INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(6):672-685, December, 2014.

Auto Adaptive Identification Algorithm Based on Network Traffic Flow

S. Dong, X. Zhang, D. Zhou

Shi Dong*
1. School of Computer Science and Technology, Zhoukou Normal University
Zhoukou, 466001, China
2. School of Computer Science & Technology, Huazhong University of Science and Technology
Wuhan, 430074, China
*Corresponding author: njbsok@gmail.com

Xingang Zhang
School of Computer and Information Technology, Nanyang Normal University
Nanyang, 473061, China
zxg@nynu.edu.cn

Dingding Zhou
Department of Laboratory and Equipment Management, Zhoukou Normal University
Zhoukou, 466001, China
zdd@zknu.edu.cn

Abstract: Traffic identification is a key task for any Internet Service Provider (ISP) or network administrator. Machine learning method is an important research method on traffic identification, while impact of the asymmetry router on the traffic identification is considered, so this paper analyzes the impact of asymmetry routing on traffic identification, and proposes an effective method to decrease the impact, and experimental results show the auto adaptive algorithm can improve the traffic identification.
Keywords: Traffic identification, Internet Service Provider (ISP), Auto Adaptive algorithm (AA), asymmetry routing.

1 Introduction

Traffic identification play an important in many fundamental network operations and maintenance activities to detect invade and malicious attacks forbid applications, bill on the content of traffics and ensure quality of service. It increasingly becomes one of the most interesting topics in network science and technology fields, especially in recent years. The current network traffic identification methods roughly five categories: (1) port-based method; (2) based on deep packet inspection (dpi) methods; (3) based on the network flow characteristic; (4) based on host behavior [1]; (5) based on machine learning methods.

The machine learning methods are divided into supervised and unsupervised machine learning. These are the more classic identification method; of course, there is also individual QOS quality of service features for identification [2]. Many share a naive assumption about the Internet that traffic on a given link is approximately symmetric, meaning that both directions of a conversation flow across the same physical link. Many developers even embed this assumption in their traffic classification tools [3, 4]. In fact, except at network edges, Internet traffic is often routed asymmetrically [5], which will impair or invalidate the results of tools and models that assume otherwise. An important cause of this asymmetry is "hot-potato routing" [6], the business practice of configuring traffic crossing one's network to exit as soon as possible, minimizing resource consumption, and thus cost, of one's own infrastructure. Particularly common in commercial settlement-free peering agreements, hot-potato routing implies that the network on the

receiving side of a packet will bear higher cost per received packet. The underlying assumption is that if both networks in a settlement-free peering agreement follow this practice, it will even out, and both sides will share evenly in carrying traffic exchanged by their customers. Another cause of asymmetric traffic is link redundancy, or alternative paths within networks. Since routing decisions occur independently for each packet, load-balancing algorithms may cause packets destined to the same endpoint to follow different paths. Other traffic engineering techniques, e.g., policy-based SPF (Shortest Path First), may also induce asymmetry in internal routing state of large provider networks, through studying on asymmetric routing, we found it had some impacts on traffic identification, and we propose auto adaptive (AA) method to improve traffic identification. Experiments results show that the AA method can achieve better accuracy than others.

The paper is structured as follows: Section 2 introduces related work of traffic identification; Section 3 proposes AA algorithm and evaluation method; in Section 4, at last, we list the proportion results which are classified by our identification algorithm, and analyze the impact of ε on traffic identification; Section 5 concludes the paper.

2 Related work

The application identification problem has been changing due the efforts of two factors that are in a continuous competition. On the one hand, the applications, and especially those that do not want to be detected (e.g., P2P applications), in order to use the network resources without control. On the other hand, a group of network operators, investigators and even ISPs who need to know the traffic characteristics of their networks to manage the resources or even charge the users depending on their consumption.

2.1 Research on traffic identification

It has become a hot research between domestic and foreign experts who take the traffic identification as research direction, which proceed distinguish, QOS, intrusion detection, traffic monitoring, billing and management. From the beginning of the study on port-based method, this method is the use for marking and identifying the traffic type by fixed port which supplied by the IANA, the other method is aim at P2P and some certain protocols, which adopt method based on deep packet detection methods, but this method has defect that can't get some encrypted information and can't get the new service type. Recently traffic identification has new method with a number of new applications come out. With appearance of the new service, the method of machine learning has been applied to the traffic identification. Identify fields on the flow, roughly divided into three research directions: one is the feature selection algorithm [7, 8], the other is identification algorithm [1,2,9], another is a category for different types of data sets, for example, all packets can be divided into flows [10–14] that are sampling NETFLOW [15]. Complementary information about related work in the field of traffic identification can be found in the survey of traffic identification techniques using machine learning in [16], in the comparison of contemporary classification methods in [13], the survey on Inter- net traffic identification in [17] and the research review on traffic identification in [18]. A critical but constructive analysis of the field of Internet traffic identification is proposed in [19], focusing on major obstacles to progress and suggestions for overcoming them. Although some articles have been studied on the identification algorithm, but the identification algorithm still exist some problems to be needed to solve, such as the neural network identification algorithm is one point worthy of study. All previous research studies in traffic identification either use insufficient network data, usually non-public, or use very few/meaningless metrics for evaluation, making it impossible to compare results shown in

different papers [17]. In addition to features selection based on flow, especially the impact of the size of packet traffic is always to be concerned. Therefore, in this article we propose AA method, and we analyze different feature metric set (bidirection feature or unidirection feature) cause different identification results.

2.2 Asymmetry routing

For a pair of hosts A and B, if the path from A to B (forward direction) is different from the path from B to A (reverse direction), we say that the pair of paths between A and B exhibit routing asymmetry. This scenario can be very common in the Internet core where asymmetric routing is an usual practice [20, 21], this asymmetry in the Internet can appear on both as level and router level paths. In fact, the path followed by packets exchanged between end points along one direction can be different from the one followed by packets going in the opposite direction. Recent reports suggest that asymmetrical routing might be moving closer to the edge of the internet than one might expect. For example, the analysis presented in [22] argues that this practice is nowadays quite common even in ISPs directly serving campus-wide networks.

2.3 Flow metric

Definition 1. The definition of flow metric, which is composed with traffic statistical feature such as flow length, flow during etc. These features have high correlation with application type. So considered as flow metric to classify traffic by machine learning. While nowadays there are two kinds of flow metric, one is unidirectional flow metric, and the other is bidirectional flow.

Unidirectional flow metric

Uniflow (Unidirectional flow)(or one-way) within your network is most likely the result of an incorrect configuration, but may also be symptomatic of a larger problem related to your overall routing architecture. Since network communications are bi-directional in nature, unidirectional traffic patterns on your network mean that the traffic flow in one direction is not following the same path as the other. By design, the least cost route to a destination should also be the desired return path. Uniclassifier (Unidirectional classifier) is classifier which use unidirectional flow metric for training set. Where unidirectional flow metric is adopted as table 1 in this paper.

Bidirectional flow metric

Biflow(Bidirectional flow): A biflow is a Flow as defined in the IPFIX Protocol document [RFC5101], composed of packets sent in both directions between two endpoints. A biflow is composed from two uniflows such that:
1.the value of each Non-directional Key Field of each Uniflow (Unidirectional flow) is identical to its counterpart in the other, and
2.the value of each Directional Key Field of each uniflow is identical to its reverse direction counterpart in the other. Biclassifier(bidirectional classifier) is classifier which use bidirectional flow metric for training set. Where bidirectional flow metric is adopted as table 2 in this paper.

Table 1: unidirectional flow feature

Feature	Feature Description
lport	low port number
hport	high port number
duration	Flow duration
Transproto	Stream transport protocol used (TCP / UDP)
TCPflags	TCP header flag,transport layer protocol is UDP,the feature is 0
pps	Packets/duration
bps	bytes/duration
Mean packets arrived time	duration/packets
tos	TOS from NETFLOW
Mean packet length	bytes/packets

Table 2: bidirectional flow feature

Feature	Feature Description
lport	low port number
hport	high port number
duration	Flow duration
Transprotocol	Stream transport protocol used (TCP / UDP)
TCPflags1	TCP header flag,transport layer protocol is UDP,the feature is 0
TCPflags2	TCP header flag,transport layer protocol is UDP,the feature is 0
pps	Packets/duration
bps	bytes/duration
Mean packets arrived time	duration/packets
Bidirectional Packets ratio	Forward packets/ backward packets
Bidirectional Bytes ratio	Forward bytes/ backward bytes
Bidirectional Packet length ratio	Bidirectional packets length ratio
Bidirectional packets	Forward packets + backward packets
Bidirectional bytes	Forward bytes + backward bytes
tos	Bidirectional TOS OR from NETFLOW
Mean packet length	Bidirectional bytes/Bidirectional packets

3 Methodology

3.1 Auto Adaptive algorithm (AA)

In this paper, we propose an algorithm which can auto adjust the flow metric to adapt the traffic identification. The algorithm is called auto adaptive algorithm(AA). The algorithm's core thought is that different traffic can select different classifier with different flow metric (unidirectional flow or bidirectional flow).
Suppose there are n flow samples, each sample has p features, then construct the n*p flow matrix, as follows:

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

When features number p of the samples are very large which enlarge dimensions of the sample, theoretically, having more features should result in more discriminating power. However, practical experience with machine learning algorithms has shown that this is not always the case. Many learning algorithms can be viewed as making an (biased) probability estimate of a set of features with the class label. This is a complex, high dimensional distribution. Asymmetric routing existing will impact on the traffic identification. So we can consider to adopt auto adaptive method to do with it. In order to depict the method, we have to introduce the H which represent the threshold.

$$H = \frac{Bidirection\_flow\_number}{total\_flow\_number} \tag{2}$$

Definition 2. Optimal threshold: which is used to evaluate the traffic accuracy, it is minimum threshold. When the traffic accuracy is maximum. H is optimal threshold ε.

According to different H, and select H as optimal threshold to enable to obtain the best traffic results, where H is random variable. When $H < \varepsilon$, it will choose unidirectional flow and generate the unidirectional classifier, conversely, it will choose directional flow and generate the directional classifier.

Algorithm 1: AA algorithm

```
// Initialize in the network
A = 0;
for each flowi ∈ [flow1, ..., flown] do
    if H < ε then
        choose unidirectional flow;
    if H >= ε then
        choose bidirectional flow;
        Return the network;
    else
        Goto exit
```

Algorithm AA presents the two kinds of flow metric. The sequence of steps that we show in Figure 1. The procedure mainly set two kinds of dataset for training and testing data set. With these data, we choose AA algorithm to train and test data. The process of machine learning identification is shown in Figure 2:

1.Collecting traffic(Input): Collecting network data from network traffic
2.Selecting traffic features and training data for building traffic classification model(Data Processing): Optimal selecting the known traffic features through the traffic feature selection algorithms. In this paper we only adopt two kinds of feature metric(unidirectional metrics and bidirectional metrics), so extra feature selection method is not added. The traffic classification model is built by training data.
3.Classified the traffic by machine learning algorithm (Output): Using the machine learning identification algorithm to classify network traffic data and generate flow with label.

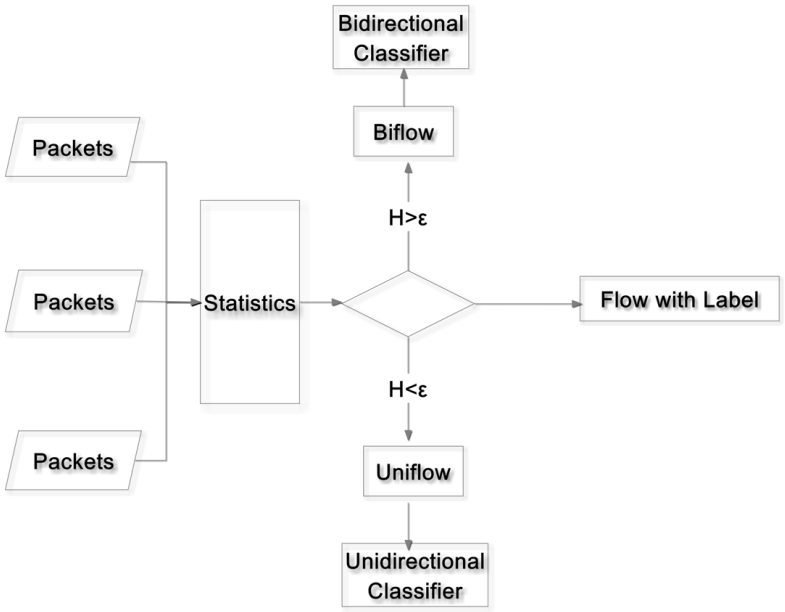


Figure 1: Traffic identification process of AA method

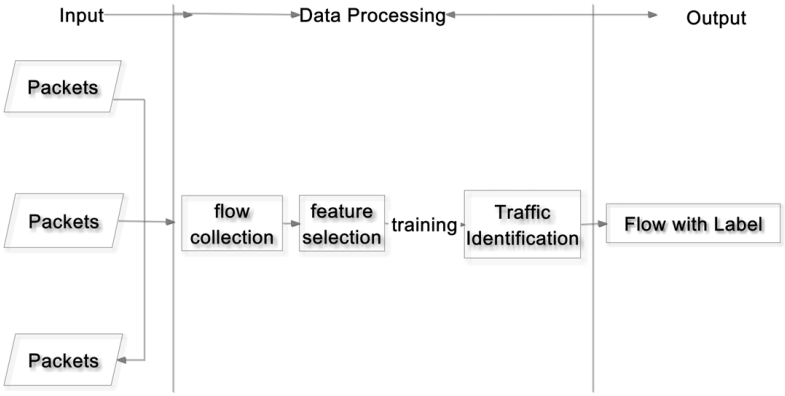


Figure 2: Process of Machine learning, traffic identification

3.2 Algorithm Evaluation

In this paper, we use the routine evaluation standard for verifying the effectiveness of our identification algorithm. The effectiveness of the current flow identification algorithm has the

Table 3: NOC_SET dataset

AppID	Application	Protocal	Flow number	Proportion(%)
1	WWW	HTTP	4943	64.6
2	Bulk	FTP	39	0.5
3	Mail	IMAP,POP3,SMTP	91	1.19
4	P2P	BitTorrent,eDonkey,Gnutella,XunLei	1414	18.5
5	Service	DNS,NTP	433	5.7
6	Interactive	SSH, CVS, pcAnywhere	6	0.08
7	Multimedia	RTSP,Real	20	0.3
8	Voice	SIP,Skype	276	3.6
9	Others	games, attacks	431	5.6

following three concepts evaluation criteria. And the concepts involved are as follows:
-TP (true positive): The flows of application A are classified as A correctly, which is a correct result for the identification;
-FP (false positive): The flows not in A are misclassified as A. For example, a non-P2P flow is misclassified as a P2P flow. FP will produce false warnings for the identification system;
-FN (false negative): The flows in A are misclassified as some other category. For example, a true P2P flow is not identified as P2P. FN will result in identification accuracy loss.

The calculating methods are as follows:
1. Precision: The percentage of samples classified as A that are really in class A

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

2. Recall: The percentage of samples in class A that are correctly classified as A

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3. Overall accuracy: The percentage of samples that are correctly classified

$$Overallaccuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \tag{5}$$

4 Experiment

4.1 Dataset

NOC_SET dataset

In order to validate the method and analyze the impact factor,we adopt NOC_SET as dataset.as shown from table 3. We collected data at southeast university,and the collecting site is a 10G backbone channel on Jiangsu Province border of CERNET. We adopt DPI method to mark flow and generate NOC_SET dataset,and use ourself l7_filter_modify software to label the flow.l7_filter_modify is developed based on L7filter [23], at last, we generate NOC_SET dataset.

LBNL_SET dataset

Table 4: LBNL_SET dataset

AppID	Category	flow number	Proportion
1	80	15000	47.69%
2	110	1400	4.45%
3	25	1350	4.29%
4	139	3300	10.49%
5	993	400	1.27%
6	443	10000	31.8%

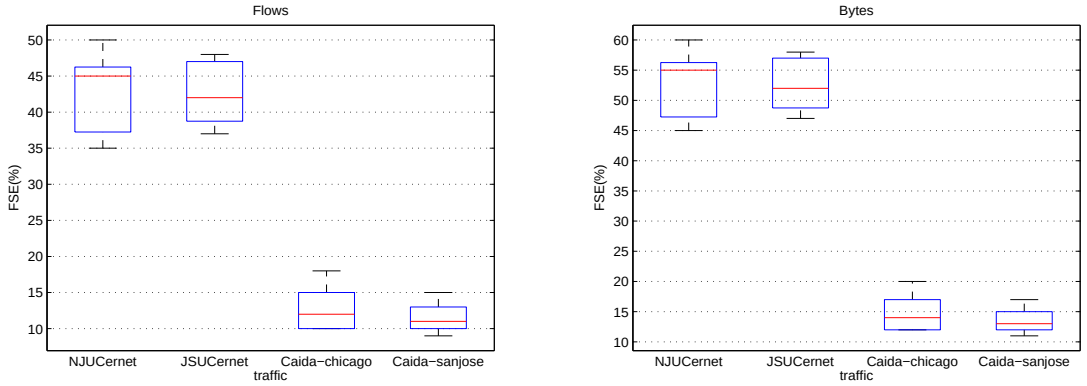
This LBNL_SET data is randomly sampled in several different periods from one node on the internet. The LBNL traffic traces are collected at the Lawrence Berkeley National Laboratory under the enterprise tracing project [24]. The packet traces are obtained at the two central routers of the LBNL network and they contain more than one hundred hours of traffic generated from several thousand internal hosts. The traffic traces are public, but they are completely anonymized, so ascertaining the "ground truth" on the application behind each recorded flow is not possible. Therefore, for this set, we built protocol sets according to the TCP destination port number of each flow, an accepted practice in these cases [25]. We use the traffic traces captured on January 6 and 7, 2005 to obtain the training and the optimization sets. Once again we perform the training by using the most frequently used port numbers in the dataset. Detail $LBNL\_SET$ dataset is shown in table 4.

CAIDA dataset

We built this data set starting from three hour long traces obtained by the Cooperative Association for Internet Data Analysis (CAIDA) [26], and collect at the AMES Internet Exchange (AIX) along an OC48 link on Mar 24, 2011. We use flows extracted from the first hour (corresponding to the interval 16:15-17:00 UTC) to build the training set the optimization set and from the third hour (18:00-18:10 UTC) to buld the evaluation set. As for the previous set, these traces are also anonymized, so port numbers are used as indicators of each protocol. The selection of flows composing the training, optimization and evaluation sets.

Table 5: CAIDA_SET dataset

AppID	Category	flow number	Flow(%)	packets(%)	bytes(%)
1	80	328091	84.69	81.74	81.58
2	110	11539	0.6	0.24	0.25
3	21	28567	3.32	0.03	0.09
4	25	2648	4.57	2.47	2.72
5	4662	2099	0.79	1.34	1.35



(a) Flows (b) Bytes

Figure 3: Comparison of FSEs for traffic

Table 6: the identification Overall accuracy rate AA, Biclassifier, Uniclassifier

Identification	Overall accuracy
AA	99.6742%
Biclassifier	88.2%
Uniclassifier	89.2%

4.2 Impact of asymmetry router on traffic identification:

In this paper, we adopt experimental data based on the NOC-SET data set and CAIDA datas set, use MATLAB tools, WEKA tools and the corresponding algorithm to identify network traffic data [27]. NOC-SET data firstly divided into two test data were 20% and 80% of the test data, and we compared our method that is AA with Biclassifier and Uniclassifier. In order to evaluate and analyze effectiveness of the method about AA. We study traffic identification distribution. In order to analyze asymmetry router, firstly we should remove from the traces any traffic that is inherently asymmetric, such as UDP and ICMP that do not always expect packet recipients to reply, and which would mislead symmetry comparisons if they appear in different magnitudes across networks. TCP background radiation, such as network scanning and probing, can also be a substantial fraction of total inherently asymmetric flows on some links, although it is usually a much lower proportion of bits. We adopt Flow-based Symmetry Estimator(FSE) [28] to evaluate impact degree on traffic, which is a simple method estimate the level of routing symmetry from passively measured flow data. From Figure 3 and Figure 4 we can see different traffic have different FSE, and CAIDA traffic is less. It indicated asymmetry router of CAIDA traffic were more obvious than NOC-SET.

From Table 6 we can see that overall accuracy of AA method traffic is better than biclassifier and uniclassifier,we adopt AA method to classify traffic based NOC-SET data,and select parameter ε=0.5(detailed analysis shown in session F). The data is divided into 9 categories, respectively, WWW, Mail, Bulk, Service, P2P, Interactive, Voice, Multimedia, Others

Table 6 indicates the AA algorithm achieved better result than Biclassifier and Uniclassifier method, moreover. P2P can be seen from Table 7 and the voice of the precision and the recall has greatly improved. The reason for high accuracy is that the proportion of P2P and voice

Table 7: Identification performance for NOC_SET(Precision and Recall)

Category	Algorithm					
	biclassifier		uniclassifier		AA	
	Precision	Recall	Precision	Recall	Precosiin	Recall
WWW	98%	100%	99%	100%	98.5%	99.2%
P2P	58%	100%	75%	100%	93.7%	91.2%
Mail	83%	91.3%	90%	99%	100%	100%
Service	58.90%	100%	70%	99%	90%	90.4%
Inter	84.5%	100%	87%	100%	80%	100%
Multimedia	100%	75%	90%	80%	60%	100%
Voice	35%	50%	45%	55%	37%	50%
Others	44%	46%	48%	77%	45%	60%

account for set of the total is relatively small,the impact of the identification results reduce to a minimum due to the collection of the specimen Caused by imbalance in the ratio.This paper also build NOC_SET dataset which is constructed by bidirectional flow characteristic.

4.3 Comparison of identification algorithm with NOC-SET dataset

Experimental data for the NOC_SET data set (Table 3 as fellows) The analysis data are actual measured IP trace [29], while the traffic flow exits about 40% biflow. NOC_SET dataset is composed by biflow feature.biflow have more information for traffic identification.if use biclassifier to classify the traffic, then the identification result will be improved. In this section, we compare AA algorithm with biclassifier and uniclassifier. Traffic identification result is shown in Table 7. As shown in Table 7, identification result indicates that AA could achieve better accuracy compared with Biclassifier and Uniclassifier.But observing from Inter and Service, identification accuracy of AA is lower than the other method. From Service to Inter types, precision of biclassifier and uniclassifier method is reduced, while the AA is in increments, so that biclassifier and uniclassifier method is easily affected by the number of training samples, while the AA is not vulnerable to the impact of the training Sample dataset. Among three identification algorithm AA, biclassifier and uniclassifier, the overall accuracy of the AA algorithm is highest.

4.4 Comparison of identification algorithm with CAIDA_SET dataset

The data set used in experimental platform: Experimental data for the CAIDA_SET data set (Table 5 as fellows). The analysis data are actual measured IP trace [29]. The two core links are part of an OC192 Tier1 backbone operated by a commercial ISP in the U.S. The first link connects Chicago and Seattle, monitored at an Equinix data center in Chicago. The other one connects San Jose and Los Angeles, monitored at a datacenter in San Jose. On those links, TCP is responsible for about 50% of flows, which was 85% of packets and 93% of bytes on average.UDP carried about 45% of flows (13% of packets and 6% of bytes). We adopted port-based method to mark Flow and generated CAIDA_SET dataset.while the traffic flow exits about 10% biflow. CAIDA_SET dataset is composed by uniflow feature. Biflow have more information for traffic identification. If use biclassifier to classify the traffic, then the identification result will be improved. In this section, we compare AA algorithm with biclassifier and uniclassifier. Traffic identification result is showed in Table 8.

Table 8: Identification performance for CAIDA_SET(Precision and Recall)

Category	Algorithm					
	biclassifier		uniclassifier		AA	
	Precision	Recall	Precision	Recall	Precision	Recall
80	92%	98%	98%	97%	96.5%	98.2%
110	63%	97%	83%	99%	95.7%	92.2%
21	82%	88.3%	92%	98%	99%	99%
25	60.80%	99%	72%	98%	92%	92.4%
4662	82.4%	99%	89%	98%	82.9%	99.2%
Overall Accuracy	65.72%		94.1342%		95.8921%	

Table 9: Identification performance for LBNL_SET(Precision and Recall)

Category	Algorithm					
	biclassifier		uniclassifier		AA	
	Precision	Recall	Precision	Recall	Precision	Recall
80	96%	98%	97%	93%	96.5%	98.2%
110	78%	90%	85%	90%	92.5%	83.2%
25	88%	82.7%	89%	87%	97%	99%
139	59.80%	98%	78%	92%	93%	91.6%
993	86.5%	99%	79%	99%	87%	99%
443	88.5%	99%	89%	99%	84%	99%
Overall Accuracy	68.83%		93.237%		95.861%	

As shown in Table 8, identification result indicates that AA could achieve better accuracy compared with biclassifier and uniclassifier. According to analysis of 4.4 section on traffic result, we can see CAIDA exists the same phenomena which is unbalance sample data. So that biclassifier and uniclassifier method is easily affected by the number of training samples, while the AA is not vulnerable to the impact of the training Sample dataset. Among three identification algorithm AA, biclassifier and uniclassifier, the overall accuracy of the AA algorithm is highest.

4.5 Comparison of identification algorithm with LBNL_SET dataset

We obtained LBNL data from the Lawrence Berkeley National Laboratory, and construct the bidirectional and unidirectional flow metric. We respectively train the two metrics and generate biclassifier and uniclassifier. We compute H value the formula 2 in section 3, and adopt AA method to select classifier which is uniclassifier or biclassifier. The experimental results is shown in table 9. From the results we can see uniclassifier and uniclassifier method is affected by unbalance sample data, while AA method can overcome the problem and improve traffic identification results.

4.6 Impact of ε on traffic identification

In this paper we propose AA method to auto adaptive select classifier(biclassifier or uniclassifier), while threshold ε is a parameter of AA method.ε decide classifiers which were selected, so it is very important for traffic identification. In this section, we will analyze the impact of ε on traffic identification. Detailed experiment method is adopting AA method proposed by varying from ε[0.1,1] based on three dataset(NOC_SET, CAIDA, LBNL_SET). From Figure 4 we can

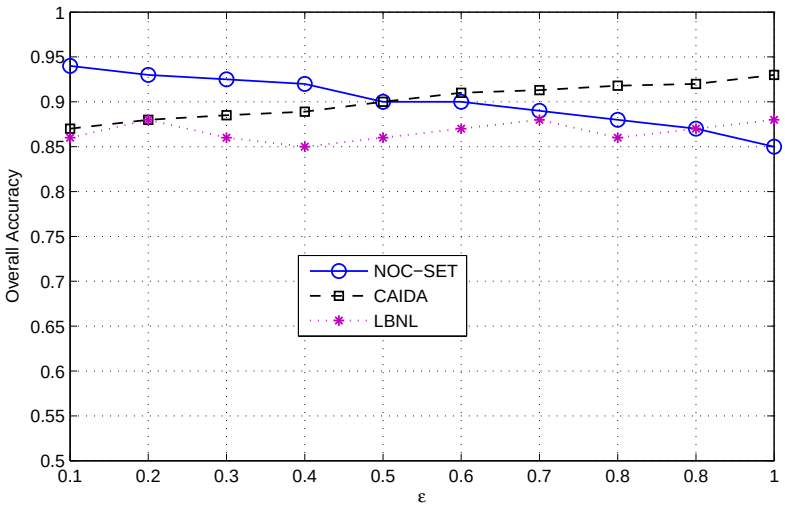


Figure 4: The identification results with ε

see overall accuracy of CAIDA and NOC_SET have biggest change happened when ε vary from 0.1 to 1. Overall accuracy of CAIDA shows an increasing tendency, while NOC_SET is descending. The possible reasons why is that CERNET network contain more symmetry routing, while asymmetry routing is less. Collection point of CAIDA data exist more asymmetry routing. Thus when threshold ε is very small, more opportunity will be selected by biclassifier. Just as mentioned that collection point of NOC_SET is CERNET network containing more symmetry routing, which will have more bidirectional flow metrics, so NOC_SET showed an descending tendency and when ε =0, overall accuracy is maximum.ε =0.5, overall accuracy of CAIDA and NOC_SET is equal. LBNL have not obvious asymmetry routing. So overall accuracy is gentle.

5 Conclusion

In this paper we propose auto adaptive algorithm, and on this basis, the introduction of biclassifier and uniclassifier, and adopt the improved AA method to classify traffic for MOORE_SET as data set, moreover, compare with two other methods which is the biclassifier and uniclassifier method, the results show that, AA method are greatly improved on identification accuracy, to further prove AA method is effective, this paper collect the data in Jiangsu provincial network border and organize trace into flow record such as data sets NOC_SET, the experimental results show that: AA method has high identification accuracy,and we analyze the impact of ε on traffic identification and find ε=0.5 which can be considered as the fixed value, traffic results will be better.

Acknowledgments

This paper is supported by Education Department of Henan Province Science and Technology Key Project Funding (14A520065) and Research Innovation of Zhoukou Normal University (zknuA201408).

Bibliography

[1] T. Karagiannis, K. Papagiannaki, M. Faloutsos (2005); Blinc: multilevel traffic classification in the dark, in: *ACM SIGCOMM Computer Communication Review*, ACM, 35: 229–240, DOI:10.1145/1080091.1080119.

[2] A. Moore, K. Papagiannaki (2005); Toward the accurate identification of network applications, *PAM'05 Proceedings of the 6th international conference on Passive and Active Network Measurement*, 41–54.

[3] A. Moore, D. Zuev (2005); Internet traffic classification using bayesian analysis techniques, in: *ACM SIGMETRICS Performance Evaluation Review*, ACM, 33:50–60, DOI:10.1145/1064212.1064220.

[4] L. Bernaille, R. Teixeira, K. Salamatian (2006), Early application identification, in: *Proceedings of the 2006 ACM CoNEXT conference*, ACM, DOI:10.1145/1368436.1368445.

[5] Wolfgang John, Sven Tafvelin (2007); Differences between in- and outbound internet backbone traffic, in: *Proceedings of Terena Networking Conference*, TERENA, 1-14.

[6] Hotpotatorouting, http://en.wikipedia.org/wiki/Hot-potato_routing.

[7] N. Williams, S. Zander, G. Armitage, Evaluating machine learning algorithms for automated network application identification, Center for Advanced Internet Architectures, CAIA, *Technical Report* 060410B, DOI:10.1.1.84.7170.

[8] N. Williams, S. Zander, G. Armitage (2006), A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification, *ACM SIGCOMM Computer Communication Review* 36(5):5–16, DOI: 10.1145/1163593.1163596.

[9] Z. Li, R. Yuan, X. Guan (2007), Accurate classification of the internet traffic based on the svm method, in: *Communications, 2007. ICC'07. IEEE International Conference on*, IEEE, ,1373–1378, DOI: 10.1109/ICC.2007.231.

[10] P. Teufl, U. Payer, M. Amling, M. Godec, S. Ruff, G. Scheikl, G. Walzl (2008), Infect-network traffic classification, in:*Networking, 2008. ICN 2008. Seventh International Conference on*, IEEE, 439–444, DOI: 10.1109/ICN.2008.42.

[11] T. Kiziloren, E. Germen (2007), Network traffic classification with self organizing maps, in: *Computer and information sciences*, 2007. iscis 2007. 22nd international symposium on, IEEE, 1–5, DOI: 10.1109/ISCIS.2007.4456852.

[12] Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, Y. Choi (2010), Internet traffic classification demystified: on the sources of the discriminative power, in: *Proceedings of the 6th International COnference*, ACM, DOI: 10.1145/1921168.1921180.

[13] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, K. Lee (2008); Internet traffic classification demystified: myths, caveats, and the best practices, in:*Proceedings of the 2008 ACM CoNEXT conference*, ACM, DOI: 10.1145/1544012.1544023.

[14] J. Erman, M. Arlitt, A. Mahanti (2006), Traffic classification using clustering algorithms, in: *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, ACM, 281–286, DOI: 10.1145/1162678.1162679.

[15] V. Carela-Espanol, P. Barlet-Ros, J. Solé-Pareta (2009), Traffic classification with sampled netflow, DOI:10.1.1.390.5780.

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

[17] A. Callado, C. Kamienski, G. Szabó, B. Gero, J. Kelner, S. Fernandes, D. Sadok (2009), A survey on internet traffic identification, *Communications Surveys & Tutorials*, IEEE, 11(3):37–52.

[18] M. Zhang, W. John, K. Claffy, N. Brownlee (2009), State of the art in traffic classification: A research review, in:*PAM '09: 10th International Conference on Passive and Active Measurement, Student Workshop*, Seoul, Korea.

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

[20] Z. Mao, L. Qiu, J. Wang, Y. Zhang (2005), On as-level path inference, in: *ACM SIGMETRICS Performance Evaluation Review*, ACM, 33:339–349.

[21] Y. He, M. Faloutsos, S. Krishnamurthy (2004), Quantifying routing asymmetry in the internet at the as level, in: *Global Telecommunications Conference*, GLOBECOM'04. IEEE, 3: 1474–1479.

[22] W. John (2008), On measurement and analysis of internet backbone traffic, Thesis for the degree of Licentiate of Engineering, a Swedish degree between M.Sc. and Ph.D., Chalmers University of Technology.

[23] J. Levandoski, E. Sommer, M. Strait, et al.(2008), Application layer packet classifier for linux, http://l7-filter.sourceforge.net/.

[24] *** Lbnl/icsi enterprise tracing project, http://www.icir.org/enterprisetracing.

[25] T. Karagiannis, A. Broido, M. Faloutsos, et al. (2004), Transport layer identification of p2p traffic, in: *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, ACM, 121–134, DOI: 10.1145/1028788.1028804.

[26] *** The cooperative association for internet data analysis(caida), http://www.caida.org.

[27] T. Nguyen, G. Armitage (2006), Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks, in: *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, IEEE, 369–376, DOI: 10.1109/LCN.2006.322122.

[28] W. John, M. Dusi, K. Claffy (2010), Estimating routing symmetry on single links by passive flow measurements, in: *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ACM, , 473–478, DOI: 10.1145/1815396.1815506.

[29] *** IP Trace Distribution System, http://iptas.edu.cn/src/system.php.