = decision3
43  SSPC_decision= decision2;
44  else if decision1 = decision2= "unknown"
45  SSPC_decision= decision3;
46  else
47  SSPC_decision = decision2;
48  end ;
49  }
```

4. EXPERIMENTS AND ANALYSIS

In order to evaluate our methodology, several experiments were done. Real time Skype and non-Skype traffic was collected from campus network. Table 1 show the considered flows for both classes, which are run manually through the monitored clients (IPs). By this way we ensure the training and testing datasets collected from same network and no need to used standard datasets. Skype traffic was generated by real communication sessions (call) between Skype Clients (SC), which are located in and out campus area. Non-Skype traffic includes http, https, FTP control, FTP data, and online game (LOL).

Table 1 considered flows for offline decision

Class (applications)	Number of the flows
Skype	2044
Non-Skype	2213

For ML training purpose, we capture traffic from some monitored clients. Offline ML classification was done to select the optimum features and algorithm. After some filtering, rule.PART algorithm within Weka [20] was selected as ML classifier; rule.PART rules were built into our statistic classifier algorithm. With the same filtering, Interarrival time and packets length (size) are used as traffic features. From these two features, some statistics factor was calculated which are shown in table 2.

Table 2 Selected features

Max of Interarrival time
Min of Interarrival time
Mean of Interarrival time
Variance of Interarrival time
Standard deviation of Interarrival time
Max of packet length
Min of Packet length
Mean of Packet length
Variance of Packet length
Standard deviation of Packet length

Before going into online decisions experiments, offline decision works was performed to make sure of the methodology. First; each classifier was run over each class dataset (table 1) separately. The result of each case was recorded. Second; by the same manner, SSPC algorithm was run over each class dataset separately. Table 3 and figure 1 show the classifiers accuracy and SSPC accuracy. For each class of the considered datasets, SSPC shows the higher accuracy compared with the other partial classifiers.

Journal of Theoretical and Applied Information Technology
31st July 2013. Vol. 53 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3 Offline Classifiers Results

	Signature	Port	Statistics	SSPC
Skype	2.35%	0.78%	87.23%	88.01%
Non-Skype	11.43%	74.74%	91.28%	93.40%

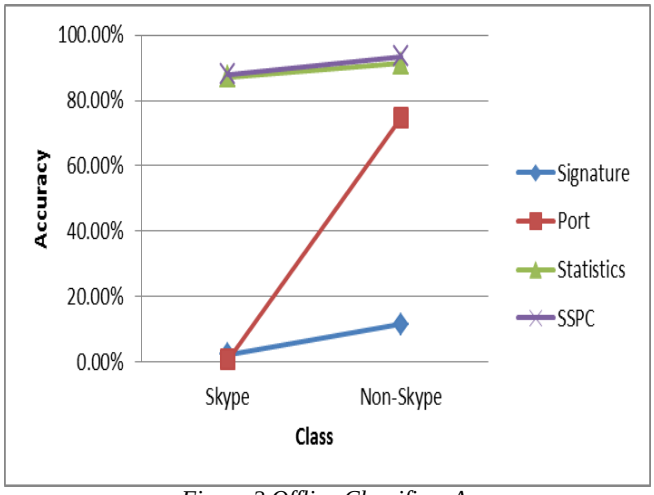

Figure 2 Offline Classifiers Accuracy

For the online decision, the same offline applications through two different experiments were considered. Same like offline decision; the applications were run in the monitored clients, which is generates dataset totally different from the training dataset. As an example in some clients, we run only www applications and then check parallel (at the same time) what is the decision of each classifier and what is SSPC decision. Table 4 shows number of flows generated by each of online experiment. Table 5 and figure 3 illustrate the accuracy of online decisions. In non-Skype, SSPC is higher accuracy in both experiments when compared with the other three classifiers. On other hand, SSPC have the higher accuracy between the other classifiers when we deal with Skype classification. The last column in table 5 shows the average of classification time (in seconds) for each flow. As an example classifying single Skype flow in experiment 1 was taken 0.06 second after end of flow capturing.

Table 4 Number Of Flows For Online Decisions

	Experiment 1	Experiment 2
Skype	217	44
Non-Skype	1911	1494

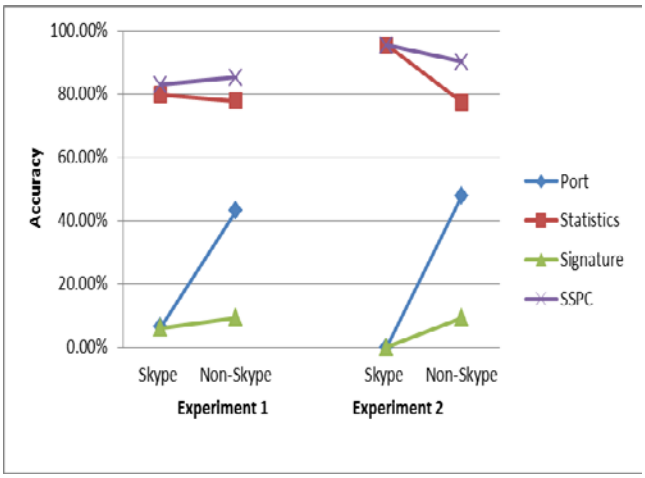

Figure 3 Online Classifiers Accuracy

5. CONCLUSION

Port based classifier has the advantage of non-complexity; however it cannot achieve high accuracy with applications of unknown port numbers. On the other hand, payload classifiers have the advantages of the accurate, but incapable to encrypted traffic. Also, statistic classifier has the benefit of classifying encrypted traffic, but it has the problem of high false positive. In this paper, Signature Statistical Port Classifier (SSPC) algorithm for online traffic classification was proposed. In parallel, each of the three partial classifiers (based on the three methods) makes decision about each traffic flow. The SSPC algorithm calculates the final decision of the three classifiers based on some priority rules.

Table 5 Online Classification Results

	Experiment 1				
	Port	Statistics	Signature	SSPC	flow/S
Skype	6.45%	79.72%	5.99%	82.95 %	0.06
Non-Skype	43.09%	77.81%	9.33%	85.26 %	0.063
	Experiment 2				
	Port	Statistics	Signature	SSPC	flow/S
Skype	0%	95.45%	0%	95.45%	0.07
Non-Skype	47.81%	77.29%	9.30%	90.14%	0.06

The proposed algorithm was used to distinguish between Skype and non-Skype traffic. Real time datasets (more than 7900 flows) were captured from

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

campus environment, which includes: non-Skype (http, https, FTP-data, FTP-control, online game), and Skype. The SSPC was tested into two stages offline and online. The results of the offline experiments show that the SSPC has higher accuracy among the three classifiers. For more validation, online classification was executed, the results were gotten shortly after each end of flow capturing, which also show higher accuracy when compared with other classifiers. Thus the SSPC achieved two objectives; first: increased in any of partial classifiers' efficiently can increase SSPC accuracy. Second: SSPC can classify non-well port and encrypted traffic, also can increased the accuracy of statistics classifier.

REFERENCES

[1]. 1. Nguyen, T.T.T., Armitage, G.: A Survey of Techniques for Internet Traffic Classification using Machine Learning. Ieee Commun Surv Tut 10(4), 56-76 (2008). doi:Doi 10.1109/Surv.2008.080406

[2]. 2. Jesudasan, R.N., Branch, P., But, J.: Generic Attributes for Skype Identification Using Machine Learning. Technical Report 100820A (2010).

[3]. 3. Alshammari, R., Zincir-Heywood, A.N.: An investigation on the identification of VoIP traffic: Case study on Gtalk and Skype. In: Network and Service Management (CNSM), 2010 International Conference on, 25-29 Oct. 2010 2010, pp. 310-313

[4]. 4. Yu, J., Lee, H., Im, Y., Kim, M.S., Park, D.: Real-time Classification of Internet Application Traffic using a Hierarchical Multi-class SVM. Ksii T Internet Inf 4(5), 859-876 (2010). doi:DOI 10.3837/tiis.2010.10.009

[5]. 5. Soysal, M., Schmidt, E.G.: Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Perform Evaluation 67(6), 451-467 (2010). doi:DOI 10.1016/j.peva.2010.01.001

[6]. 6. Gu, R., Wang, H., Ji, Y.: Early traffic identification using Bayesian networks. In: 2010, pp. 564-568

[7]. 7. Skype.website. www.skype.com (2012). Accessed accessed at 6/5/2012

[8]. 8. Xu, C., Tang, H., Zhao, G.F.: Traflow: Design and complementation of a real time Traffic Measurement System in High-Speed Networks. 2008 Ifip International Conference on Network and Parallel Computing, Proceedings, 341-344 (2008). doi:Doi 10.1109/Npc.2008.10

[9]. 9. Xu, T., Qiong, S., Xiaohong, H., Yan, M.: A Dynamic Online Traffic Classification Methodology Based on Data Stream Mining. In: Computer Science and Information Engineering, 2009 WRI World Congress on, March 31 2009-April 2 2009 2009, pp. 298-302

[10]. 10. Hong, M.-h., Gu, R.-t., Wang, H.-x., Sun, Y.-m., Ji, Y.-f.: Identifying online traffic based on property of TCP flow. The Journal of China Universities of Posts and Telecommunications 16(3), 84-88 (2009). doi:http://dx.doi.org/10.1016/S1005-8885(08)60231-9

[11]. 11. Gu, C., Zhang, S., Huang, H.: Online internet traffic classification based on proximal SVM. Journal of Computational Information Systems 7(6), 2078-2086 (2011).

[12]. 12. Gu, C., Zhang, S., Xue, X., Huang, H.: Online wireless mesh network traffic classification using machine learning. Journal of Computational Information Systems 7(5), 1524-1532 (2011).

[13]. 13. Nguyen, T.T.T., Armitage, G., Branch, P., Zander, S.: Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic. Networking, IEEE/ACM Transactions on PP(99), 1-1 (2012). doi:10.1109/tnet.2012.2187305

[14]. 14. Sun, M.F., Chen, J.T.: Research of the traffic characteristics for the real time online traffic classification. Journal of China Universities of Posts and Telecommunications 18(3), 92-98 (2011).

[15]. 15. Chen, Z.X., Yang, B., Chen, Y.H., Abraham, A., Grosan, C., Peng, L.Z.: Online hybrid traffic classifier for Peer-to-Peer systems based on network processors. Appl Soft Comput 9(2), 685-694 (2009). doi:DOI 10.1016/j.asoc.2008.09.010

[16]. 16. Branch, P.A., Heyde, A., Armitage, G.J., Acm: Rapid Identification of Skype Traffic Flows. Nossdav 09: 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video. Assoc Computing Machinery, New York (2009)

[17]. 17. A, H.M.: Correlation-based feature selection for machine learning. Waikato University (1998)

[18]. 18. Chen, Y., Yang, B., Dong, J.: Nonlinear system modelling via optimal design of neural trees. Int J Neural Syst 14(2), 125-137 (2004).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[19]. 19. Min, D., Xingshu, C., Jun, T.: Online Internet traffic identification algorithm based on multistage classifier. Communications, China 10(2), 89-97 (2013). doi:10.1109/cc.2013.6472861

[20]. 20. Witten, I.H., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques. Diane Cerra, (2005)