Efficient Flow based Network Traffic Classification using Machine Learning

Jamuna .A*, Vinodh Ewards S.E**
*(Department of Computer Science and Engineering, Karunya University, Coimbatore-114)
** (Assistant Professor, Department of Computer Science and Engineering, Karunya University)

Abstract

Traffic classification based on their generation applications has a very important role to play in network security and management. The port-based prediction methods and payload-based deep inspection methods comes under Traditional methods. The standard strategies in current network environment suffer from variety of privacy issues, dynamic ports and encrypted applications. Recent research efforts are focused on traffic classification and Machine Learning Techniques and of which machine learning is used for classification. This paper conducts a flow based traffic classification and comparison on the various Machine Learning (ML) techniques such as C4.5, Naïve Bayes, Nearest Neighbor, RBF for IP traffic classification. From this C4.5 Decision Tree gives 93.33% accuracy compare with other algorithms. The two methods are used Full Feature selection and Reduced feature set for classification. From this classification the Reduced feature selection gives good result.

Keywords – Bayes Net, C4.5 Decision Tree, Navie Bayes,, Machine Learning (ML), Nearest Neighbor(NN), Payload based – Deep Inspection Methods, RBF, Traffic Classification.

I. INTRODUCTION

Network Traffic classification has strained important consideration over the past few years. Classifying traffic flows by their generation applications plays very essential task in network security and management, such as, lawful interception and intrusion detection, Quality of Service (QoS) control. Conventional traffic classification methods [1] include the port-based prediction methods and payload-based deep inspection methods. In current network environment, the conventional methods suffer from a number of practical problems such as dynamic ports and encrypted applications. Recent research efforts have been absorbed on the application of machine learning techniques to traffic classification constructed on flow statistical features. It can instinctively search for and describe practical structural patterns in a supplied traffic dataset, which is helpful to logically conduct traffic classification. The flow statistical feature founded traffic classification can be understood by using supervised classification algorithms or unsupervised classification (clustering) algorithms. Every ML algorithm has a different approach to classify and order the set of features, which runs to different dynamic behaviors during training and classification. ML has two categories, namely Unsupervised and Supervised Learning**.** The supervised traffic classification is classified into two different types: parametric classifiers, such as C4.5 decision tree [2], SVM [3], Naïve Bayes, Bayesian network [4], Naïve bayes Tree [5] [19] and non-parametric classifiers such as Nearest Neighbor (k-NN) [6]. The Unsupervised clustering techniques include basic K-Means, DBSCAN, and EM. By using supervised classification algorithms or unsupervised cluster classification algorithms, the flow based statistical feature traffic classification can be done.

In this paper 5 ML algorithms are compared with full featured and reduced feature dataset. We also provide some vision into the qualities of ML traffic classification by describing 43 practical flow features. These 43 practical flow features are used within IP traffic classification, and further reduce the number of features using Correlation-based feature reduction algorithms. We strengthen that a similar level of classification accuracy can be achieved when using several different algorithms with the similar set of features and training/testing data. On the basis of computational performance the algorithms can be differentiated.

1. Feature reduction improves the computational performance by reducing the number of features needed to identify traffic flows.
2. Feature reduction does not reduce the classification accuracy as much as it improves the performance.

3. We find that different ML algorithms (Bayes Net, Naïve Bayes Tree, Nearest Neighbor and C4.5, RBF) present very similar classification accuracy. [2]

II. RELATED WORK

2.1 Supervised Learning

The supervised traffic classification is classified into two different types: parametric classifiers, such as C4.5 decision tree [2], SVM [3], Naïve Bayes, Bayesian network [4] [19], Naïve bayes Tree [5] and non-parametric classifiers such as Nearest Neighbor (k-NN) [6]. The supervised traffic classification procedures analyze the supervised training data and yield an inferred function which can calculate the output class for any testing flow. In supervised traffic classification, appropriate supervised training data is a general assumption. To address the difficulties suffered by payload-based traffic classification, such as user data privacy and encrypted applications, Moore and Zuev [7] realistic the supervised naive Bayes techniques to classify network traffic created on flow statistical features. Bernaille and Teixeira [8] intended to use only the packet size of an SSL connection to identify the encrypted applications. Este et al. [9] applied one class SVMs to traffic classification and turned-out an easy optimization algorithm for every set of SVM working parameters. These works use parametric Machine Learning algorithms, which require a severe training procedure for the classifier parameters and need the re-training for new discovered applications. There are a few works using non-parametric machine learning algorithms. Nguyen and Armitage [10] intended to conduct traffic classification based on the recent packets of a flow for real-time purpose. Auld et al. [4] protracted the work of [7] with the application of Bayesian neural networks for exact traffic classification. Roughan et al. [6] have verified NN and LDA methods for traffic classification using five categories of statistical features. Kim et al. [3] broadly compared host behavior based BLINC method, ports-based CorelReef method and seven mutual statistical feature based methods using supervised algorithms on seven distinct traffic traces.

2.2 Unsupervised Learning

The Unsupervised clustering techniques are basic K-Means, DBSCAN, AutoClass and EM. The unsupervised methods (or clustering) finds cluster structure in unlabeled traffic data and allocate any testing flow to the application-based class of its nearest cluster. Erman et al. [11] compared the k-means, DBSCAN and AutoClass algorithms for traffic clustering on two experimental data traces. The experimental research displayed that traffic clustering can produce high-purity clusters when the number of clusters is set as much superior as the number of real applications. In general, the clustering techniques can be used to find out traffic from before unknown applications [12]. Zander et al. [13] exhausted AutoClass to group traffic flows and intended a metric called intra-class homogeneity for cluster evaluation. Erman et al. [14] intended to help a set of supervised training data in an unsupervised approach to address the difficult of mapping since flow clusters to real applications. However, the mapping method will yield a great proportion of 'unknown' clusters, particularly when the supervised training data is very small. Wang et al. [15] intended to incorporate statistical feature founded flow clustering with payload signature matching method, so as to eliminate the constraint of supervised training data. Finamore et al. [16] shared flow statistical feature based clustering and payload statistical feature based clustering for mining unidentified traffic. But, the clustering methods suffer from a problem of mapping from a large number of clusters to real applications. This problem is very complicated to solve without knowing any information about real time applications.

III. SYSTEM MODEL

Several research papers in different mechanisms have used ML classifier [17]. Comparison is made between five ML algorithms (Nearest Neighbor, C4.5, Bayes Net, Naive Bayes, RBF). The development of real time internet traffic dataset is used to classify seven applications, which are WWW, E-MAIL, CHAT, P2P, FTP, IM (Instant Message), VoIP. This work uses Tcpdump tool for capturing the packet from the network. The Netmate tool is used for calculating flow statistical features. The first process is to capture the IP packets crossing a computer network and constructs traffic flows by IP header inspection. A flow consists of Src_ip, Dest_ip, Protocol, Dest_port, Sorc_port. Features Extraction/Selection is used to extract the statistical features to represent each flow. Classification with WEKA is done with various ML algorithms.

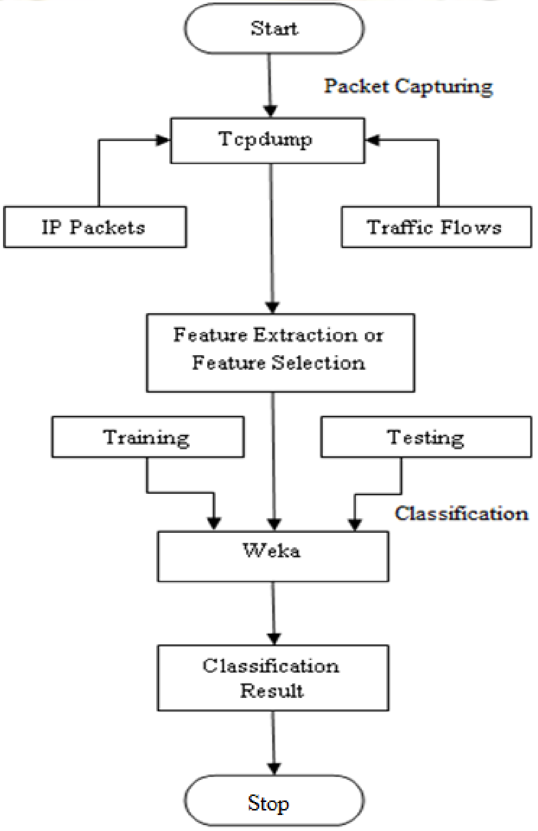


Fig 1: Experimental Flow

Network traffic is represented using flow-based features. In this case, each network flow is described by a set of statistical features. Here, a feature is a descriptive statistic that can be calculated from one or more packets. To this end, NetMate [18] is employed to process data sets, generate flows and compute feature values. Flows are bidirectional and the first packet seen by the tool determines the forward direction. Additionally, flows are of limited duration. UDP flows are terminated by a flow timeout. TCP flows are terminated upon proper connection teardown or by a flow timeout, whichever occurs first. The TCP flow time out value employed in this work is 600 seconds. The flows as defined by the features, we extract a similar set of features as shown in Table 1.

Table 1: Features used by the selected previous ML works for traffic classification

Ref	Features employed for classification
[12]	Number of packets, packet length and inter arrival statistics, flow duration.
[6]	Flow duration statistics, data volume, number of packets per flow
[5,19,20]	249 features including port, flow duration and inter-arrival statistics, packet length
[2,21]	Protocol, duration, volume in bytes and packets, packet length and inter arrival statistics.
[22]	Percentage of IP packets with certain sizes and percentage of flows with certain packet sizes and duration, protocol layer ports
[6][23]	Packet length statistics, window size, byte ratio of received packets, number of packets per flow, transport protocol

IV. INTERNET TRAFFIC DATASET

In this work, a packet capturing tool, Tcpdump is used to capture the real time internet traffic from Karunya University net. Tcpdump tool is a network packet analyzer which captures the network packets. In this process of developing datasets, two datasets are obtained 1. Full Feature dataset, 2. Reduced dataset. In both the datasets seven internet applications are taken into account such as WWW,E-MAIL,FTP,P2P,IM,VOIP,CHAT. These datasets include 3500 samples.

4.1 Feature Selection

Our complete full feature data set contains 44 different features as found by using Netmate tool, each of which has varied distribution in the datasets and different collection/computation cost has been associated with it. The features such as, minimum, maximum, mean, standard deviation number of packets sent in forward and backward direction, average packets, packet size, duration etc. In reduced feature, dataset is obtained from full feature dataset using cfsSubsetEval evaluator and Best First search in the attribute selection filter of Weka tool. For our work, we have used 2.27 GHz core i3 CPU workstation with 4 GB of RAM and Ubuntu 12.04 operating system.

V. IMPLEMENTATION AND RESULT ANALYSIS

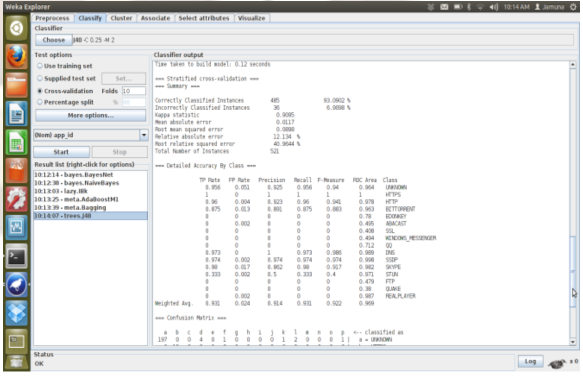
A. *Methodology*

In this work, Weka tool [24], which is a well known data mining tool, is used for implementing IP traffic classification with five different ML algorithms. The Classification accuracy, Training time, recall and precision values [25] of individual internet application samples are employed in order to evaluate performance of all these 5 machine learning or classifiers. All these parameters are defined as follows:

- Accuracy: Bytes of packets are carried by the correctly classified flows
- Training Time: The total time occupied for training of a machine learning classifier
- True positive (TP): The Recall or True Positive rate (TP) is the proportion of positive cases that were correctly identified.
- True Negative (TN): The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly
- False Positive (FP): The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive
- False Negative (FN): The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative.
- Recall =TP/(TP+FN)*100 %
- Precision (P) is the proportion of the predicted positive cases that were correct, Precision = TP/(TP+FP)*100%

B. *Results and Analysis*

The Result shows that, in case of full feature dataset, Bayes Net Classifier provides the better accuracy which is 82.33 %, C4.5 provides the higher accuracy which is 93.33%. The algorithms such as Bayesian Network, Naïve Bayes, Nearest Neighbor, C4.5, RBF Decision Tree are shown in Fig1. Among these algorithm C4.5 decision tree gives high accuracy (93.55%) for traffic classification. The Training Time of this algorithm is 0.05 seconds when compared with other algorithms.



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) app_id

10:12:14 - bayes.BayesNet
10:12:38 - bayes.NaiveBayes
10:13:03 - lazy.IBk
10:13:25 - meta.AdaBoostM1
10:13:39 - meta.Bagging
10:14:07 - trees.J48

```
Time taken to build model: 0.12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         485               93.0902 %
Incorrectly Classified Instances        36                6.9098 %
Kappa statistic                          0.9095
Mean absolute error                      0.0117
Root mean squared error                  0.0898
Relative absolute error                 12.134  %
Root relative squared error             40.9644 %
Total Number of Instances              521

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.956     0.051      0.925     0.956     0.94       0.964    UNKNOWN
                 1         0          1         1         1          1        HTTPS
                 0.96      0.004      0.923     0.96      0.941      0.978    HTTP
                 0.875     0.013      0.891     0.875     0.883      0.963    BITTORRENT
                 0         0          0         0         0          0.78     EDONKEY
                 0         0.002      0         0         0          0.495    ABACAST
                 0         0          0         0         0          0.408    SSL
                 0         0          0         0         0          0.494    WINDOWS_MESSENGER
                 0         0          0         0         0          0.712    QQ
                 0.973     0          1         0.973     0.986      0.989    DNS
                 0.974     0.002      0.974     0.974     0.974      0.998    SSDP
                 0.98      0.017      0.862     0.98      0.917      0.982    SKYPE
                 0.333     0.002      0.5       0.333     0.4        0.971    STUN
                 0         0          0         0         0          0.479    FTP
                 0         0          0         0         0          0.38     QUAKE
                 0         0.002      0         0         0          0.987    REALPLAYER
Weighted Avg.    0.931     0.024      0.914     0.931     0.922      0.969

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   <-- classified as
 197   0   0   4   0   1   0   0   0   0   1   2   0   0   0   1 |   a = UNKNOWN
```

Status
OK

Fig2: Traffic Classification using C4.5 Decision Tree

The Fig2 shows the analysis of Metrics True Positive, True Negative, False Positive, Precision and Recall.

Table 2: Classification Accuracy and Training Time - Full Feature dataset

ML Classifiers	NN	C4.5	RBF	Bayes Net	Naïve Bayes
Classification	76.23	80.23	71.12	82.33	68.31

Accuracy %					
Training Time (sec)	15	7	125	20	17

Table 3: Classification Accuracy and Training Time - Reduced Feature set

ML Classifiers	NN	C4.5	RBF	Bayes Net	Naïve Bayes
Classification Accuracy %	80.2	93.33	68.25	78.32	70.13
Training Time (sec)	5	1	53	15	7

Table 2 and 3 shows the result analysis of Accuracy of traffic classification using Nearest Neighbor,C4.5,RBF,Bayes Net, Naïve Bayes algorithm.

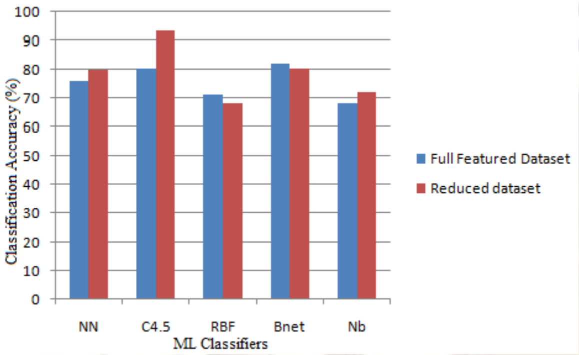


Fig3: Classification accuracy of full featured dataset and Reduce Feature dataset

Fig3 shows comparison of full feature dataset and reduced feature dataset. From these results, for the full featured dataset the Nearest Neighbor algorithm gives better performance than other RBF and Naïve Bays.

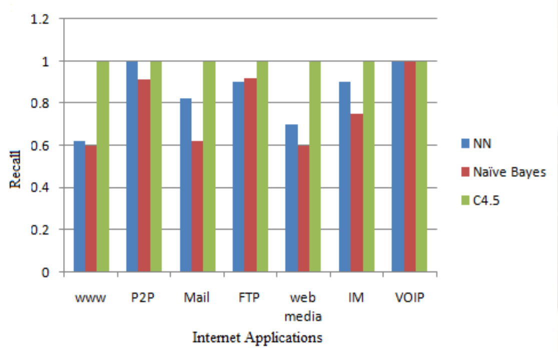


Fig4: Recall – Full Feature Dataset

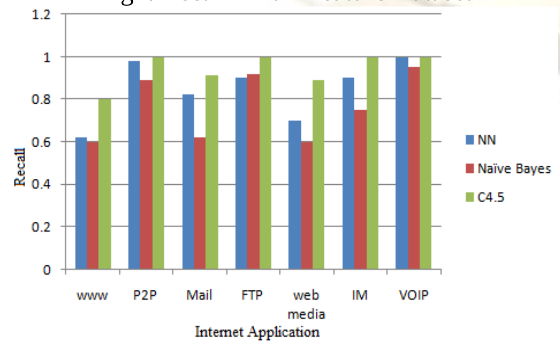


Fig5: Recall – Reduced Feature Dataset

Fig 4 and 5 shows the result for recall for both full feature dataset and reduced feature dataset. In both the cases C4.5 decision tree have high value compare to other two ML algorithms.

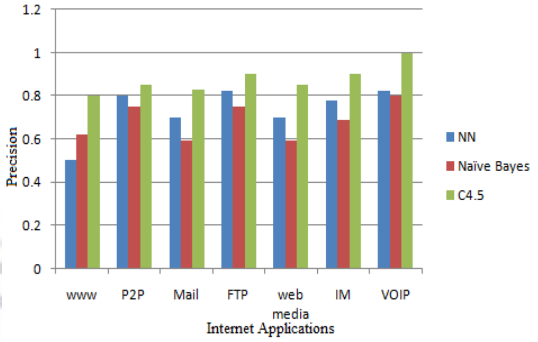


Fig 6: Precision – Full Feature Dataset

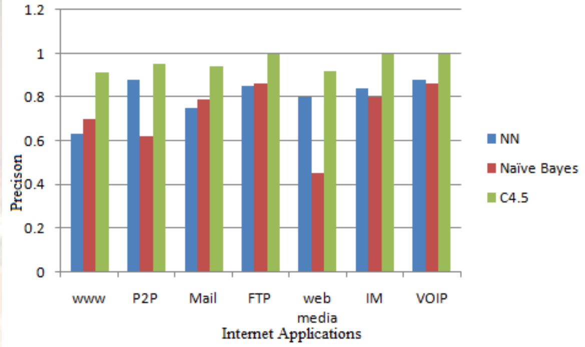


Fig 7: Precision – Reduced Feature set

The fig 6 and 7 shows the result Precision values for full featured datasets. In both the results the C4.5 decision tree algorithms have good accuracy.

VI. CONCLUSION

In this paper we identify that real-time traffic classifiers will work under constraints, which limit the number and type of features that can be calculated. On this source we define 43 flow features that are simple to compute and are well implicit within the networking community. We estimate the classification accuracy and computational performance of C4.5, Nearest Neighbor,Bayes Network, and Naïve Bayes algorithms using the 43 features and with one reduced feature sets and feature selection set. We find that better difference of algorithms can be found by examining computational performance metrics such as build time and classification speed. In comparing the classification speed, we find that C4.5 is intelligent to classify network flows faster than the remaining algorithms. We found RBF to have the slowest classification speed compared with Bayes Net, NB, NN and C4.5.Build time found RBF to be slowest by a considerable margin. Early, ML techniques relied on static offline analysis of previously captured traffic. More recent work begins to address the requirements for practical ML-based real-time IP traffic classification in operational networks. It shows the various Data Traces and Features. The Table 1 and 2 shows how ML techniques out performs the other algorithms with the accuracy.

REFERENCES

[1] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: multilevel traffic classification in the dark,” *SIGCOMM Comput. Commun. Rev.*, vol. 35, pp. 229–240, August 2005.

[2] N. Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification,” *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 5–16, October 2006.

[3] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, “Internet traffic classification demystified: myths, caveats, and the best practices,” in *Proceedings of the ACM CoNEXT Con- ference*, New York, NY, USA, 2008, pp. 1–12.

[4] R. Kohavi, “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid”, in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 1996.

[5] T. Auld, A. W. Moore, and S. F. Gull, “Bayesian neural networks for internet traffic classification,” *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 223–239, January 2007.

[6] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, “Class-ofservice mapping for QoS: a statistical signature-based approach to IP traffic classification,” in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, 2004, pp. 135–148.

[7] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” *SIGMETRICS Perform. Eval. Rev.*, vol. 33, pp. 50–60, June 2005.

[8] L. Bernaille and R. Teixeira, “Early recognition of encrypted applications,” in *Proceedings of the 8th international conference on Passive and active network measurement*, Berlin, Heidelberg, 2007, pp. 165–175.

[9] A. Este, F. Gringoli, and L. Salgarelli, “Support vector machines for tcp traffic classification,” *Computer Networks*, vol. 53, no. 14, pp. 2476–2490, September 2009.

[10] T. Nguyen and G. Armitage, “Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks,” in *Local Computer Networks, Annual IEEE Conference on*, Los Alamitos, CA, USA, 2006, pp. 369–376.

[11] J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,” in *Proceedings of the SIGCOMM workshop on Mining network data*, New York, NY, USA, 2006, pp. 281–286.

[12] J. Erman, A. Mahanti, and M. Arlitt, “Internet traffic identification using machine learning,” in *IEEE Global Telecommunications Conference*, San Francisco, CA, 2006, pp. 1–6.

[13] S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classification and application identification using machine learning,” in *Annual IEEE Conference on Local Computer Networks*, Los Alamitos, CA, USA, 2005, pp. 250–257.

[14] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, “Offline/ realtime traffic classification using semi-supervised learning,”*Performance Evaluation*, vol. 64, no. 9-12, pp. 1194–1213, October 2007.

[15] Y. Wang, Y. Xiang, and S.-Z. Yu, “An automatic application signature construction system for unknown traffic,” *Concurrency Computat.: Pract. Exper.*, vol. 22, pp. 1927–1944, 2010.

[16] A. Finamore, M. Mellia, and M. Meo,”Mining unclassified traffic using automatic clustering techniques,” in *TMA InternationalWorkshop on Traffic Monitoring and Analysis*, Vienna, AU, April 2011, pp. 150–163.

[17] Singh .K and S. Agarwal, Comparative Analysis of Five Machine Learning Algorithms for IP traffic classification 2011

[18] http://sourceforge.net/projects/netmate-meter/ Netmate(viewed August 2006)

[19] A.W.Moore, D. Zuev, Internet Traffic Classification using Bayesian analysis techniques, in:ACM SIGMETRICS 05,2005

[20] A.W. Moore,, D. Zuev, Traffic Classification suing a statistical approach, in Passive and Active Measurement Workshop, PAM05,2005

[21] N.Williams, S. Zander, G. Armitage, Evaluating machine learning algorithms for automated network application identification in Center for Advanced Internet Architectures, CAIA, Technical Report 060410B,2006

[22] E.G. Schmidt, M. Soysal, An intrusion detection based approach for the scalable detection of P2P traffic in the national academic network backbone, in : International symposium on Computer Networks, ISCN06,2006,pp.128-133

[23] R. Yuan, Z. Li, X. Guan, An SVM- based machine learning method for accurate Internet traffic classification 2008, Information Systems Frontiers.

[24] Waikato Environment for Knowledge Analysis (WEKA) 3.4.4, http://www.cs.waikato.ac.nz/ml/weka/ (viewed August 2006).

[25] Nguyen T. T. and Armitage G., (2008.)“A survey of techniques for internet traffic classification using machine learning,” *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, Fourth Quarter.