Journal of Southeast University (English Edition) Vol. 30, No. 3, pp. 289 - 295 Sept. 2014 ISSN 1003—7985

Early-stage Internet traffic identification based on packet payload size

Wu Tong[1] Han Zhen[1] Wang Wei[1] Peng Lizhi[2]

([1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

([2]Provincial Key Laboratory for Network Based Intelligent Computing, University of Jinan, Jinan 250022, China)

Abstract: In order to classify the Internet traffic of different Internet applications more quickly, two open Internet traffic traces, Auckland Ⅱ and UNIBS traffic traces, are employed as study objects. Eight earliest packets with non-zero flow payload sizes are selected and their payload sizes are used as the early-stage flow features. Such features can be easily and rapidly extracted at the early flow stage, which makes them outstanding. The behavior patterns of different Internet applications are analyzed by visualizing the early-stage packet size values. Analysis results show that most Internet applications can reflect their own early packet size behavior patterns. Early packet sizes are assumed to carry enough information for effective traffic identification. Three classical machine learning classifiers, i. e., the naive Bayesian classifier, naive Bayesian trees, and the radial basis function neural networks, are used to validate the effectiveness of the proposed assumption. The experimental results show that the early stage packet sizes can be used as features for traffic identification.

Key words: pattern recognition; network measurement; traffic classification; traffic feature

doi: 10.3969/j.issn.1003-7985.2014.03.006

With the explosion of Internet traffic, an accurate traffic classification has become progressively important for network management (e.g., deploying quality-of-service-aware mechanisms, bandwidth budget management, and intrusion detection). Two effective classical techniques are used under traditional network conditions: port-based and payload-based methods. However, these traditional techniques are becoming ineffective for the modern Internet because of dynamic port numbers and encryption techniques. Machine learning techniques introduced by traffic classification research have proven to be promising techniques in recent years[1-4]. Machine learning-based traffic classification techniques are effective in modern traffic classification because they can identify traffic according to macro-patterns instead of micro-features, especially for cases that traditional techniques suffer from. Many supervised and unsupervised machine learning algorithms have been successfully applied in traffic classification in the past few years. Moore and his research group have made significant contributions in this area[5]. They first built a traffic classification data set in 2005 based on 248 statistical features. They used Bayesian analysis techniques[6] and Bayesian neural networks[7] in traffic classification, and achieved high identification accuracies. Este et al.[8] studied the information stability carried by traffic flow features at the packet level, including packet size, round-trip time (RTT), location and inter-arrival time (IAT). The packet size provides the most significant contribution in discriminating application protocols. They also proposed a support vector machine (SVM) based classification frame, which achieved high classification accuracies on three publicly released data sets[9]. Li et al.[10-11] also used SVM. Crotti et al.[12] proposed protocol fingerprinting, which was a novel statistical method[12]. Du et al.[13] used the *k*-nearest neighbors algorithm as the classifier integrated with a binary particle swarm optimization algorithm. They constructed a multistage traffic classifier[14]. Unsupervised learning techniques have also attracted much research attention. Bernaille et al.[15] stated that a supervised learning model should be based on a pre-labeled set of samples. They also stated that an unsupervised learning was appropriate for traffic classification because this type did not rely on predefined classes. Erman et al.[16] used *K*-means and density-based spatial clustering of applications with noise clustering algorithms to find traffic patterns. They proposed a semi-supervised traffic classification model that found and mapped traffic clusters to applications by using ground truths[17]. Other machine learning techniques, such as the Gaussian mixed model[18] and flexible neural trees[19], were also applied in traffic classification.

Early-stage traffic identification has drawn considerable interest from the research community in recent years[20].

Received 2013-12-26.

Biography: Wu Tong (1979—), male, doctor, lecturer, wutong@bjtu.edu.cn.

Foundation items: The Program for New Century Excellent Talents in University (No. NCET-11-0565), the Fundamental Research Funds for the Central Universities (No. K13JB00160, 2012JBZ010, 2011JBM217), the Ph. D. Programs Foundation of Ministry of Education of China (No. 20120009120010), the Program for Innovative Research Team in University of Ministry of Education of China (No. IRT201206), the Natural Science Foundation of Shandong Province (No. ZR2012FM010, ZR2011FZ001).

Citation: Wu Tong, Han Zhen, Wang Wei, et al. Early-stage Internet traffic identification based on packet payload size[J]. Journal of Southeast University (English Edition), 2014, 30(3): 289 - 295. [doi: 10.3969/j.issn.1003-7985.2014.03.006]

Most traditional machine learning-based traffic classification techniques use an instance's statistical features to identify traffic. In real cases, classifying Internet traffic when they have ended is useless. Therefore, many researchers have turned to finding effective models that are able to identify early-stage Internet traffic. In 2009, Este et al. [8] proved that early-stage packets of an Internet flow can carry enough information for traffic classification. They analyzed RTT, packet size, IAT, and packet direction of early-stage packets. Packet size was the most effective feature for early-stage classifications. In 2008, Huang et al. [21] studied the early-stage application characteristics and used them for effective classification. They recently extracted early-stage traffic features by analyzing the negotiation behaviors of different applications. They applied these features to machine learning-based classifiers with high performances[22]. Hullár et al. [23] proposed an automatic machine learning-based method that consumes limited computational and memory resources for early-stage P2P traffic identification. Dainotti et al. [24] constructed highly effective hybrid classifiers and applied a hybrid feature extraction method to early-stage traffic classification. Nguyen et al. [25] used statistical features from sub-flows for timely voice-over-Internet protocol traffic identification. A sub-flow is a small number of the most recent packets taken at any point in a flow's lifetime. Hence, they made the early stage "timely."

In this study, the payload sizes of the eight earliest non-zero packets are used as the early-stage flow features. First, the packet payload size is directly extracted from its header without any extra computation. Secondly, the payload size is the most important packet-level feature with the lowest network environment correlation. Other packet-level features (e. g., RTT and IAT) heavily depend on the network environment. The application negotiation procedure is the main factor that determines the payload size. First, the behavior patterns of different Internet applications are analyzed by visualizing these early packet size values. This study employs two open Internet traffic traces, namely Auckland Ⅱ and UNIBS traffic traces. The phenomenon that most Internet applications show their own early-stage packet size behavior patterns is found. It is assumed that early packet sizes carry enough information for effective traffic identification. Three classical machine learning classifiers are applied to our validation experiments. The early stage packet sizes are used as effective traffic identification features.

1 Data Sets

Two open Internet traffic data sets are employed in this study: Auckland Ⅱ captured in New Zealand in 2000 and UNIBS captured in Italy in 2009. The two data sets are selected to analyze and validate the early-stage Internet traffic packet size features.

1.1 Auckland traffic traces

Auckland Ⅱ is a collection of long GPS-synchronized traces obtained by using a pair of DAG 2 cards at the University of Auckland. This data set is available in Ref. [26]. From November 1999 to July 2000, 85 trace files were captured. Most traces were targeted at 24-hour runs. However, hardware failures resulted in significantly shorter traces. Two trace files captured on Feb. 14, 2000 (i. e., 20000214-185536-0. pcap and 20000214-185536-1. pcap) are selected. The traces included only the header bytes with a maximum amount of 64 bytes for each frame. The application payload was fully removed. All IP addresses were used with anonymity by using the Crypto-Pan AES encryption. The header traces were captured with a GPS-synchronized mechanism by using a DAG3. 2E card connected to a 100 Mbit/s Ethernet hub that interconnects the University's firewall to the border router.

Deep packet inspection tools are invalid for obtaining ground truths because the application payloads are not recorded in Auckland Ⅱ. The only way to obtain the original application type is to use port numbers. In this study, only the transmission control protocol (TCP) case is used because the TCP is the predominant transport layer protocol. Each flow is assigned to the server port-identified class. Eight main packet types are selected from the Auckland Ⅱ traces and filter "mice flows" with no more than eight packets with payload. All selected types and their instance and total byte distributions are listed in Tab. 1.

Tab. 1 Selected types of Auckland Ⅱ trace

Type	Number of instances	Total bytes
ftp	251	136 241
ftp-data	463	5 260 804
http	23 721	13 942 1961
imap	193	86 455
pop3	498	98 699
smtp	2 602	1 230 528
ssh	237	149 502
telnet	37	21 171

1.2 UNIBS traffic traces

The UNIBS is another open traffic trace developed by Prof. F. Gringoli and his research team. This data set is available in Ref. [27]. They developed a useful system named GT[28] to apply ground truths of captured Internet traffic. The traces were collected from the University of Brescia campus network's edge router for three consecutive working days (Sept. 30, Oct. 1, and Oct. 2, 2009). These traces were composed of traffic generated by a set of 20 workstations that run the GT client daemon. Traffic is collected by running Tcpdump[29] on the faculty's router. The router was a dual Xeon Linux box that connected

the network to the Internet through a dedicated 100 Mbit/s uplink. In UNIBS, 99% are TCP flows. Hence, the research reuses the TCP flows in this data set. The UNIBS traces record each captured flow's application information by using GT. The application ground truths are obtained by using both TCP port numbers and GT records. Eight main packet types are chosen in UNIBS (see Tab. 2). Two popular P2P applications included in this data set (i.e., BitTorrent and eDonkey) are recorded by GT. Skype is also selected as an import Internet application. Flows with no more than eight packets of payload were filtered. Each type's instance and total byte distributions are listed in Tab. 2.

Tab. 2 Selected types of UNIBS traces

Type	Numbers of instances	Total bytes
bittorrent	3 571	6 393 487
edonkey	379	241 587
http	25 729	107 342 346
imap	327	860 226
pop3	2 473	4 292 419
skype	801	805 453
smtp	120	43 566
ssh	23	39 456

2 Feature Analysis Using Visualization

In this study, the paper analyzes the effectiveness of the early-stage packet size features by using the visualization method. The research selects the eight earliest non-zero payload packets for each TCP traffic instance in both the Auckland Ⅱ and UNIBS traces. The transport layer payload sizes of these packets are used as features. As stated in Section 1, the traffic samples whose number of non-zero payload size packets is less than eight were filtered as mice flow and were not taken into account. Four scatters for both the Auckland Ⅱ and UNIBS traces are drawn. The first one is the payload size view of packets 1 and 2. The second one is the payload size view of packets 3 and 4, and so forth. A logarithmic scale is used for both horizontal and vertical axes in these scatters. The horizontal axis represents the former packet's payload size. The corresponding vertical axis represents that of the latter packet (e.g., the horizontal axis of Fig. 1 represents the payload size of packet 1 and the vertical axis represents that of packet 2). We visualize the early-stage payload size features and preliminarily validate these features' effectiveness by using the scale.

The visualization results of Auckland Ⅱ traces are shown in Fig. 1. All traffic samples are centralized in a few values for the first packet's payload size (see Fig. 1).

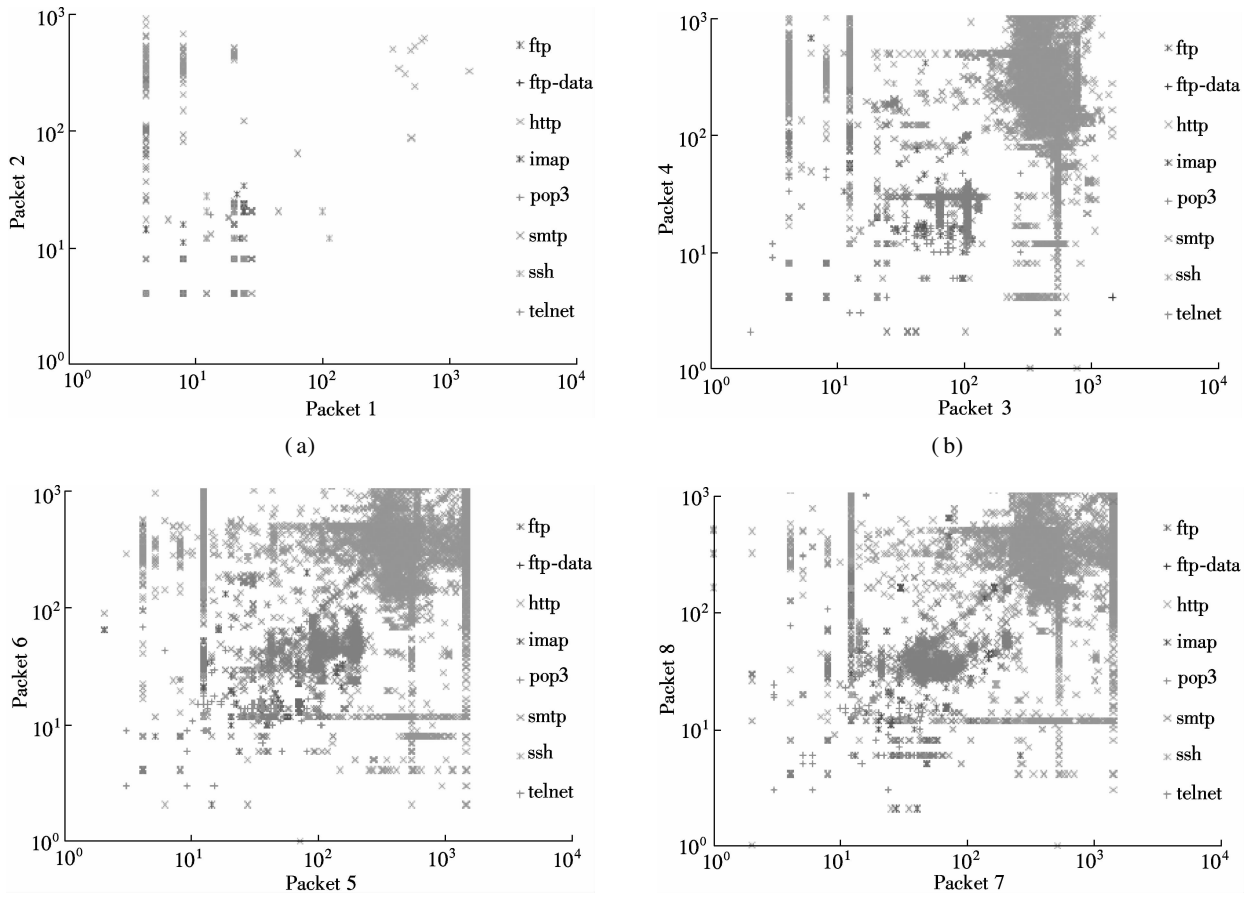
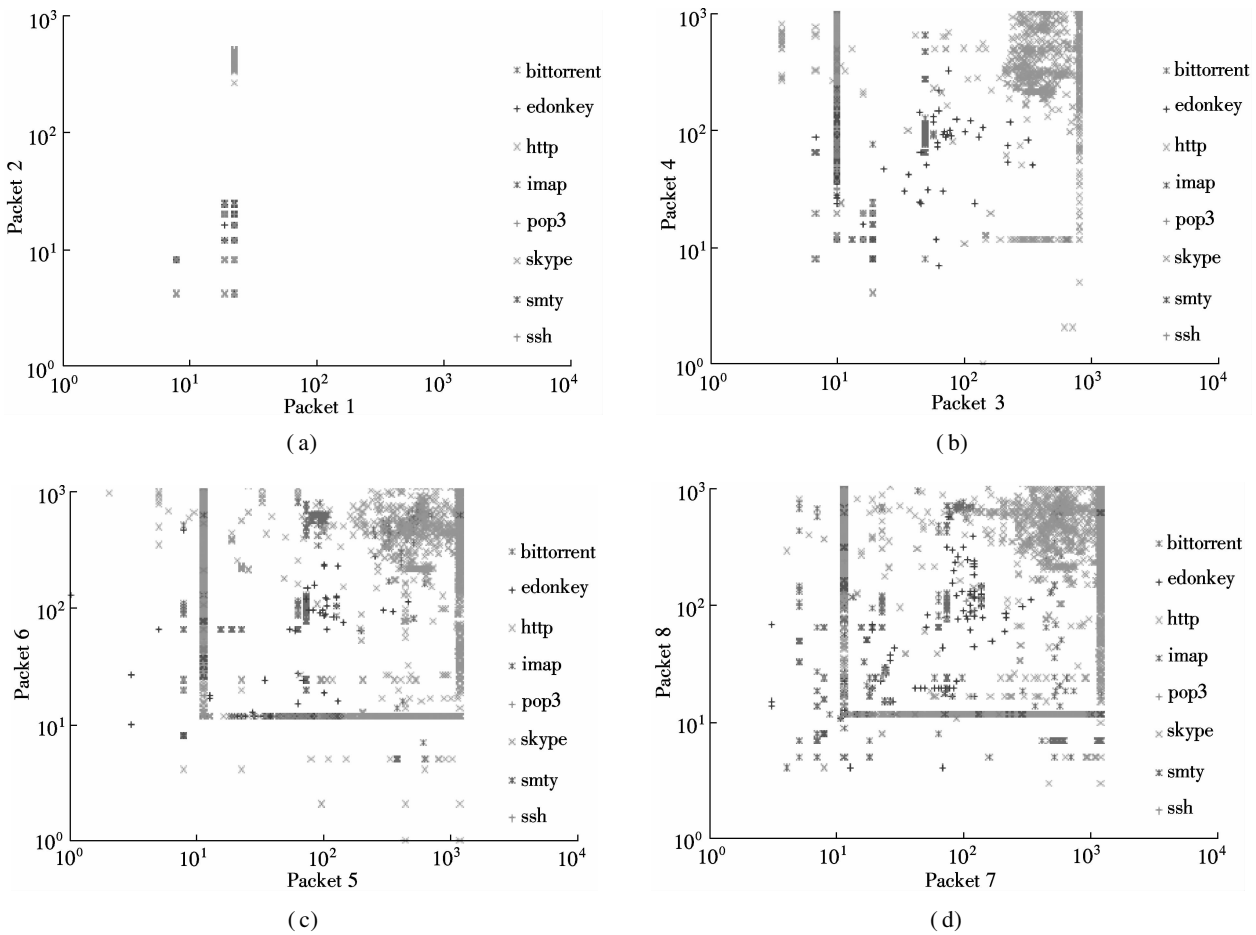


Fig. 1 View of the payload sizes of Auckland Ⅱ trace

The second packet also does not scatter on many values for the payload size. Hence, making a distinction between traffic types is relatively difficult. Therefore, the first and the second packets are unsuitable for traffic pattern recognition in this data set. Most types show their unique patterns in Figs. 1(b) to (d). Most hypertext transfer protocol (HTTP) instances congregate on the top-right corner of these figures as the dominant traffic type. Most HTTP sessions carry relatively heavy payloads from the third packet. Many Internet applications are layered on the HTTP protocol (e.g., Internet streaming media and Web mail). This structure causes various HTTP traffic patterns. Some HTTP instance scatter is observed on different areas in Figs. 1(b) to (d). Another important type in the Auckland traces is the simple mail transfer protocol (SMTP). In the same figures, SMTP show a different behavior pattern. Most SMTP instances cluster at the middle areas and they are represented by the cross area in each figure. The POP3 instances are fewer than that of the SMTP. Most POP3 instances cluster in Figs. 1(b) and (c). However, POP3 does not create a clear cluster in Fig. 1(d). POP3 does not have its clear pattern from packet 7.

An interesting phenomenon, which we called the "12-byte size packet" phenomenon, is observed in Fig. 1. Many packets have a payload size of 12 bytes despite their traffic types. We cannot explain why various traffic kinds generated excessive 12-byte size packets. The reasons behind this phenomenon are difficult to discover without the payload content information in the traffic trace.

The UNIBS trace visualization results are shown in Fig. 2. In general, various kinds of traffic in the UNIBS data set do not make clear clusters like those in the Auckland Ⅱ data set. The main traffic types are still easily distinguished from each other. Some characteristics that are easily observed from these figures are in accordance with Fig. 2. First, packets 1 and 2 in Fig. 2(a) show few payload size values. Furthermore, the traffic behavior patterns of packets 1 and 2 in the UNIBS trace are simpler than those in the Auckland Ⅱ trace. Secondly, HTTP also occupies the top-right corners of the views of packets 3 and 4, packets 5 and 6, and packets 7 and 8. The pattern is in accordance with the Auckland Ⅱ trace attribute. As P2P traffic types, BitTorrent and eDonkey have a considerable number of instances in the UNIBS trace. These two P2P applications do not show clear behavior patterns in the visualization views except in packets 5 and 6. Some BitTorrent and eDonkey instances mix with the HTTP instances in Figs. 2(c) and (d). However, these instances clearly cluster together in Fig. 2(b). The P2P traffic has its own early-stage packet payload size characteristics.



(a) (b) (c) (d)

Fig. 2 View of the payload sizes of UNIBS trace

3 Classification Experiments

The naive Bayesian classifier (NB), naive Bayesian classification trees (NBTree), and radial basis function neural networks (RBF) are applied for classification tests to validate the effectiveness of the early-stage packet size features in traffic classification. Five-folder crossover validation is used on both Auckland Ⅱ and UNIBS data sets. Each data set is uniformly split into five subsets. The first subset is used for the testing set. The other four subsets are used for the training set. The second subset is used for the testing set and so forth. Each time, the NB, NBTree, and RBF are applied, and their classification results are compared.

3.1 Performance measures

The confusion matrix is the classifier measurement basis in which rows denote the actual instance class and the columns denote the predicted class. A typical binary classification confusion matrix is shown in Fig. 3. The true positive (TP) is the number of positive instances that are correctly classified. The false positive (FP) is the number of positive instances incorrectly classified as negative samples. The true negative (TN) is the number of correctly classified negative instances. The false negative is the number of negative instances incorrectly classified as positive samples. Many measure types are conducted based on the confusion matrix to evaluate the classifier performance. The following measures are mostly used for general classification tasks.

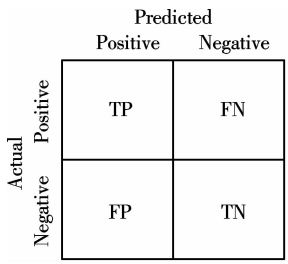


Fig. 3 Confusion matrix

The true positive rate (TPR) is the ratio between the correctly classified positive instances and all actual positive instances:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

The false positive rate (FPR) is the ratio between the incorrectly classified negative instances and all actual negative instances:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

3.2 Results and analysis

The five-folder crossover validation results on the Auckland Ⅱ data set are shown in Tab. 3. The best value for each measure is marked in bold. From the classifiers' point of view, the NBTree obtains the best performance among the three classifiers. The NBTree obtains the highest true positive rate (TPR) values for all classes except for the FTP data. All NBTree TPR values are greater than 0.9 except for Telnet. Most NB and RBF TPR values are greater than 0.85. All false positive rate (FPR) values in the table are low. The classifiers are able to accurately identify other samples for each traffic class. In most cases, the selected classifiers can effectively identify the traffic instances in the Auckland Ⅱ data set by using the early-stage packet size features. Both the NB and the RBF obtain very low TPR values for the FTP and Telnet classes. In addition, the NBTree obtains a relatively low TPR value for telnet. The sample distributions in Tab. 1 show that 251 and 37 instances for FTP and Telnet accounted for 0.90% and 0.13%, respectively, of the entire data set. Therefore, the highly imbalanced Auckland Ⅱ class distribution is an important factor that causes the low identification performances for the FTP and Telnet. As the most predominant traffic type on the Internet, the HTTP should be accurately distinguished from other types. Tab. 3 shows that all NB, NBTree, and RBF achieve high TPR and low FPR values for HTTP traffic instances. Therefore, the early-stage packet size features are able to carry enough information for HTTP traffic identification. This finding is in accordance with the analysis in Section 2.

Tab. 3 5-folder crossover validation results of Auckland Ⅱ data set

Class	NB		NBTree		RBF	
	TPR	FPR	TPR	FPR	TPR	FPR
ftp	0.323	0.003	**0.940**	0.001	0.327	**0.000**
ftp-data	**0.952**	0.009	0.942	**0.000**	0.933	0.002
http	0.968	**0.012**	**0.998**	0.025	0.991	0.035
imap	0.850	0.006	**0.917**	**0.000**	0.539	0.004
pop3	0.906	0.017	**0.958**	**0.001**	0.876	0.007
smtp	0.867	0.007	**0.972**	**0.002**	0.905	0.013
ssh	0.886	0.002	**0.962**	**0.000**	0.886	0.001
telnet	0.162	0.007	**0.649**	**0.000**	0.108	0.000
Average	0.949	**0.011**	**0.992**	0.021	0.969	0.031

The five-folder crossover validation results on the UNIBS data set are shown in Tab. 4. The best value for each measure is marked in bold. The NBTree obtains the best class values for both TRP and FPR with no exception. Therefore, the NBTree is primarily concluded as a good classifier for Internet traffic identification. The NB does not perform well for the UNIBS data set. The NB obtains five TPR values, which are less than 0.8, especially for HTTP. The NB TPR value is 0.746, which is less than that of the NBTree and the RBF (0.999 and 0.992, respectively). The NBTree and RBF obtain TPR values greater than 0.99 and FPR values lower than 0.04. The early-stage packet size features are effectively

used to identify HTTP traffics. All Skype TPR values are low. Traffic identification on Skype is difficult. The Skype instances do not show clear behavior patterns in the visualization views. The instances scatter on many areas in Figs. 2(b) to (d). Therefore, the early-stage packet size features are not enough for Skype traffic identification. Other effective features should be used to identify Skype traffic.

Tab. 4 5-folder crossover validation results of UNIBS data set

Class	NB		NBTree		RBF	
	TPR	FPR	TPR	FPR	TPR	FPR
bittorrent	0. 715	0. 010	**0. 996**	**0. 002**	0. 925	0. 017
edonkey	0. 673	0. 040	**0. 873**	**0. 002**	0. 042	0. 000
http	0. 746	0. 026	**0. 999**	**0. 006**	0. 992	0. 038
imap	0. 498	0. 024	**0. 976**	**0. 001**	0. 691	0. 002
pop3	0. 982	0. 013	**0. 994**	**0. 000**	0. 982	0. 008
skype	0. 395	0. 152	**0. 407**	**0. 000**	0. 000	0. 000
smtp	0. 942	0. 001	**0. 983**	**0. 000**	0. 967	0. 000
ssh	0. 957	0. 000	**1. 000**	**0. 000**	0. 957	0. 000
Average	0. 757	0. 024	**0. 995**	**0. 005**	0. 967	0. 032

A set of comparison experiments were carried out among the early-stage payload size feature set, the early-stage hybrid feature set, and the long-term payload size feature set. These experiments are used to validate the effectiveness of the payload size features. The early-stage hybrid feature set contains the payload sizes, IAT, and RTT of the first eight packets. The long-term payload size feature set contains the payload sizes of the first 20 packets. The model building time is taken as an important performance measure.

The comparison results are shown in Tab. 5, where 8-ps represents the early-stage payload size feature set, 8ps + iat + rtt represents the early-stage hybrid feature set, and 20-ps represents the long-term payload size feature set. The best value for each measure is marked in bold. The 8-ps feature set obtains the best time performance for all classifiers and both data sets. The same feature achieves the best TPR and FPR performances for most cases. The 20-ps feature set result is somewhat unexpected. This feature set does not show advantages over the 8-ps feature set for most cases. The 8ps + iat + rtt feature set performs quite well in the comparison using the NB classifier, especially for the UNIBS data set.

Tab. 5 5-folder crossover validation results using different feature sets on UNIBS data set

Class		Auckland Ⅱ			UNIBS		
		8-ps	8ps + iat + rtt	20-ps	8-ps	8ps + iat + rtt	20-ps
NB	TPR	**0. 949**	0. 0941	0. 940	0. 757	**0. 904**	0. 774
	FPR	**0. 011**	0. 012	0. 015	**0. 024**	0. 035	0. 034
	time	**0. 07**	0. 10	0. 27	**0. 09**	0. 15	0. 41
NBTree	TPR	**0. 992**	**0. 992**	**0. 992**	**0. 995**	0. 992	0. 991
	FPR	0. 021	0. 011	**0. 008**	0. 005	0. 010	**0. 004**
	time	**13. 98**	44. 92	282. 73	**16. 81**	69. 15	568. 67
RBF	TPR	0. 969	**0. 975**	0. 963	**0. 967**	0. 960	0. 947
	FPR	0. 031	**0. 024**	0. 032	**0. 032**	0. 055	0. 094
	time	**289. 39**	549. 87	1 013. 49	**148. 39**	321. 07	973. 22

4 Conclusion

This paper studies Internet traffic identification using the early-stage packet payload size as features. By visualizing the early-stage packet payload size values, different Internet traffic types are found to have their own early-stage behavior schemas. Therefore, constructing effective Internet traffic identification models is possible by using the simple payload sizes of the early-stage packets. The existing early-stage traffic identification techniques use packet size and time features. Our method can easily and rapidly extract features because it uses the packet payload size only. Speed is important for real early-stage traffic identification. Hence, an identification model is required to extract features as fast as possible. Therefore, using the packet payload size as the traffic feature is a feasible and effective feature-extracting method. Three classical classifiers are applied to validate the effectiveness of the feature of the early-stage packet payload size. For most Auckland Ⅱ and UNIBS traffic trace types, the classifiers obtain high identification rates by using the simple payload size features.

References

[1] Callado A, Kamienski C, Szabó G, et al. A survey on internet traffic identification[J]. *IEEE Communications Surveys & Tutorials*, 2009, **11**(3): 37 - 52.

[2] Hu Bin, Shen Yi. Machine learning based network traffic classification: a survey[J]. *Journal of Information & Computational Science*, 2012, **9**(11): 3161 - 4170.

[3] Nguyen T T T, Armitage G. A survey of techniques for Internet traffic classification using machine learning[J]. *IEEE Communications Surveys & Tutorials*, 2008, **10**(4): 56 - 76.

[4] Valenti S, Rossi D, Dainotti A, et al. Reviewing traffic classification[C]//*Data Traffic Monitoring and Analysis*. Berlin: Springer, 2013: 123 - 147.

[5] Moore A, Zuev D, Crogan M. Discriminators for use in flow-based classification[EB/OL]. (2005-08-17)[2013-11-16]. http://www.cl.cam.ac.uk/~awm22/publications/RR-05-13.pdf.

[6] Moore A, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]//*Proceedings of the* 2005

ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM, 2005: 50 - 60.

[7] Auld T, Moore A, Gull S. Bayesian neural networks for Internet traffic classification [J]. *IEEE Transactions on Neural Network*, 2007, **18**(1): 223 - 239.

[8] Este A, Gringoli F, Salgarelli L. On the stability of the information carried by traffic flow features at the packet level [J]. *ACM SIGCOMM Computer Communication Review*, 2009, **39**(3): 13 - 18.

[9] Este A, Gringoli F, Salgarelli L. Support vector machines for TCP traffic classification [J]. *Computer Networks*, 2009, **53**(14): 2476 - 2490.

[10] Li Z, Yuan R, Guan X. Accurate classification of the Internet traffic based on the SVM method[C]//*IEEE International Conference on Communications*. Glasgow, USA, 2007: 1373 - 1378.

[11] Lu Gang, Zhang Hongli, Sha Xuefu, et al. Tcfom: a robust traffic classification framework based on oc-svm combined with mc-svm[C]//2010 *International Conference on Communications and Intelligence Information Security*. Nanning, China, 2010: 180 - 186.

[12] Crotti M, Dusi M, Gringoli F, et al. Traffic classification through simple statistical fingerprinting [J]. *ACM SIGCOMM Computer Communication Review*, 2007, **37**(1): 5 - 16.

[13] Du Min, Chen Xingshu, Tan Jun. Online Internet traffic identification algorithm based on multistage classifier[J]. *China Communications*, 2013, **10**(2): 89 - 97.

[14] Du Min, Chen Xingshu, Tan Jun. A novel P2P traffic identification algorithm based on BPSO and weighted k-nearest-neighbor[J]. *China Communications*, 2011, **8**(2): 52 - 58.

[15] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly[J]. *ACM SIGCOMM Computer Communication Review*, 2006, **36**(2): 23 - 26.

[16] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms[C]//*Proceedings of the* 2006 *SIGCOMM Workshop on Mining Network Data*. New York: ACM, 2006: 281 - 286.

[17] Erman J, Arlitt M, Mahanti A, et al. Offline/realtime traffic classification using semi-supervised learning [J]. *Performance Evaluation*, 2007, **64**(9): 1194 - 1213.

[18] Qian F, Hu G, Yao X. Semi-supervised Internet network traffic classification using a Gaussian mixture model [J]. *Int J Electron Commun*, 2008, **62**(7): 557 - 564.

[19] Peng Lizhi, Zhang Hongli, Yang Bo, et al. Traffic identification using flexible neural trees[C]//2010 18*th International Workshop on Quality of Service*. Beijing, China, 2010: 5542729-1 - 5542729-5.

[20] Dainotti A, Pescape A, Claffy K C. Issues and future directions in traffic classification[J]. *IEEE Network*, 2012, **26**(1): 35 - 40.

[21] Huang N, Jai G, Chao H. Early identifying application traffic with application characteristics[C]//*IEEE International Conference on Communications*. Beijing, China, 2008: 5788 - 5792.

[22] Huang N, Jai G, Chao H, et al. Application traffic classification at the early stage by characterizing application rounds[J]. *Information Sciences*, 2013, **232**: 130 - 142.

[23] Hullár B, Laki S, Gyorgy A. Early identification of peer-to-peer traffic [C]//*IEEE International Conference on Communications*. Kyoto, Japan, 2011: 5963023-1 - 5963023-6.

[24] Dainotti A, Pescape A, Sansone C. Early classification of network traffic through multi-classification [C]//*Lecture Notes in Computer Science*. Berlin: Springer, 2011, 6613: 122 - 135.

[25] Nguyen T T T, Armitage G, B ranch P, et al. Timely and continuous machine-learning-based classification for interactive IP traffic[J]. *IEEE/ACM Transactions on Networking*, 2012, **20**(6): 1880 - 1894.

[26] Waikato Internet Traffic Storage (WITS)[EB/OL]. (2006-06-17)[2012-10-13]. http://www.wand.net.nz/wits.

[27] UNIBS: Data sharing[EB/OL]. (2011-07-21)[2013-09-14]. http://www.ing.unibs.it/ntw/tools/traces/.

[28] Gringoli F, Salgarelli L, Dusi M, et al. GT: picking up the truth from the ground for internet traffic[J]. *ACM SIGCOMM Computer Communication Review*, 2009, **39**(5): 12 - 18.

[29] Tcpdump/Libpcap[EB/OL]. (2013-11-20)[2013-12-16]. http://www.tcpdump.org.

基于有效载荷大小的早期网络流量识别

吴 同[1] 韩 臻[1] 王 伟[1] 彭立志[2]

([1] 北京交通大学计算机与信息技术学院, 北京 100044)

([2] 济南大学山东省网络智能计算技术重点实验室, 济南 250022)

摘要:为快速将网络应用的流量进行分类,以 Auckland Ⅱ和 UNIBS 两个数据集的网络流量包为研究对象,选取网络应用程序流量中最初的 8 个有效载荷大小作为识别特征进行研究. 由于这类特征可在早期流量阶段快速提取,因此效果显著. 通过将早期载荷大小可视化的方式,分析了不同网络应用的行为模式. 分析结果表明,多数网络应用程序可通过早期有效载荷大小显示出它们特有的行为模式,根据早期有效载荷大小的信息可对流量进行有效识别. 在此基础上,选用 3 种典型的机器学习分类器, 即朴素的贝叶斯分类器、朴素的贝叶斯树和径向基函数神经网络进行验证分析. 实验结果显示,早期有效载荷大小可作为特征对流量进行有效识别.

关键词:模式识别;网络测量;流量分类;流量特征

中图分类号:TP393