ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 3, Issue 5, May 2018

Real-time Network Traffic Classification for Educational Academy Using Machine Learning

[1] Hari Krishna Mishra, [2] Manas Pratim Dutta, [3] Sunit Kumar Nandi
[1][2][3] Department of Computer Science & Engineering National Institute of Technology Yupia, Arunachal Pradesh – 791112 India

***Abstract*— Over the last few years, there has been a rapid increase in the variety and volume of Internet traffic. The regular method of classifying this traffic relies on known IANA assigned port numbers or inspecting the payload. However, these methods are not effective as the used port numbers can differ from well-known or official ones. Payload inspection is not effective either because applications encrypt their data before sending. A new machine learning technique for network traffic classification is proposed by us to overcome this drawbacks. Services on the Internet can be grouped into different classes. There are a number of of websites in an educational institutions. They are educational websites and non-educational. Educational services are used for learning and research purposes, whereas non-educational services include entertainment, social networking and communication. Our goal is to classify the traffic effectively and ensure optimal and fair bandwidth allocation among Internet users in the institution giving higher priority to educational services. In this paper, we have used kNN classifiers to classify the accessed services at different values of K [1]. For classification accuracy in terms of dataset, the value of K should be found. For that we have collected many results and datasets. We need good features to get better classification accuracy [2], [3].**

Keywords— Network traffic classification, machine learning, k-NN, wireshark.

I. INTRODUCTION

In this century, the number of users as well as applications on the Internet are increasing. This has lead to drastic increase in IP traffic. The major contributors to Internet traffic is due to various services like YouTube, Hotstar, BitTorrent, etc. The second major contributors to Internet traffic are VoIP applications such as Skype, Google Hangouts, Facebook Messenger, etc. for text messaging and audio-video calls. Port number based and payload inspection based methods are the traditional techniques for traffic classification. Methods which is based on port number do not work properly because non-standard port numbers can be used. Payload based method are becoming ineffective because 99% of the time applications generate encrypted payloads [4]. In our paper, the creation of web services which includes educational and non-educational for network traffic classification is done. Classifications is desperately needed in our work. For that we have used a special method. That excellent property of the method can classify the network traffic in educational and non-educational services such that social media, blogs, video sharing sites, proxy servers, pirate sites, VOIP [5]. To classify the network traffic, we capture data packets of these various web services using a packet capturing tool wireshark. Then, we extract various features from the data and develop the dataset with pyshark. After that, training and testing is done with k-NN classifier with different values of K. Finally, comparison is performed between different results obtained from the various experiments [6].

II. RELATED WORK

A. Port number Based Technique

This is the most common technique for network traffic classification. We know that all packets carry port numbers. Internet Assigned Numbers Authority (IANA) has a duty to assign the port number . Most server applications have known registered port numbers but this is not a necessity. Newer applications like online gaming, peer-to-peer (P2P) file transfer and streaming do not use registered port numbers. As these new applications do not use specified port numbers, it may not be easy to predict the application using the port number. [7].

B. Payload based Technique:

The solution for the drawback of port-number based technique is payload based technique. The signature of the known packets is matched with the payload of packets. This specific technique is known as deep packet inspection. This technique gives the best results (approximately 100%) if encryption is not done in the packets. Henceforth, the accuracy of the technique is very high. Still, there are two vital problems. Firstly, it can not identify encrypted packets. This makes it ineffective widely as 99% of the Internet traffic is encrypted today. Secondly, it takes too much time and processing power to perform the classification [8]–[10].

C. Machine Learning Techniques:

ML algorithms are an interesting way to classify network traffic. There are two types of ML techniques: (1) unsupervised learning (clustering) and (2) supervised learning (classification). [11]

1) Supervised Learning Technique: This technique is based on the attributes of a label or class. The label is chosen on the basis of attributes obtained on the whole data. The dataset is allocated with a set of train and test instances, pre-classified into labels or classes. This approach has 2 steps: training and testing. Training phase examines the data (training dataset). The model that is constructed in the training phase is used to predict new and unseen data. In our work, we have employed supervised ML.

2) Unsupervised Learning Technique: This technique uses the concept of clustering. We generate clusters of same attributes but clustering is not provided with direction. There is no need for the testing phase and training phase.

III. MANUAL FEATURE SELECTION TECHNIQUES

Extracting attributes from a session is the main job of Feature Extraction. SSL/TLS is best way for safe communication in the traffic which is encrypted. Splitting the encrypted data coming from the upper layer is the job of TCP layer. Splitting the data into chunks is only happened if the packets overshoot the Maximum Segment Size(MSS). Every TCP segment is created for each of the chunk. After that the encapsulation is done to each TCP segment which is converted to IP datagram. No session identifier is used for TCP packets. Attributes is the best way to recognize a session. There is a backward and a forward flow in a session. In every complete session, the description of flow is time-ordered sequence in TCP packets. The description of forward flow is depending on the number of bytes received in the incoming packets. On the other end, the definition of backward flow is depending on the bytes sent through the outgoing packets [1], [8], [12]. The features extracted from raw data is as follows:

1) No. of Forward packets

2) Total no. of Forward total Bytes
3) Minimum value of forward inter arrival time difference
4) Maximum value forward inter arrival time difference
5) Mean value forward inter arrival time difference
6) STD forward inter arrival time difference
7) Mean value forward packets
8) STD value of forward packets
9) Number of Backward packets
10) Number of Backward total Bytes
11) Min value backward inter arrival time difference
12) Max value backward inter arrival time difference
13) Mean value backward inter arrival time difference
14) STD value of backward inter arrival time difference
15) Mean value backward packets
16) STD value backward packets
17) Mean value forward TTL value
18) Minimum number of forward packet
19) Minimum number of backward packet
20) Maximum number of forward packet
21) Maximum number of backward packet
22) Total Number of packets
23) Minimum size of packet
24) Maximum size of packet
25) Mean size of packet
26) Packet size variance
27) Average TTL value
28) Window size
29) Window size scaling factor
30) Forward packet data-rate
31) Backward packet data-rate [13]

IV. MACHINE LEARNING APPROACHES FOR INTERNET TRAFFIC CLASSIFICATION

A. K-Nearest Neighbour (K-NN)

K-nearest neighbour (k-NN) is an algorithm used for the classification of the things or objects based on the Euclidean distance between the things or objects. Classification is done using the dataset into multiple dimensions, where every dimension represents an attribute of the dataset. This algorithm stores the values of attributes and class of learning data. In the classification step, the same attributes are evaluated for test data (for which class is not known). The distance of new data for all of its attributes' value is calculated, and the number K which is closest closest to the data provides the outcome. K-NN algorithm accuracy is determined by the absence or presence of attributes [5], [14].

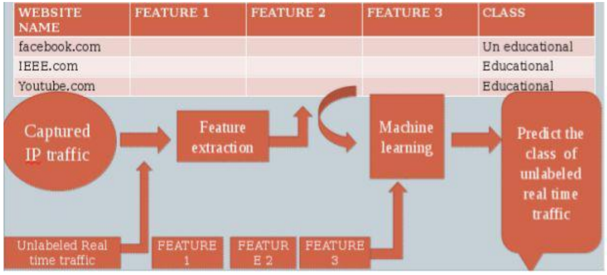
K-NN p using the Euclidean distance formula is given below.

$$D = \sqrt{(x_1 - y_1)^2) + (x_2 - y_2)^2)}$$

$$where x_1, x_2, y_1 and y_2 are data points.$$

V. METHODOLOGY

For. this research work, a research methodology has been designed and shown in Fig. 1



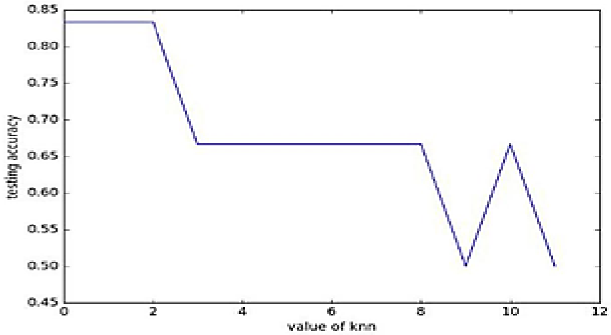
WEBSITE NAME	FEATURE 1	FEATURE 2	FEATURE 3	CLASS
facebook.com				Un educational
IEEE.com				Educational
Youtube.com				Educational

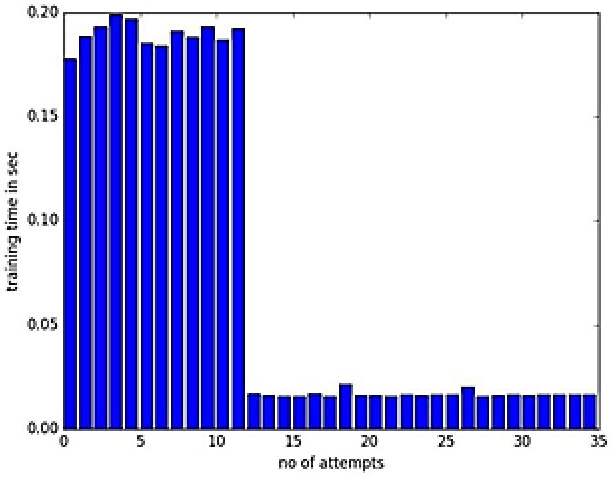
Fig. 1. Flow-chart for network traffic classification

In this paper, Wireshark is used for capturing internet traffic of known and required websites traffic. Then the features of all services traffic are extracted and a dataset is prepared. The dataset is divided into 2 sets: data samples for training and testing purpose in both cases. The Real-Time Internet traffic is filtered website wise and then Train the dataset withML. The feature of raw data is extracted and the dataset is created to predict the class of new traffic data using ML. Classification accuracy is the percentage of correctly classified test data over all samples. Training time is the total time taken for training of ML classifier. Which is measured in seconds [7], [15].

VI. RESULTS AND ANALYSIS

K-NN algorithm is trained and tested for their presentation using test data. Classification accuracy, training time, and values of K-NN classifiers are shown in table 1.The maximum classification accuracy provided by K-NN classifier at certain value of K is 96% . From Table 1, it contains minimum accuracy of only 72. Therefore for all values of k, some of the k values are not suitable for this classification. Thus we again clear that K-NN gives better performance at certain undefined K value but here we checked all possible K value and noticed the better performance in terms accuracy.





VII. CONCLUSION

In our work, we use a tool which is generally used for capturing packet data. The name of the tool is Wireshark. Many educational and non educational data is captured using this tool which is related to network traffic. So, for the network traffic classification we use K-NN classifiers. This classifier gives best result for a particular value of K. A high classification accuracy of 86 is given by K-NN classifier.

REFERENCES

[1] T. Wiradinata and A. S. Paramita, “Clustering and feature selection technique for improving internet traffic classification using k-nn,” 2016.

[2] S. Lee, K. Levanti, and H. S. Kim, “Network monitoring: Present and future,” Computer Networks, vol. 65, pp. 84–98, 2014.

[3] M. Joshi and T. H. Hadi, “A review of network traffic analysis and prediction techniques,” arXiv preprint arXiv:1507.05722, 2015.

[4] K. Singh, S. Agrawal, and B. Sohi, “A near real-time ip traffic classification using machine learning,” International Journal of Intelligent Systems and Applications, vol. 5, no. 3, p. 83, 2013.

[5] J. Kaur, S. Agrawal, and B. Sohi, “Internet traffic classification for educational institutions using machine learning,” International Journal of Intelligent Systems and Applications, vol. 4, no. 8, p. 37, 2012.

[6] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, “A survey of methods for encrypted traffic classification and analysis,” International Journal f Network Management, vol. 25, no. 5, pp. 355–374, 2015.

[7] K. Singh and S. Agrawal, “Comparative analysis of five machine learning algorithms for ip traffic classification,” in Emerging Trends in Networks and Computer Communications (ETNCC), 2011 International Conference on. IEEE, 2011, pp. 33–38.

[8] T. T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56–76, 2008.

[9] T. Bujlow, V. Carela-Español, and P. Barlet-Ros, “Independent comparison of popular dpi tools for traffic classification,” Computer Networks, vol. 76, pp. 75–89, 2015.

[10] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller, and K. Hanssgen, “A survey of payload-based traffic classification approaches,” IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 1135–1156, 2014.

[11] D. M. Divakaran, L. Su, Y. S. Liau, and V. L. Thing, “Slic: Self-learning intelligent classifier for network traffic,” Computer Networks, vol. 91, pp. 283–297, 2015.

[12] J. Muehlstein, Y. Zion, M. Bahumi, I. Kirshenboim, R. Dubin, A. Dvir, and O. Pele, “Analyzing https encrypted traffic to identify user’s operating system, browser and application,” in Consumer Communications & Networking Conference (CCNC), 2017 14th IEEE Annual. IEEE, 2017, pp. 1–6.

[13] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, “Toward an efficient and scalable feature selection approach for internet traffic classification,” Computer Networks, vol. 57, no. 9, pp. 2040–2057, 2013.

[14] S. Katal and A. Singh, “A survey of machine learning algorithm in network traffic classification,” Int. J. Comput. Trends Technol.(IJCTT), vol. 9, no. 6, 2014.

[15] N. Namdev, S. Agrawal, and S. Silkari, “Recent advancement in machine learning based internet traffic classification,” Procedia Computer Science, vol. 60, pp. 784–791, 2015.