Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Online Statistical Internet Traffic Classification Approaches

Max Bhatia
University Institute of Engineering & Technology.
Panjab University, Chandigarh, India
maxbhatia.cse@gmail.com

Sakshi Kaushal
University Institute of Engineering & Technology.
Panjab University, Chandigarh, India

Abstract --- **Internet traffic classification into specific network applications is essential for managing network resources and from security point of view. Traditional classification techniques such as port based and payload based are having significant limitations. Hence newer statistical approach is adopted for classifying internet traffic into specific applications. This paper reviews modern statistical classification techniques which may or may not use Machine Learning approach, but still classify traffic with great accuracy.**

Keywords **--- online traffic classification, classification with Machine Learning, classification without Machine Learning, real time classification techniques.**

I. INTRODUCTION

Network traffic classification is one of the major challenging task in past few years. The tasks of network engineers include meeting application performance, meeting bandwidth requirement of customers, managing bandwidth consumption, apply security rules, fault diagnosis, performing accurate accounting for billing, etc. In order to accomplish all these tasks, it is necessary to understand network traffic properties, which would help to improve network performance by developing better architecture. Therefore, network traffic classification is of great importance. With the help of classification results, an enterprise or a service provider can protect and manage network resources. All traffic classification techniques use some metrics to evaluate the result. These classification techniques can be differentiated by using criterion known as predictive accuracy (i.e how correctly a technique can make decision about data). The common metrics which are used: False Negative(FN), False Positive(FP), True Negative(TN), True Positive(TP). A good classifier minimizes FN & FP. Some other evaluation metrics used are: Accuracy, Recall and Precision. The remainder of this paper is organized as follows: Section II presents traditional classification techniques used. Section III presents new classification technique used. Section IV presents related work done using new classification technique. At last, we conclude the paper in Section V.

II. TRADITIONAL TRAFFIC CLASSIFICATION TECHNIQUES

Traditional IP traffic classification relies on inspection of TCP or UDP port nos. (Port based classification) or reconstruction of protocol signatures from its payload (Payload based classification). Both suffered from no. of limitations with little advantages as well, which are discussed below.

A. Port based classification:

This approach matches port numbers to applications where an application is associated with a well defined port number (by IANA). For example, HTTP traffic is associated with port number 80**,** DNS with port number 53, etc. This approach utilizes packet headers only. Historically, most of the applications utilized 'well-known' port numbers to which other hosts initiate communication. Then, a classifier which is placed in the middle of network looks for SYN packets which are TCP packets utilized during 3-way handshake process, to identify server side of the TCP connection; and since this packet also contain target port number, the application also gets identified by it. In similar way, UDP also utilizes port numbers, even though there is neither connection establishment nor maintenance of connection state.

Advantages:

It is fast method as no calculations as involved. Its implementation is simple since it requires adding port nos. for new applications in database.

*Limitations***:**

Some applications (such as p2p) may not have their ports registered with IANA and use dynamic port nos. Many applications masquerade and hide their traffic behind well known port nos. such as port no. 80 which is used for HTTP traffic to get through firewalls. Sometimes IP layer

encryption may also obfuscate TCP or UDP port nos. thus making it impossible to recognize actual port no. Also, sometimes specific applications use non-default port numbers.
Moore and Papagiannaki [7] observed that at most 70% of the byte accuracy could be achieved using port-based classification by utilizing official IANA list. Sen et al. [8] observed that only 30% of the total traffic (in bytes) for Kazaa P2P protocol could be found using default port number.

B. Payload based classification:

This approach not only look into the packet headers but also into the packet payloads. Here the packet payloads are examined bit by bit to locate the bit streams that contain signatures (pre-defined byte sequences) of certain network protocol. If such bit streams are found, then packets can be accurately labeled. Then, stored-signatures are compared directly to packets of network applications in order to accurately perform classification. For example, web traffic can be identified with '\GET' string, eDonkey P2P traffic contains 'xe3\x38' string, etc. This approach is commonly employed for P2P traffic detection [9, 10, 11] and network intrusion detection [12].

Advantage:

It is able to perform traffic classification fairly accurately.

Limitations:

It imposes significant complexity and processing load on traffic identification device. The device also needs to be kept updated with application protocol semantics. It is difficult or almost impossible when dealing with encrypted traffic or proprietary protocols. Direct analysis of packet payload breaches organizational privacy policies or violate relevant privacy legislation. It is difficult to maintain signatures with high hit ratio and low false positive ratio. For example, '\GET' string finds both HTTP and Gnutella applications which leads to ambiguity. To reduce trace file size, packets are recorded with limited length. Hence, signature may not be contained in that part. Packet fragmentation also leads to computational complexity.

III. NEW TRAFFIC CLASSIFICATION TECHNIQUE

Due to number of limitations of traditional techniques, newer approaches have been found, which rely on traffic's statistical characteristics to identify applications.

A. Statistical based classification:

It uses network or transport layer which has statistical properties such as distribution of flow duration, flow idle time, packet inter-arrival time, packet lengths, etc. These are unique for certain classes of applications and hence help to distinguish different applications from each other. Some statistical features of packet-level-trace are captured which are then used to classify network traffic. For example, sudden jump in rate of packets may be an indication of P2P applications or BGP updates or worm propagation. This method is feasible to determine application type but not generally the specific application/client type. For example, it can't determine if flow belongs to Skype or MSN messenger voice traffic specifically.

B. Machine Learning approach for Statistical Traffic Classification:

Due to the need to deal with traffic patterns, large datasets and increasing number of features, it becomes more difficult to specify a mapping between the features and the respective traffic classes. Hence, there arises the need for introduction of Machine Learning (ML) techniques for classification, where different algorithmic procedures can be applied to construct a classifier that groups data instances into different classes based on their feature values. ML is therefore, a powerful technique for data mining and knowledge discovery which searches and describes useful structural patterns in data. It has wide range of applications including search engines, handwriting recognition, medical diagnosis, etc. In 1994, ML was first utilized for Internet flow classification in the context of intrusion detection.
ML takes input in the form of dataset of instances which are individual and independent examples of datasets. Each instance is characterized by value of its features that measures different aspects of the instance. The dataset is ultimately presented as matrix of instances versus features. The output is description of knowledge that is learnt. There are following 4 types of learning:

- Classification (supervised learning)
- Clustering (unsupervised learning)
- Association
- Numeric Prediction

Classification technique involves learning from pre-labeled examples, from which a set of ruled are generated that are used to classify unknown examples. Clustering technique involves grouping instances based on their similarities. In association technique, any association between features is sought. Numeric prediction outcomes numeric quantity rather than a discrete class.

C. Feature selection:

Feature selection is one of the most critical step for the performance of ML algorithms. It is the process of selecting smallest necessary set of features required to achieve one's accuracy goals. There are many features available which are used to classify traffic, but using irrelevant or redundant features often lead to negative impact on accuracy of most ML algorithms and can make the system computationally expensive since the amount of information to be stored and processed also increases. Therefore, it becomes necessary to select only important subset of features. For this purpose, feature selection algorithms are employed, which are classified into following 2 methods:

Filter method:

It make independent assessment based on general characteristics of data and rely on certain metrics to rate & select best subset before learning commences. Its results are not biased towards particular ML algorithm used.

Wrapper method:

It evaluates performance of different subsets using ML algorithms that will finally be employed for learning. Hence, its results are biased towards particular ML algorithm used. Examples of algorithms which can be employed for this process are: Correlation based Feature Selection (CFS), Genetic Algorithm (GA), etc.

IV. WORK DONE USING NEW CLASSIFICATION TECHNIQUES

All current traffic classification techniques make use of statistical characteristics of flows in combination with ML techniques. However, there are some other approaches as well which makes use of statistical characteristics of flows but may not use ML technique to classify traffic but instead may use traditional techniques in combination, as we will see in the following sections.

A. *Statistical Classification using ML techniques*:

Min-huo et al. in [1] adopted an approach for online traffic classification by observing first n TCP packets and then use Bayesian network method to build a classifier. Here, the supervised learning mechanism is used, in which classifier can classify TCP flow dynamically as packets passed through it by deciding if a TCP flow belongs to a given application. The approach is depicted in Figure 1.

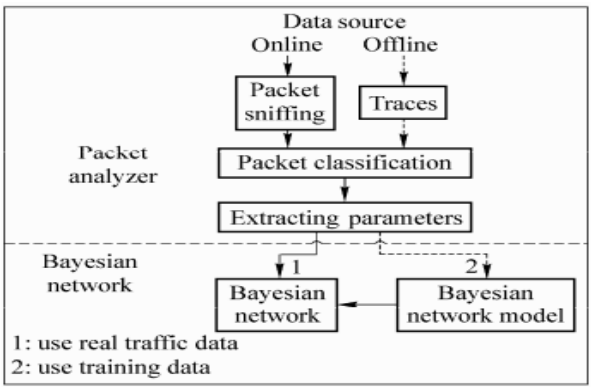


Fig. 1 Online traffic classifier [1]

Here, the classifier structure includes 2 parts: Packet analyzer and Bayesian network classifier. Packet analyzer separates incoming traffic into TCP flows by making use of: source IP, source port, destination IP, destination port and protocol and gains the statistics of length and inter-arrival time of first n packets. It can deal with both online as well as offline traffic. Further, training data for the Bayesian classifier is obtained manually which includes <packet length, inter-arrival time, application type>. Then, Bayesian network model is constructed using training data to get Bayesian network structure and joint probability distribution P_B(packet length, inter-arrival time, application type). Finally, posterior probability P_B(type|packet length, inter-arrival time) is calculated and application with highest P_B is considered as TCP flow application.

To remove irrelevant and redundant features, this method makes use of CFS (Correlation-based Feature Selection) algorithm to identify the optimal subset of features. TCP-dump was used for collection of data from authors' campus network and applications considered for datasets were: FTP(control), FTP(data), Mail, Http and IPTV. The average accuracy of classification across all traces reached 98.4%. Hence, this technique can be used for online traffic classification and saves storage space as it utilizes only first n packets of TCP connection.

Mei-feng et al. in [2] adopted the approach for real time online traffic classification using two characteristics: (ACK-len ab and ACK-len ba) which is that data sent continuously by the communicating parties. These two characteristics depend only on data's total length of first few packets on the flow. In this technique, total data length is taken, which is sent by the host A before it receives first ACK packet from another host B. Here, ACK-len ab represents total data length sent from A to B before first ACK packet arrived (and similarly for ACK-len ba). To verify the effectiveness of ACK-len ab and ACK-len ba, decision tree algorithm C4.5 is utilized as classifier to identify four types of applications: WWW, FTP, Email and P2P. Here also, the authors utilized 5-tuples (source IP, source port, destination IP, destination port, protocol) to define the flow; which is considered to be expired if no more packets belonging to the flow is observed for period of 60 seconds.

Authors identified the behavior of internet applications based on TCP flow which show periodic characteristics. For example, when WWW is browsed, hyperlinks are clicked/visited continuously and produce similar kind of traffic at every click. This can be explained by considering Figure 2, which provides a flow fragment of WWW. In the fig.3, it can be identified that packet no. 236-238 is a three-way handshake process. Then, packet 239-243 is one period; which begins with a Http request packet (239) of data volume 327 bytes (ACK of packet 240 minus sequence no. of packet 238), which is sent from client (IP address: 192.168.0.3, port: 1328) to server (IP address: 58.192.140.19, port: 80). After that, server sends 3 packets (240-242) of data volume 2035 bytes (ACK of packet 243 minus sequence no. of packet 240) for this request. Similarly, packets 244-247 represents another period. From here, it can be inferred that both sides have data volume to send, but data size sent by server is larger than that of the client, which is consistent with data interactive characteristic of WWW.

No.	Source address	Source port	Destination adress	Destination port	protocol	Info
236	192.168.0.3	1328	58.192.140.19	80	TCP	ewall > http [SYN] Seq=0 Win=65535
237	58.192.140.19	80	192.168.0.3	1328	TCP	http > ewall [SYN, ACK] Seq=0 Ack=1
238	192.168.0.3	1328	58.192.140.19	80	TCP	ewall > http [ACK] Seq=1 Ack=1 Win=
239	192.168.0.3	1328	58.192.140.19	80	HTTP	GET / HTTP/1.1
240	58.192.140.19	80	192.168.0.3	1328	TCP	http > ewall [ACK] Seq=1 Ack=328 Wi
241	58.192.140.19	80	192.168.0.3	1328	TCP	[TCP segment of a reassembled PDU]
242	58.192.140.19	80	192.168.0.3	1328	HTTP	HTTP/1.1 200 OK (text/html)
243	192.168.0.3	1328	58.192.140.19	80	TCP	ewall > http [ACK] Seq=328 Ack=2036
244	192.168.0.3	1328	58.192.140.19	80	HTTP	GET /images/css.css HTTP/1.1
245	58.192.140.19	80	192.168.0.3	1328	TCP	[TCP segment of a reassembled PDU]
246	58.192.140.19	80	192.168.0.3	1328	HTTP	HTTP/1.1 200 OK (text/css)
247	192.168.0.3	1328	58.192.140.19	80	TCP	ewall > http [ACK] Seq=596 Ack=4877
248	192.168.0.3	1328	58.192.140.19	80	HTTP	GET /images/face_01.jpg HTTP/1.1
251	58.192.140.19	80	192.168.0.3	1328	TCP	[TCP segment of a reassembled PDU]
252	58.192.140.19	80	192.168.0.3	1328	TCP	[TCP segment of a reassembled PDU]

Fig. 2 A fragment of WWW flow [2]

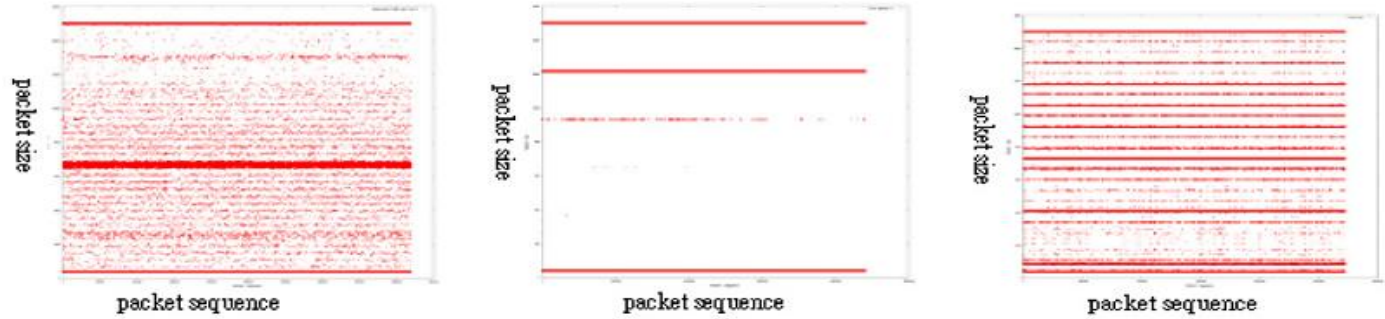
Similar inference can be drawn for FTP which has 2 types of flows: control flow and data flow. In control flow, server sends small data (say 28 bytes) and ACK packet of client has no data. In data flow, after 3-way handshake, server sends large amount of data (say 2920 bytes) to client which again send no data, but only acknowledgement. Similarly, for P2P, both download and upload are supported; in which one peer first sends its data to other, then after receiving the data, receiver sends its own data to the sender and piggybacking acknowledgement. The data volume here is different for different P2P applications and in general are non-zero. Hence, to sum up, data first sent continuously can certainly reflect patterns of network applications, which is different for different applications. This technique obtained good results with accuracy over 97% for dataset acquired on authors' working environment.

Hence, this technique can be used for real-time classification and it doesn't rely on datasets and saves storage space as it only require to store size of first early data packets with no requirement for arrival order. But, this technique is currently limited only for TCP flow and depends on payload length which can be changed by the attacker by padding data at end of the packet.

B. Statistical Classification without using ML techniques:

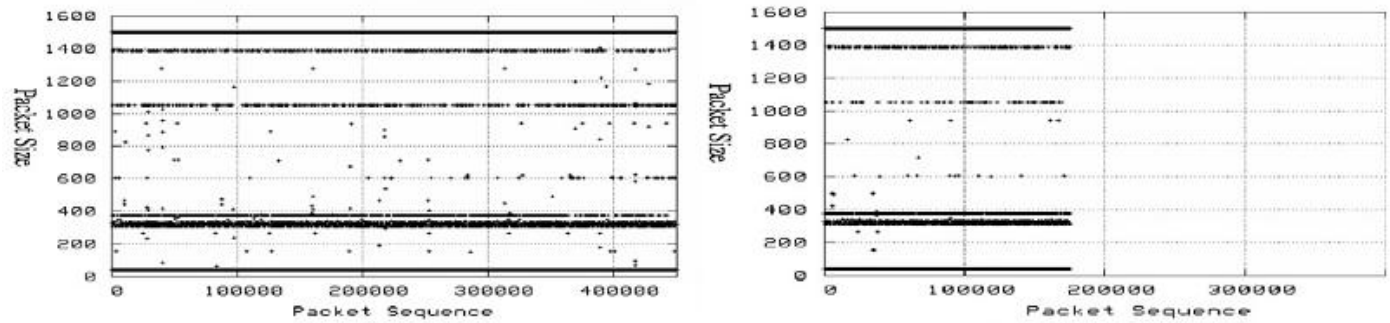
Chun-Nan Lu et al. in [3] adopted the approach of session level flow classification (SLFC) to classify network flows as a session, which comprises of flows in same conversation. SLFC contains 2 parts: flow classification and flow grouping. The former classifies flows into applications by Packet Size Distribution (PSD) and latter groups related flows as sessions by utilizing port locality. This technique also identify flow with 5-tuple: source IP, source port, destination IP, destination port and protocol.

The authors observed that flows of same application have similar PSDs which is represented in Figure 3; whereas different applications have diverse PSDs which is represented in Figure 4.



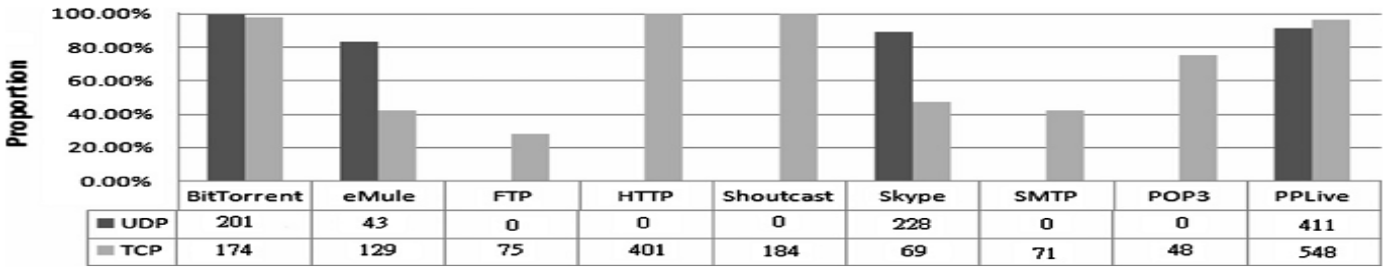
(a) PSD of one Shoutcast instance (b) PSD of one FTP instance (c) PSD of one eMule instance

Fig. 3 Different applications having distinct PSD [3]



(a) BitTorrent Instance 1 (b) BitTorrent Instance 2

Fig. 4 Same application(BitTorrent) having similar PSD [3]



	BitTorrent	eMule	FTP	HTTP	Shoutcast	Skype	SMTP	POP3	PPLive
UDP	201	43	0	0	0	228	0	0	411
TCP	174	129	75	401	184	69	71	48	548

Fig. 5 Different applications have different most frequent packet size [3]

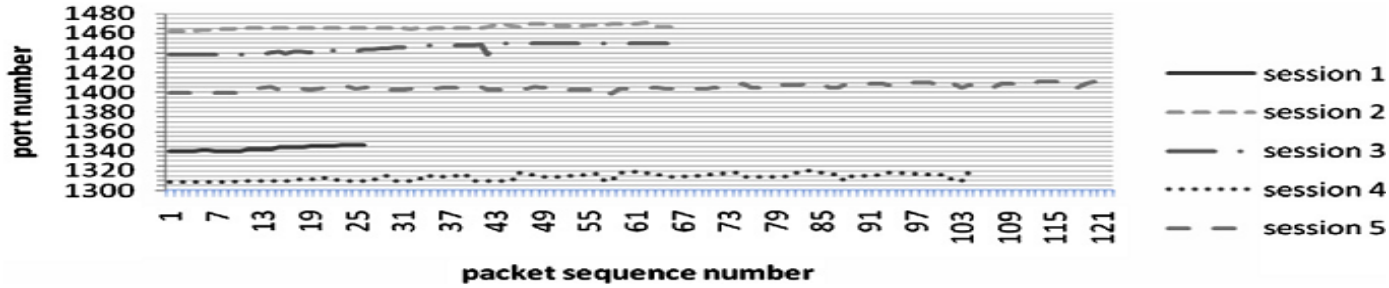


Fig. 6 Port numbers used along with packets during Http sessions [3]

Also, different applications produce unequal packet sizes due to different operational requirements. Figure 5. shows most frequently used packet size of each pre-selected application (except packets without payloads and packets with size of maximum transmission unit). PSD of network application can be obtained from all of its flows. To capture traffic of a specific application, its traces are captured manually in a crafted environment. When PSD of one flow is determined, it is compared with each representative of pre-selected application to identify which application it belongs to. Further, flows are grouped into sessions by checking port locality because operating system often allocate consecutive port numbers for an application when it establishes multiple connections with remote host. For example, Figure 6. shows port numbers used by flows of multiple Http sessions.

Figure 7. shows overview of this technique. SLFC consists of 2 phases: an offline training phase and an online session classification phase. The former finds out application representatives by first collecting set of traffic traces and then extracting representatives from it. The latter first extracts 5-tuple (source IP, source port, destination IP, destination port and protocol) and PSD from all real world flows. Then, PSD is transformed into 2-D space point. Next, the flow classification module compares flows with application representatives and classifies it into application having minimum distance. Afterwards, session grouping module tries to group flows as a session based on port locality. If 2 or more flows of a session are classified as different applications, the application arbitration module is invoked to solve the conflict by treating all flows of a session as an application having largest amount of flows in that session.

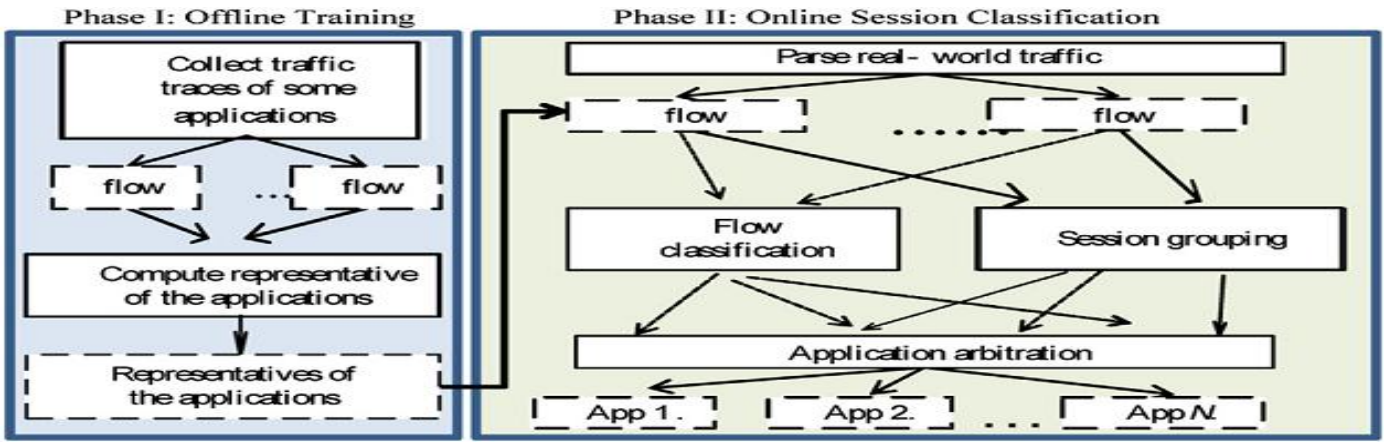


Fig. 7 Components of SLFC [3]

V. CONCLUSION

This paper reviewed three modern statistical classification techniques. First technique utilizes first n TCP packets along with Bayesian network ML algorithm to classify traffic. Second technique classifies traffic based on 2 characteristics: ACK-Len ab and ACK-Len ba, which is the data sent continuously by the communicating parties and utilizes C4.5 decision tree ML algorithm. Third technique classifies traffic using the PSD approach and then classifying flows into sessions based on port nos. but doesn't utilize ML algorithm. All of the mentioned techniques are able to classify traffic in real time, which is of prime importance these days of growing internet traffic.

REFERENCES

[1] Min-huo HONG, Ren-tao GU, Hong-xiang WANG, Yong-mei SUN, Yue-feng JI, "*Identifying online traffic based on property of TCP flow*", The Journal of China Universities of Posts and Telecommunications, vol. 16, no. 3, pp. 84-88, June 2009.

[2] Mei-feng SUN, Jing-tao CHEN ,"*Research of Traffic characteristics for real time online traffic classification*", The Journal of China Universities of Posts and Telecommunications, Comuter Networks,vol. 18, no. 3, pp. 92-98, June 2011.

[3] Chun-Nan Lu, Chun-Ying Huang, Ying-Dar Lin, Yuan-Cheng Lai ,"*Session level flow classification by packet size distribution and session grouping*", Computer Networks, vol. 56,no.1, pp. 260-272, Jan 2012.

[4] Nguyen, T.T.T., Armitage, G., "*A survey of techniques for internet traffic classification using machine learning*", in: IEEE Communications Surveys & Tutorials, vol. 10, no.4, Jan 2009.

[5] Callado, A.; Kamienski, C.; Szabo, G.; Gero, B.; Kelner, J.; Fernandes, S.; Sadok, D.,"*A Survey on Internet Traffic Identification*", in: IEEE Communications Surveys & Tutorials, vol. 11, no. 3, pp. 37-52, Aug 2009.

[6] Murat Soysal, Ece Guran Schmidt,"*Machine Learning Algorithms for accurate flow based network traffic classification: Evaluation and Comparison*", Performance Evaluation, vol. 67, no. 6, pp. 451-467, June,2010.

[7] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in Proc. Passive and Active Measurement Workshop (PAM2005), Boston, MA, USA, March/April 2005.

[8] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in network identification of P2P traffic using application signatures," in WWW2004, New York, NY, USA, May 2004.

[9] T. Karagiannis, A. Broido, N. Brownlee, K.C. Claffy, M. Faloutsos, "Is P2P dying or just hiding", in: IEEE Global Telecommunications Conference, GLOBECOM 04, 2004.

[10] S. Sen , C. Spatscheck, D. Wang, "Accurate, scalable innetwork identification of P2P traffic using application signature", in: 13th International Conference on World Wide Web, 2004.

[11] T. Karagiannis, A. Broido, M. Faloutsos, K.C. Claffy, "Transport layer identification of P2P traffic", in: 4th ACM SIGCOMM Conference on Internet Measurement, 2004.

[12] K. Wang, S.J. Stolfo, "Anomalous payload-based network intrusion detection", in: Lecture Notes in Computer Science, Springer, Berlin, 2004.

