The Study of Network Traffic Identification Based on Machine Learning Algorithm

DONG Shi*‡§, ZHOU DingDing†, DING Wei‡*
*School of Computer Science and Engineering
Southeast University,Nanjing, China
Email:shdong@njnet.edu.cn
†Dept.of Laboratory and equipment management
Zhoukou Normal University,Zhoukou,China
Email: zdd@zknu.edu.cn
‡Eastern China (North) Regional Network Center
§School of Computer Science and Technology
Zhoukou Normal University,Zhoukou,China

***Abstract*—Network traffic identification is one of the hot research fields for network management and network security; machine learning is an important method during the network traffic identification research.this paper describes the current situation and common methods of network traffic identification, at the same time this paper also states the currently popular Machine learning methods. We compared and evaluated the supervised and unsupervised classification and clustering algorithms, the experiment results show that feature selection algorithm has great effect on supervised machine learning and DBSCAN algorithm which belongs to unsupervised clustering algorithm has great potential in precision.**

Keywords—network management; traffic identification; Machine learning; DBSCAN

I. INTRODUCTION

Network traffic identification is an important application research direction for network management and measure, the current network traffic identification methods roughly can be classified into four categories.(1)port-based method;(2)DPI(Deep packets inspection);(3)host behavior method [1];(4)flow-based method based on machine learning.Besides the machine learning methods are divided into supervised and unsupervised machine learning.Of course,there are also individual QOS quality of service features for classification [2]. Through the discussion and research of machine learning which has applied to network traffic identification, the paper has compared with the different identification method based on supervised and unsupervised machine learning.Besides it has analyzed the element which has effect on the supervised machine learning.This paper is organized as follows. Section II presents related work. In Section III and IV,algorithm evaluation is introduced and experiment results are prsented.finally,Section V draws our conclusions.

II. RELATED WORK

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 port-based method, this method used wellknown port numbers to identify Internet traffic.This was successful because many traditional applications use fixed port numbers assigned by IANA,but This technique has been shown to be ineffective for some applications such as the current generation of P2P applications.So Payload-based Analysis technology was proposed to overcome the shortcoming of port-based, which adopt method based on deep packet detection methods, but this method has still drawbacks that it can't cope with some encrypted traffic and can't obtain the new service type. Recently traffic identification and classification have new method with a number of new applications and service increasing, Machine learning methods have been applied to the traffic identification.

A. ML Algorithm

The definition of machine learning:Studying computer how to simulate or realized human learning behavior and obtaining the new knowledge or skills, reorganizing the existing knowledge structure so as to improve their performance continuously. Nowadays,the popular identification method mainly use the machine learning to identify traffic.ML can be divided into supervised, unsupervised, semi-supervised machine learning algorithms according to labeled training dataset and unlabeled training dataset. Supervised machine learning:the training data contains input vectors and the corresponding target(labeled sample) such as classification, association rules,and regression. Unsupervised machine learning: the training data does not contain the labeled sample such as clustering (Cluster), Density estimation and Visualization.According to the classification method,it also can be divided into Bayesian classification method, decision tree classification, neural network classification methods and clustering methods.

Introduce machine learning algorithms:

Bayes:It includes Naive Bayes, Bayes Network and so on. Moore [3] uses the method to classify network traffic earlier which is belong to Naive Bayes algorithm. The paper [4]

TABLE I
MODELING TIME AND RATE OF CLASSIFICATION FOR COMMON MACHINE LEARNING ALGORITHMS

Machine learning algorithm	Type of machine learning	The average modeling time	Classification accuracy
BAYES	Supervised machine learning	4.5s	95%
SVM	Supervised machine learning	404.83s	99.3194%
C4.5	Supervised machine learning	94.52s	99.7815%
KMEANS	Unsupervised machine learning	60s	95.1%
Two-stage classification algorithm	Unsupervised machine learning	–	97.8%
Neural network (SOM)	Supervised machine learning	–	96.0%

adopted several Bayesian methods to identify P2P,experiment results show that K2,TAN and BAN has high precision and less time, it is an ideal classification algorithm. However, this method is a learning method based on probability, it has potential instability when over-reliance on the distribution of sample space. Moore adopts FCBF feature selection method and use the estimated technology,the result show that classification accuracy has improve from 65% to 95%. [5].

SVM (Support Vector Machine):paper [5] compared the FCBF + NBK and SVM algorithms and drew the conclusion that SVM algorithm has better classification accuracy than NBK without using any feature filtering strategy,and it can avoid interference which cause by the uncertainty factor effectively, besides it has obvious advantage in dealing with traffic classification. The paper [6]discussed and compared two forecasting methods between AR and SVM, experiments show that the function of improved SVM mold attack identification has lower false alarm rate than AR mold. The paper [7] adopts 15 kinds of algorithm to identify network traffic, through comparing modeling time,test time and described simplicity; it concludes that C4.5 is most suitable for network traffic identification. Alshammari R[8] use RIPPER and C4.5 method to classify traffic,experiment show the C4.5 algorithm has better advantage in inspecting speed and rate of fault definition than RIPPER algorithm. In addition, C4.5 algorithm can avoid the effect which bring by change of network flow distribution, but it can't achieve genuine online classification in network traffic area[9].

Kmeans, EM, FCM:Kmeans clustering algorithm is one of the fastest and simplest. FCM is a very effective method for unsupervised fuzzy clustering, It can obtain the satisfied result even for the variable which is hard to classify clearly and easy to realize.[10]. Matti Hirvonen [11] used network traffic classification method of two-phases which had been training by Kmeans method, the experiment shows that the accuracy of classification has achieved 97.8%. Identification system proposed in paper[12]includes two phases(offline learning phase and online identification phase). This system supports three clustering algorithms including Kmeans, DBSCAN and Kmedoids, it also supports different traffic statistical information.Experiment shows that the identification rate could reach about 90% for application layer traffic,especially for P2P application. it is more efficient than port-based and is superior to the method based on payload-based.

Neural network: Peter Teufl and his colleague [13] proposed a Infect framework identification tools, this software can achieve feature selection method, using two supervised machine learning algorithms: SVM is a supervised variant of self-organizing map (SOM). Generally, SOM is a better choice if it is visualized or analyzes all database which used in training, The paper [14] proposed a classifier based on SOM neural network which is used for distinguish three kinds of network traffic type,port scan,high traffic download and other types. Experiments show that it has a better classification result.Paper[15] review recent achievements and discuss future directions in traffic classification, along with their trade-offs in applicability, reliability, and privacy. The process of machine learning classification is shown in Figure 1:

1.Collecting traffic: Getting network data from network traffic.
2.Selecting the characteristics of traffic: Optimal selecting the known traffic attribute through the traffic feature selection algorithms.
3.Classified the traffic sample by machine learning algorithm: Using the machine learning classification algorithm to classify network traffic data.

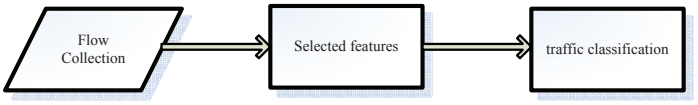

Fig. 1. Process of traffic classification based on Machine learning

III. 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 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 classification;
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 classification 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:
Precision:The percentage of samples classified as A that are

really in class A

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Overall accuracy:The percentage of samples that are correctly classified

$$Overallaccuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \quad (2)$$

IV. Experiment

This section evaluates the effectiveness of both the supervised and unsupervised classification algorithms.and the data sets used in this study are outlined.

A. Dataset

MOORE-SET: The dataset described originally by Moore el [16] was used for the experiment. This data was randomly sampled in several different periods from one node on the internet. This site was shared by about 1,000 researchers, technicians and management staff of three research institutions, and connected to the Internet through a full-duplex Gigabit Ethernet link. All full-duplex traffic on this connection was captured in a full 24 hours period, so the original traffic-set contained all full duplex traffic connected the node in both link directions.The number of flows and the proportion of the various types of network traffic are shown in Table II.

Auckland-SET: The Auckland-dataset [17] consists of all traffic measured during the 72 hour period on March 16, 2001 at 06:00:00 to March 19, 2001 at 05:59:59 from the Auckland IV trace.Detailed data are shown in Table IV.

B. Comparison of supervised classification

In this section we studied several supervised classification algorithms(Bayesian,Decision tree,SVM).The comparison of classification accuracy and time efficiency, which adopt the selection strategy and doesn't adopt the selection strategy.MOORE-SET has 249 features totally, but not all the features are effective for for supervised machine learning classification algorithms, so it become a topic in traffic classification by choosing several superiority attributes, different attributes of selection algorithm has different effect on accuracy rate and time efficiency of traffic classification.Through adopting SymmetricalUncertAttributeSetEval estimation algorithm and FCBF feature selection method,Finally,the method select feature number(4,72,108,91,155,202,113,50 and 266) from 249 features in MOORE dataset. From Table III,we can see classification time also has been shorten with the selection feature reducing, when it adopts feature selection algorithm, the efficiency of classification time has been greatly improved. But the classification accuracy depends on the feature selection algorithm whether match and suitable or not. Such as classification algorithm adopts Bayesian and selected algorithm adopt SymmetricalUncertAttributeSetEval + FCBF method, the accuracy rate has been greatly improved. But if the SVM and decision tree algorithm use this algorithm, the accuracy rate will be reduced.This is because SVM and decision tree have feature reducing strategy which can bring conflict with FCBF method.

C. Comparison of unsupervised clustering algorithm

When adopt flow characteristic (Total packets,Average size of packets,Average payload,Mean Inter-arrive time,Translate bytes) from Auckland-SET to evaluate the unsupervised clustering algorithm, it gets a better result. The experimental platform used size of dataset is less than 8,000 connections, because this numerical can ensure that the modeling time which set up by Auto-class algorithm can maintain at 4 to 10 hours. From Figure 2,By comparing with HTTP, P2P, POP3

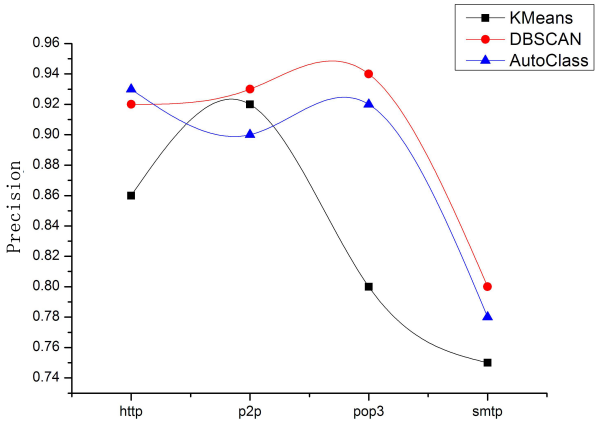

Fig. 2. Precision of Unsupervised clustering algorithm

and SMTP,we can get the conclusion that DBSCAN algorithm has higher precision and three precisions of them have passed 90%.Besides another experiment show that constructing model time of three algorithms is different. Model K-Means algorithm established needs 1 minute, DBSCAN needs 3 minutes ,meanwhile, Autoclass needs 4.5 hours

V. Conclusion

This paper has studied and analyzed the machine learning algorithm for network traffic identification and mainly studied unsupervised and supervised machine learning. Through experiment on the classification algorithm of two different datasets, comparing the classical unsupervised and supervised algorithm; the experiment result show that the supervised machine learning is greatly influenced by feature selection algorithm. The suitable feature selection algorithm can improve the accuracy and time efficiency of classification algorithm. By comparing several unsupervised machine learning algorithm(cluster algorithm), results show that DBSCAN algorithm has great potential and has more advantage than other two kinds of algorithms in precision, besides the modeling time is between the K-Means method and DBSCAN method.

Acknowledgment

This project was supported in part by: State Scientific and Technological Support Plan Project of China under Grant No.2008BAH37B04, National Basic Research Program

TABLE II
$MOORE\_SET$ DATASET

AppID	Category	Application	Flow number	Proportion
1	WWW	HTTP,https	328091	86.91
2	BULK	FTP	11539	3.056
3	MAIL	pop3,Imap,Smtp	28567	7.567
4	DB	Sqlnet,Oracle	2648	0.701
5	SERV	Dns,Ntp,Ldap	2099	0.556
6	P2P	Kazaa,Bittorrent,Gnutella	2094	0.555
7	ATTACK	Worm and virus Attacks	1793	0.475
8	MULT	Media Player,Real	1152	0.305
9	INT	ssh,klogin,Telnet	110	0.029
10	GAME	Halflife	8	0.002

TABLE III
OVERALL ACCURACY AND TIME EFFICIENCY OF CLASSIFICATION

Classification method	Overall accuracy		time	
	Full features	Optimized features	Full features	Optimized features
Bayes	57.8933%	96.9835%	6.36s	0.27s
SVM	99.1353%	87.8052%	356.53s	4.88s
C4.5	99.6742%	99.638%	58.67s	0.77s

TABLE IV
AUCKLAND DATASET

AppID	Application	flow number	Byte	Byte ratio
1	HTTP	1132920	23,693,723,103	47.3%
2	P2P	41478	17,578,995,934	35.1%
3	IMAP	2955	228,156,060	0.5%
4	POP3	3674	72,274,560	0.1%
5	SMTP	46882	2,997,244,939	6.0%
6	MYSQL	8105	23,824,936	0.0%
7	OTHER	41239	658,046,156	1.3%

of China (973) under Grant No.2009CB320505, and National Natural Science Foundation of China under Grant No.60973123.

REFERENCES

[1] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark," in *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 229–240, ACM, 2005.

[2] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for qos: a statistical signature-based approach to ip traffic classification," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 135–148, ACM, 2004.

[3] A. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, pp. 50–60, ACM, 2005.

[4] L. jun and Z. shunyi, "Peer-to-peer traffic identification using bayesian networks," *Journal of Applied Sciences*, vol. 27, no. 2, pp. 124–130, 2009.

[5] X. Peng, L. Qiong, and L. Sen, "Internet traffic classification using support vector machine [j]," *Journal of Computer Research and Development*, vol. 3, 2009.

[6] Z. Li, R. Yuan, and X. Guan, "Accurate classification of the internet traffic based on the svm method," in *Communications, 2007. ICC'07. IEEE International Conference on*, pp. 1373–1378, IEEE, 2007.

[7] Y. Ma, Z. Qian, G. Shou, and Y. Hu, "Study on preliminary performance of algorithms for network traffic identification," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 1, pp. 629–633, IEEE, 2008.

[8] R. Alshammari and A. Zincir-Heywood, "Investigating two different approaches for encrypted traffic classification," in *Privacy, Security and Trust, 2008. PST'08. Sixth Annual Conference on*, pp. 156–166, IEEE, 2008.

[9] X. peng and lin sen, "Internet traffic classification using c4.5 decision tree," *soft journal*, vol. 20, no. 10, pp. 2692–2704, 2009.

[10] Y. Feng and Z. shunyi, "Based on semi-supervised learning network traffic classification," *Computer Engineering*, vol. 35, no. 12, pp. 90–94, 2009.

[11] M. Hirvonen and J. Laulajainen, "Two-phased network traffic classification method for quality of service management," in *Consumer Electronics, 2009. ISCE'09. IEEE 13th International Symposium on*, pp. 962–966, IEEE, 2009.

[12] S. Xin and Y. Jianhua, "Traffic identification system for comparative analysis of clustering algorithm," *Computing Technology and Automation*, vol. 27, no. 3, pp. 1–6, 2008.

[13] P. Teufl, U. Payer, M. Amling, M. Godec, S. Ruff, G. Scheikl, and G. Walzl, "Infect-network traffic classification," in *Networking, 2008. ICN 2008. Seventh International Conference on*, pp. 439–444, IEEE, 2008.

[14] T. Kiziloren and E. Germen, "Network traffic classification with self organizing maps," in *Computer and information sciences, 2007. iscis 2007. 22nd international symposium on*, pp. 1–5, IEEE, 2007.

[15] A. Dainotti, A. Pescape, and K. Claffy, "Issues and future directions in traffic classification," *Network, IEEE*, vol. 26, no. 1, pp. 35–40, 2012.

[16] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," *Passive and Active Network Measurement*, pp. 41–54, 2005.

[17] "Auckland data sets." http://www.wand.net.nz/wand/wits/auck/.