A Proposal for IP Traffic Classifier for Educational Institutions

Jaspreet Kaur
University Institute of Engineering & Technology
Panjab University
Chandigarh, India

S. Agrawal
University Institute of Engineering & Technology
Panjab University
Chandigarh, India

ABSTRACT

Now a days internet traffic classification is an emerging research field since 1990's because of its use in a large number of network activities. Traditional techniques of internet traffic classification that relied on well known TCP/UDP port numbers or payload based are rarely used because of use of dynamic port numbers instead of fixed port numbers and due to various cryptographic techniques which inhibit inspection of packet payload. Recent trends are use of ML (machine learning) algorithms for internet traffic classification. In our research work we propose a technique to classify the internet traffic into two classes, one for educational websites and another for non-educational websites. In educational institutes for the optimum use of network resources and for the welfare of the students, the use of non-educational websites should be banned while only the educational websites should be allowed to open. To classify the internet traffic we propose a technique to capture data packets first, related with various educational and non-educational websites, using a packet capturing tool Wireshark. Then using feature selection algorithm, a reduced feature dataset will be developed. After that training and testing of various ML algorithms will have to be performed. Finally comparative analysis of the different classifiers from the obtained results is to be performed.

General Terms

IP Traffic Classification, Educational websites, Non-educational websites.

Keywords

Machine Learning, Features.

1. INTRODUCTION

In recent years with rapid growth in internet users both for educational and non-educational purposes, the internet traffic is going to be increased at drastic rate. Due to the use of a number of internet applications by users in different fields, the internet traffic increases day by day. Internet traffic classification is of great importance in various network fields like security, monitoring to accounting, to detect network intrusions, to detect network misuse by internal and external users and many more.

In our research work internet traffic is to be classified into two classes, one for educational websites and another for non-educational websites. There are infinite websites in educational institutes like educational and non-educational, that one can access, but for the optimum use of network resources in the educational institutes , the use of non-educational websites should be banned while only the educational websites should be allowed to open.

Educational websites e.g. www.math.com is used for solving mathematical problems, www.novelguide.com is used for literary analysis, www.sparknotes.com is used for study guides for literature, poetry, history, film and philosophy etc. Non educational websites like www.facebook.com, www.yahoomassenger.com are used for chatting purposes and different websites used for songs, movies and games download also come under the category of non-educational institutes.

The major contribution in internet traffic is done by peer to peer (P2P) applications such as Bit Torrent, Emule, Kaaza etc. leads 80% rise in internet traffic [1].

Internet Traffic classification can be either offline or online. In online classification, analysis is performed while data packets flowing through the network are captured; but in case of offline classification technique, firstly data traces are captured and stored and then analyzed later [1].

Traditional IP traffic classification techniques are direct packet inspection based techniques such as port number based and payload based techniques [2], [3]. But presently these techniques are rarely used. Traditional techniques are payload based and port number based packet inspection techniques. In payload based technique , payload of few TCP/IP packets are analyzed in order to identify any particular application which is not possible today because of use of cryptographic techniques used to encrypt data in packet payload and privacy policies of governments which do not allow any unaffiliated third party to inspect each packets payload.

In port number based packet inspection technique, well-known port numbers are provided in header of IP packets which are reserved by IANA (Internet Assigned Numbers Authority) for particular applications e.g. port number 80 is reserved for web based applications [4]. Unfortunately, this method is also rarely used due to the use of Dynamic port numbers instead of Well-known port numbers for various applications.

Now number of researchers are looking for internet traffic classification techniques without deep inspection of packets. Most of these techniques come under the category of Machine Learning (ML) techniques. In ML techniques, first, features are defined to identify and differentiate future unknown internet traffic data. These features are attributes of flows calculated over multiple packets (such as maximum or minimum packet lengths in each direction, flow durations or inter-packet arrival times, data rate of traffic, traffic volume etc) [5].

Most of the students waste their precious time on surfing the non-educational websites therefore for the optimum use of network resources in the educational institutes these non-educational websites should be banned. For banning these websites internet traffic identification is required. In this paper we will be proposing a technique to identify the educational and non-educational websites and to prepare a dataset related to them.

The remaining paper is organised as follows: section II gives some information about related work done by various researchers in the field of IP traffic classification. Section III includes proposed approach. Section IV gives conclusion.

2. RELATED WORK

Various researchers have shown their interest in internet traffic classification over last few years . some previous work done in this field by some researchers is discussed as follows.

In [6], Hyunchul Kim et al. have discussed CoralReef approach which is a port number based internet traffic classification technique. Results show that the overall accuracy of CoralReef on the traces ranges from 71.4% to 95.9%. Finally, authors conclude that Port-based approach still accurately identifies most legacy applications; its weakness is in identifying applications that use dynamic ports or traffic masquerading behind a port typically used for another application. Thus, ports still possess significant discriminative power in classifying certain types of traffic.

In [7], Madhukar and Willionson have conducted a longitudinal study over a 2-year period on the effectiveness of port-based classification using empirical Internet traces taken from the University of Calgary. The authors compared port-based classification with a classification technique that relies on a set of transport layer heuristics. Their trace only had SYN, FIN, and RST packets due to the longitudinal nature of their trace, and thus, validation of their classification results using payload-based technique was not feasible. They found that 30% to 70% of the traffic is classified as unknown with port-based analysis. In addition, they found that the amount of unknown traffic was typically from 10% to 30% in the September 2003 to April 2004 portion of their trace. It has since increased from 30% to 70% by the Spring of 2005. They provide strong circumstantial evidence that this increase in unknown traffic is highly correlated to the increase in P2P traffic found with their transport-layer heuristic.

In [8], Moore and Papagiannaki have discussed content based or payload based and port number based internet traffic classification technique. They describe a content-based methodology to classify network traffic. The first step of their classification methodology uses IANA assigned port numbers to create an initial classification and after that, using an iterative procedure, they use increasingly more information at later steps. This approach allows the traffic to be classified with increased confidence. The last step concludes the process by relying on manual analysis of the traffic for any remaining unclassified traffic. They have compared the effectiveness of port-based classification to this content based approach. To facilitate this comparison the authors collected a 24-hour trace of the traffic generated from approximately 1,000 users. This comparison found that approximately 30% of the bytes in the traffic are either misclassified or unclassified when using just the IANA port assignments. However, with the content-based approach 99.9% of the traffic were identified confidently.

In [9], C. Dews et al. have given a look at the network traffic dynamics of Internet Chat Systems. The authors focus on IRC and web-based chat systems. Their paper describes a port and payload-based methodology for identifying the chat flows and filtering out non-chat traffic. Their approach uses well-known port numbers to filter out traffic that is most likely non-chat such as Gnutella traffic on port 6346. After this filtering has taken place they use payload signatures to separate the web-based chat flows from the regular non-chat traffic.

In [10, 11], Zander et al. have discussed their work by looking at maximizing intra-cluster homogeneity (or cluster purity) by investigating which set of features separate the flows from different applications with greatest accuracy. The traces used in this analysis are from a publicly available archive of traces and port-based analysis was used to establish the “base truth". The authors have continued this work and recently used the C4.5 supervised machine learning algorithm to estimate the traffic trends in archival traces.

In [12], Amina El Gonnouni et al. have discussed a non linear system identification problem. A Support Vector Regressor has been used to solve the Internet traffic identification problem. They have given a basic idea underlying Support Vector (SV) machine for regression, which is a novel type of learning machine based on statistical learning theory. Furthermore, they described how SV regressor can be applied for non linear system identification. In their simulations results they have presented two type of kernel functions, the Radial Basis Function (RBF), and the hyperbolic tangent, which are compared with the classical two-layer MLP (Multi-Layer Perceptron) Neural Networks, trained to minimize a quadratic error objective with the Back-Propagation (BP) algorithm. The SV regressor outperforms the MLP and demonstrates its effectiveness for solving non linear system identification problems. This paper concludes that SV regression is attractive for non linear system identification because high dimensional expansions are not explicitly required to solve for linear parameters. From other hand, SVM have the advantage that the number of kernels is found automatically and optimally. Other parameters associated with kernels can be optimized relatively easily because data do not have to be tested to judge performance. But dataset used in [12] is limited to video packet transmission only.

In [13], Soysal and Schmidt have presented a systematic approach for investigating and evaluating the internet traffic classification performance of three supervised Machine Learning (ML) algorithms namely Bayesian Networks (BNs), Decision Trees (DTs) and Multilayer Perceptrons (MLPs), using flow traces. The performance results indicate that DTs have both a higher accuracy and a higher classification rate than BNs. However, DTs require a larger build time and are more susceptible in the case of incorrect or small amounts of training data. A detailed analysis of traffic classification with MLPs that are trained by back propagation is carried out to identify the drawbacks of this algorithm. As a result, it is not possible to simultaneously achieve acceptable recall values for these traffic types when the MLP algorithm is used.

In [14], Arya and Mishra have proposed multilevel classifiers based on the performance of multiple classifiers for internet traffic classification. Five classifiers namely J48, Random Tree, Random Forest, Bagging and boosting algorithms are evaluated over single benchmark dataset. Proposed multilevel classifiers give better performance than single classifier. Performance of classifier for P2P class increases by using classifier combinations using Bagging and Multiboosting. Multiboosting outperforms the Bagging approach.

In [15] Shijun huang et. Al have demonstrated the statistical features based approach to classify internet traffic using supervised ML. The simplified statistical features and the easy-to-use k-Nearest Neighbor (KNN) estimator result in lower space and time complexity, which is worth mentioning. They carried out several data sets including 9 flows of MAIL, 100 flows of WWW, 34 flows of BULK, 100 flows of IM, 100 flows of P2P and 5 flows of STREAM (full-flow) are collected in the way mentioned in section III, all of which are used to train the k-Nearest Neighbor estimator. They inferred that the classifier model works perfectly when classifying only MAIL, WWW and BULK flows. But with IM flows added, classification results of MAIL flows drop greatly, which breaks the principle of fairness in KNN algorithm. More problems are discovered when P2P flows are added.

In [16] Youngli Ma et.al integrated theory with actual needs on the measure works, and made use of the characteristics of the network traffic that were understood easily in internet network. First, they used the CFS and genetic search method to select three subsets from three full sets. Then they primarily selected 15 kinds of algorithms from more than 50 ones which involved in decision tree, rules, Bayes, neural network algorithms, finally proposed the multivariate evaluation method (MEM) to assess these algorithms on accuracy, memory consumption, CPU utilization, construction model time and test time.

In [17] Singh and Agrawal captured firstly real time internet traffic using Wire shark software which is a packet capturing tool. After that, Internet traffic is classified using five ML classifiers. Results show that Bays' Net gives better classification of internet traffic data in terms of classification accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications. After that, the no. of features used to characterize each internet application data sample of this dataset are further reduced to make a reduced feature dataset. Their results show that with reduced feature dataset, performance of these classifiers is improved to large extent. In this case, C4.5 classifier gives very much accurate results. Thus it is evident that Bays' Net and C4.5 are effective ML techniques for IP traffic classification with accuracy in the range of 94 %.

3. PROPOSED APPROACH

In our research work we will be exploring the educational and non-educational websites. We will classify them into two classes. So we will adopt a general research methodology for internet traffic classification which is shown by a flow chart in Fig.1

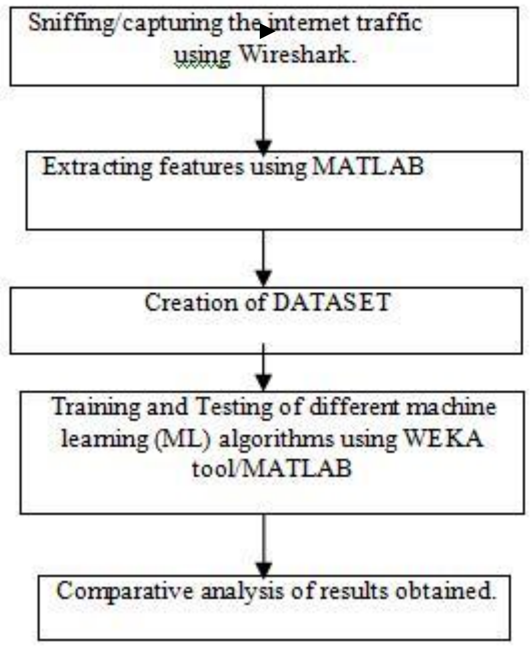
Sniffing/capturing the internet traffic using Wireshark.
Extracting features using MATLAB
Creation of DATASET
Training and Testing of different machine learning (ML) algorithms using WEKA tool/MATLAB
Comparative analysis of results obtained.

Fig.1 General research methodology-A Flowchart

- Capturing Real Time Internet Traffic: In this research work, two real time internet traffic datasets have to be developed for packet capture duration of 1 minute and 1 second respectively using packet capturing software, Wireshark.

Wireshark, [18] which is a well-known packet sniffer, is used to capture internet traffic. From this captured traffic, a number of features can be extracted and datasets can be developed. Wireshark is an open source network packet analyzer which can capture network packets and tries to display that packet data as detailed as possible. It has many remarkable features like available for UNIX and Windows, capture live packet data from a network interface, display packets with very detailed protocol information, open and Save packet data captured, import and export packet data from and to a lot of other capture programs, etc.

- Feature Extraction: After capturing internet traffic traces for different packet capture time durations, various flow based statistical features will be extracted using MATLAB [19]. The number of features and their type extracted from the dataset will be decided based on the final understanding.
- Dataset Creation: After extracting features for each website, two datasets will be developed: one for packet capture duration of 1 minutes and other for packet capture duration of 1 second. These datasets will be over sampled using Weka tool [20] to increase number of samples in each dataset.
- Training and testing of ML Algorithms: We will employ different ML algorithms for internet traffic classification using various internet traffic datasets. For this purpose, MATLAB and Weka tool will be employed.

- Comparative Analysis of Results: In our research work, results of internet traffic classification from various classifiers then will be analyzed on the basis of different parameters.

4. CONCLUSION

In this paper we conclude that till now no research work on the classification of internet traffic of educational and non-educational websites has been done , so we propose a technique for this type of classification. Most of the research work has been done in the field of internet traffic classification using ML techniques. In our research work we propose a technique that capture data packets first, related with various educational and non-educational websites, using a packet sniffing tool Wireshark. Then using feature selection algorithm, a reduced feature dataset will be developed for both educational and non-educational websites. After that training and testing of various ML algorithms will have to be performed. Finally comparative analysis of the different classifiers from the obtained results will be performed.

5. REFRENCES

[1] Arthur Callado, Carlos Kamienski, Géza Szabó, Balázs Péter Ger″o, Judith Kelner,Stênio Fernandes ,and Djamel Sadok, “A Survey on Internet Traffic Identification,” IEEE Communications Survey & tutorials, vol. 11, no. 3, pp. 37-52, Third Quarter 2009.

[2] Thuy T.T. Nguyen and Grenville Armitage, “A Survey of Techniques for Internet Traffic Classification using Machine Learning,” IEEE Communications Survey & tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008.

[3] Runyuan Sun, Bo Yang, Lizhi Peng, Zhenxiang Chen, Lei Zhang, and Shan Jing, “Traffic Classification Using Probabilistic Neural Network,” in Sixth International Conference on Natural Computation (ICNC 2010), 2010, pp. 1914-1919.

[4] http:/www.iana.org/assignments/port numbers.

[5] Andrew W. Moore, Denis Zuev, Michael L. Crogan, “Discriminators for use in flow-based classification,” Queen Mary University of London, Department of Computer Science, RR-05-13, August 2005.

[6] Hyunchul Kim, kc claffy, Marina Fomenkov, Dhiman Barman, Michalis Faloutsos, and KiYoung Lee, “Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices,” in ACM CoNEXT 2008, December 10-12, 2008, Madrid, SPAIN.

[7] A. Madhukar and C. Williamson. “A Longitudinal Study of P2P Traffic Classification.” In MASCOTS'06, Monterey, USA, August 2006.

[8] A.W.Moore and D.papagiannaki, “Toward the accurate Identification of network applications”, in poc. 6th passive active measurement. Workshop (PAM), mar 2005, Vol.3431, pp 41-54.

[9] C. Dews, A. Wichmann, and A. Feldmann. “An Analysis of Internet Chat Systems” In IMC'03, Miami Beach, USA, October 2003.

[10] S. Zander, T. Nguyen, and G. Armitage. “Automated Traffic Classification and Application Identification using Machine Learning”. In LCN'05, Sydney, Australia, November 2005.

[11] S. Zander, T. Nguyen, and G. Armitage. “Self-Learning IP Traffic Classification Based on Statistical Flow Characteristics”. In PAM'05, Boston, USA, March 2005.

[12] Amina Lyhyaoui, “Support Vector Machine for Internet Traffic Identification,” in IEEE, 2007,pp. 351-354.

[13] Murat Soysal, and Ece Guran Schmidt, “Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison,” Performance Evaluation Elsevier Journal, Vol. 67, 2010, pp. 451-467.

[14] Indra Bhan Arya, and Rachna Mishra, “Internet Traffic Classification: An Enhancement in Performance using Classifiers Combination,” International Journal of Computer Science and Information Technologies, Vol. 2 (2), 2011, pp. 663-667.

[15] Shijun Huang Kai Chen Chao Liu, Alei Liang, Haibing Guan, “A Statistical-Feature-Based Approach to Internet Traffic Classification Using Machine Learning” 9781-4244-3941-6/09/$25.00 ©2009 IEEE

[16] Yongli Ma, Zongjue Qian, Guochu Shou, Yihong Hu“*Study on Preliminary Performance of Algorithms for Network Traffic Identification”* 978-0-7695-3336-0/08 $25.00 © 2008 IEEE DOI 10.1109/CSSE.2008.1277

[17] Kuldeep Singh and Sunil Agrawal, Comparative Analysis of five Machine Learning Algorithms for IP Traffic Classification, *Internation Conference on Emerging Trends in Networks and Computrt Communications (ENCTT-2011)*, Udaipur, Rajasthan, India, April 22-24, 2011.

[18] Wireshark,Available:http://www.wireshark.org/

[19] MATLAB,Available:www.mathworks.com