I.J. Wireless and Microwave Technologies, **2012, 2, 7-11**
Published Online April 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012.02.02

Available online at http://www.mecs-press.net/ijwmt

The Comparison of Machine Learning Algorithms on Online Classification of Network Flows

Keji Wei[a], Shaolong Cao[b],.Jian Yu[a]

[a]*School of Electronic and Information Engineering, Xi'an Jiaotong Univ, P.R.China*
[b]*School of Science, Xi'an Jiaotong Univ, P.R.China*

Abstract

Online classification of network flows is a process that captures packets generated by network applications and identifies types of network applications (or flows) in real time. There are three key issues about online classification: observation window size, feature selection, and classification algorithms.

In this paper, by collecting five types of typical network flow data as the experiment sample data, the authors found observation window size 7 is the best for the sample data and most classifiers. The authors proposed a full feature set based on the standard feature set which reflects statistical features of network flows. Using five commonly used feature selection methods, the authors identified the most effective features could be reduced from 56 original features to 11 effective features. Lastly, according to special need for online classification, the authors studied 11 different classifiers on their classification accuracy, model construction time, and classification speed. The results show that C4.5 and JRip are the two best algorithms for online classification.

Index Terms: Online classification; network flow; statistical feature; feature selection; classifier

1. Introduction

In the modern information society, network has become a necessary implement in our daily lives. The common classification method is based on the use of well known ports. However, this method is inclined to get inaccurate estimates of the number of traffic carried by different applications because of technology of dynamically allocation of ports or users intention to adopt non-default port numbers[1,2]. A classification method that relies on the full packet payload can approach almost 100% accuracy. But the limitations of this method are obvious: the privacy problem about checking users' private data; the high price of space and computing time.

Considering the flaws of above mentioned classification methods, machine learning based on statistical features of network flows was applied to network traffic classification. Offline classification only takes care of the accuracy of classifiers but neglects the speed and response time. Online classification highlights practical application to all important indicators of online classification performance, including accuracy, latency, and throughput, are taken into account. In addition, it is critical for online classification to meet latency and throughput requirements.

The remainder of the paper is organized as follows. In Section 2, experimental sample data and statistical features of network flow data are formulated. In Section 3 the authors finished the preliminary work about observation window size and identifying the most effective features of network flow data. In section 4, the authors compare the classification results of 11 classifiers and propose the best classifiers for online classification. In last section, the authors summarize the findings of this paper and give directions for future work.

2. Experimental Data and Statistical Feature of Network Flows

In order to ensure the generalization of our research, the authors select 5 different data sets from 3 different locations over 4 different years. Auckland-vi-20010611 comes from four trace documents including 20010611-120000-0,20010611-120000-1,20010611-180000-0 and 20010611-180000-1, which costs about 12 hours to capture. Loc3-200311-1925, costs 86 hours to capture, is merged from 20 trace documents from loc3-20031119 to loc3-20031125. WaikatoI-20031210, WaikatoI-20041208, and WaikatoI-20050812 cost 24 hours each[12].

The authors extend standard feature set based on two assumptions: first, the features based on effective load size of packets can better represent the behavior of network flows than the features based on IP packet size. Second, it is generally accepted that a few of packets with effective load after TCP three handshakes protocol represent the negotiation stage of application layer protocol. Following experiments will justify the extended feature set.

So the authors use a full feature set including 56 features (which was shown in [3,4,5,6,7,8]) in this paper.

3. Perliminery Work

In this section, the authors explore the problem of observation window size. The authors compare various experiment results of different OWS(observation window size) configuration. Next, we extend the standard feature set and identify the most effective features of network flow data given feature selection methods and sample data.

Generally, the bigger observation window size is, the more trace information can be conveyed. However, big observation window size will aggravate the retard of a system. That is, big OWS tends to increase latency. So we will determine an appropriate OWS without sacrificing classification accuracy significantly.

After several experiments, the authors find that observation window size 7 attains a good balance between accuracy and efficiency for all 11 online classifiers.

Feature selection aims at selecting a subset of feature set so that the accuracy of classifiers based on selected or reduced feature set will be close to or a little lower than the accuracy based on full feature set. The key is to search an optimal subset within the full feature set.

In this paper firstly the authors use specific searching strategies to select feature subset. Secondly the authors evaluates the selected subset based on evaluation functions. In order to enhance the speed of online classification, the authors use filter category of evaluation functions which is shown in Table 1.

Table 1 evaluation functions and subset searching strategy

Algorithm(evaluation function)	subset searching strategy
CON	Greedy
CFS	Best First
CHI	Genetic
GR	Ranker
ReliefF	

Using the combination of 5 experimental data and 5 evaluation functions, we have 25 experimental settings to make feature selection on 56 features. The authors can see there are 11 features that are chosen as effective features more than 13 times in 25 settings. So our finding is that these 11 features are effective features we should use in online classification.

4. Classification Algorithms

Firstly, for each experiment set, the authors take CFS、CON、CHI、GR、ReliefF feature choose method to get experiment set. Then from each of these experiment sets we take 11 different kinds of machine learning method by get accuracy, precision, recall, classification speed, testing time to get the more suitable one.

From figure1, this comparison means for each machine learning method to give many feature choose method in different data set to get the average accuracy. We could see C4.5(98.507%),RandomTree(98.505%),BayesNet(96.862%), PART(98.522%),JRip(98.209%), AdaBoost+C4.5(99.014%), IBK(98.28%) have a higher accuracy(>96%). In our experimental environment, except NB, NBK, Adaboost+NB, other methods have little differences in classification accuracy.

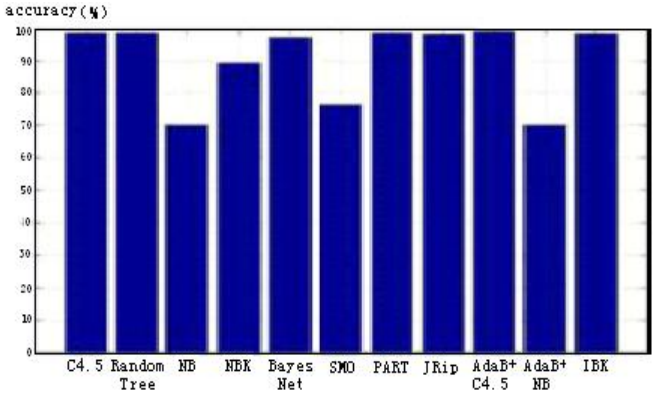

Figure 1 Average Accuracy of 11 classifiers

Classification accuracy is only one part of our evaluation system, therefore, the authors need to illustrate precision and recall of machine learning method to get further study to make it reliable. From experiment we can find that the precision of each C4.5, RandomTree, PART, JRip, AdaBoost+C4.5, IBK except telnet are more than 95%. After all, in our experiment, all of machine learning method have a better classification precision.

Recall reflects the proportion of classifier forecast correctly to all. From result, give us different learning method for each category's average classification recall. The results approach the precision above. In this way, we can believe that each machine leaning method have a well recall.

From table 2, we get 11 machine learning methods in different data set to get the average classification speed and classifier average construction time. The classification speed of 11 method in descending order is : JRip, C4.5, PART, RandomTree, AdaBoost+C4.5, SMO, BayesNet, NB, AdaBoost+NB, NBK, IBK: the classifier average construction time in asending order is IBK, NB, NBK, RandomTree, BayesNet, C4.5, PART, AdaBoost+NB, SMO, AdaBoost+C4.5, JRip. The authors can find that the classification speed of JRip, C4.5, PART, RandomTree is no less that 5000 per second, and is much better that others in online classification.

Table 2 Average construction time and average classification speed of 11 classifier

Classifier	description	Construction time (second)	Classification speed (N/sec)
C4.5	C4.5 Decision Tree	1.2676	62931
RandomTree	Random Tree	0.532	47966
NB	Naive Bayes	0.216	13536
NBK	Naive Bayes based on Kernel density	0.2528	992.95
BayesNet	Bayesian Networks	0.95	27563
SMO	Sequential Minimal Optimization	6.524	35088
PART	Partitions Decision Tree	3.3468	55824
JRip	a fast and efficient RIPPER algorithm	14.707	76814
AdaBoost+C4.5	AdaBoost algorithm based on C4.5	10.757	37621
AdaBoost+NB	AdaBoost algorithm based on Naive Bayes	5.5016	10973
IBK	K-Nearest Neighbor	0.0296	493.64

All in all, we find that classifiers such as C4.5, RandomTree, PART, JRip, AdaBoost+C4.5, IBK get a classification accuracy greater than 96%, precision more than 95% and recall more than 94% at the same level. However, classifiers have large difference in classification speed and construction time. C4.5 and JRip have highest classification speed. The construction time of RandomTree, NB, NBK, BayesNet, IBK is less than 1 second. C4.5 is 1.2676 second and JRip is 14.707 second. From all of the analysis made above, it is easy to identify that C4.5 and JRip have a high classification accuracy（>98%）, and the classification speed is fast（>60000per /sec）, even though Jrip have a longer classifier construction time that others, but only 14.707 second. At hence, the authors synthesize accuracy, speed and construction time, in our experimental environment, C4.5 and JRip is more suitable for online network traffic classification.

5. Summary and Future Work

In this paper the authors set up an experimental environment in which the authors collected five typical

network flow datasets and proposed 56 features to represent statistical characteristics of network flow data. Under this experimental environment the authors studied three key issues of online classification: OWS, feature selection, and classification algorithms.

Firstly the authors analyzed the relationship between OWS and classification accuracy. Our experimental results show that OWS 7 is the best for online classification. Secondly, the authors extended standard feature set to a full feature set by including more flow-related features. The authors use five feature selection methods to reduce a full feature set of 56 features to only 11 effective features. Lastly, the authors evaluated the accuracy, construction time, and classification speed of 11 commonly used classifiers in machine learning field. Our experimental results show that C4.5 and JRip are two more suitable classifiers for network flow online classification.

The findings of this paper are valuable and promising but there are some avenues to be further explored. Apart from 11 classifiers in this paper, there are more classifiers which are worth further studying. In addition, our model building stage is based on offline trace data, we need to explore the possibility of building models using real time feeding network flow data. This way the authors can get a more practical online classification system.

Reference

[1] A Moore,K Papagiannaki.Toward the Accurate Identification of Network Applications[C].Passive and Active Measurements Workshop, Boston, USA, 2005.
[2] A Moore,D Zuev.Internet traffic classification using Bayesian analysis techniques[C].Proceedings of the 2005 Conference on Measurement and Modeling of Computer Systems,New York,2005: 50-60.
[3] N Williams,S Zander,G Armitage.Evaluating Machine Learning Algorithms for Automated Network Application Identification[R]. CAIA Technical Report, April 2006.
[4] YL Ma,ZJ Qian.Study of information Network Traffic Identification Based on C4.5 Algorithm[C]. WiCOM '08. 4th International Conference,2008:1-5.
[5] W Yu, SZ Yu.Supervised Learning Real-time Traffic Classifiers[J].Journal of Networks, 2009, 4(7):622-629.
[6] W Yu, SZ Yu.Machine Learned Real-time Traffic Classifiers[C].Intelligent Information Technology Application,2008,3:449-454.
[7] YL ma,ZJ Qian,GC Shou.Study on Preliminary performance of Algorithms for Network Traffic Identification[C].Computer Science and Software Engineering,2008,1:629-633.
[8] N Williams,S Zander.A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification[C].Special Interest Group on Data Communication Computer Communication Review,2006,36(5):5-16.
[9] J Teixeira.Feature Selection with a General Hybrid Algorithm[D].Ottawa:SITE,2004.
[10] NetMate[OL]. http://www.ip-measurement.org/.
[11]IANA[EB/OL].http://www.iana.org/assignments/port-numbers
[12] NLANR traces[OL]: http://www.wand.net.nz/wits/.
[13] Weka 3.6.1[OL]. http://www.cs.waikato.ac.nz/ml/weka/.