IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1178)
Available online at: https://www.ijariit.com

IP traffic classification of 4G network using Machine Learning techniques

S. Mahammad Rafi
rafirinky@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

T. Lavanya
lavanya0948@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

B. Shamitha
baggidishamitha28@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

S. Phaneeswar
phaneeshsriram999@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

N. Chandu
chanduroyal231@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

ABSTRACT

In today's world, the number of online services and users are growing rapidly. This leads to a huge increase internet traffic. Therefore, the task of separating IP traffic is approx. it is important for Internet service providers or ISPs, as well as the variety government and the private sector for the better network management and security. IP traffic separation includes identifying user activity using network traffic flowing into the system. This will also help to improve the network performance. Use of traditional IP traffic Classification methods based on the evaluation of packet capacity and hole numbers dropped significantly because there are so many online apps today use naturally incorrect port numbers than well-known port numbers. Also, there are several encryption strategies today as a result of when testing the package payload is blocked. Currently, various machine readings techniques commonly used to differentiate IP traffic. However,not much research has been done on IP fragmentation 4G network traffic. During this study, we did a new database by downloading real-time Internet traffic packets 4G network data using a tool called Wireshark. After that,we have released the considered features of the packaged packages using the python script. Then we used five typewriters models, namely, Decision Tree, Vector Support Equipment, K Very Near Neighbors, Random Forest, and Naive Bayes IP splitting traffic. It was noted that Random Forest offered the best almost 87% accuracy.

Keywords—*IP Traffic Classification; Port Number; Deep Packet Inspection; Packet Capturing; Feature Extraction; Machine Learning.*

1.INTRODUCTION

Significant increase in the number of internet users around world due to low internet prices and easy access to it cell phones and other devices and services have led I a powerful increase in the amount of existing IP traffic is distributed worldwide. This increase in IP traffic could be resulting from the use of various applications by internet users in their daily lives like email, World Wide Web, text sending text messages, audio or video calls, and various other internet applications. Therefore, splitting IP traffic is very important Internet service providers (ISPs) and various governments and NGOs. Separating IP traffic can help several network management functions such as analysis I Quality of Service (QoS) for online service, diagnostics of any network error, etc. It can also help in several networks security functions such as access detection [1].

2. RELATED WORKS

Many research projects have been undertaken in this field to differentiate IP traffic, taking into account different types of internet applications. Many researchers have come up with various theories segmentation strategies in the sector to differentiate IP traffic. I The following paragraphs describe some of these research activities:

A.Hole Number Based on Arrangement

In this process, for the first time, ports for online applications registered Online Assigned Number Authority or IANA.Then, IP traffic is categorized using the IANA list of registered port numbers [21]. For example, port numbers in some of the online applications registered with IANA given in Table I.

Table1: Iana assigned port numbers for some internet applications

Application	Port Number
FTP	21
Telnet	23
SMTP	25
DNS	53
HTTP	80
IRC	194

As discussed in Section I, this approach does not work because there are so many online apps (e.g.P2P applications) that use the correct port numbers naturally dynamic than known hole numbers.

B. Payment-Based Planning

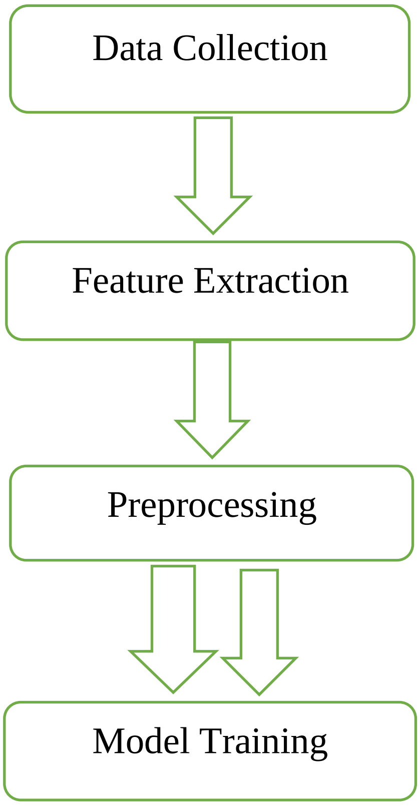
This method is also known as Deep Packet Inspection (DPI) method. This way, the Internet traffic packet the content of the payment is analyzed, as well as the direct signature of known applications are searched.

C. Mechanized Learning Based Planning

As discussed in Section I, this process is based on to train a machine learning model using a variety of calculations Independent package loading features and using this trained model to distinguish IP traffic. Great benefit that this method provides that packet port test the number or volume of the package is not required. So far, several machine learning techniques are often used for profit IP traffic separation function.

3. PROPOSED APPROACH AND EXPERIMENTAL DESIGN

Steps to follow the implementation of our research function is shown in Fig. 1. All these steps and operations the steps used are discussed in the following paragraphs.



Internet Application

A. Internet Traffic Data Collection

For the purpose of our research work, we have created a new one data by capturing 4G real-time internet traffic data the network uses a popular package scanning tool called Wireshark. The source holes and destination must be IPv6 to determine if the received network is 4G.

B. Feature Domain

After the packages are captured in PCAP file format using the Wireshark, the packages are separated by themselves flow respectively. Flow is a series of a one app that can be seen with the packages it has local IP address and the same source and ports as well protocols. Flow is both natural and forward the direction is indicated by the first packet in the flow. This was made in group packs of the same type to flowfeatures can be extracted that can be used for training model.

C. Data Preprocessing

After extracting the flow elements from the PCAP file,static features such as Source and IP address of the destination as well ports were removed before model training. It flows again containing only one package is excluded as similar features arrival time requires at least 2 packs for a certain flow by their calculation. Thus, the features were calculated in 1899it flows. After this, the database is measured using the Standard Scale making the data uniform so that no element can control it other features while separating packets. Train inspection a 3: 1 separation ratio was performed for model training and testing.

D. Model Training

The previously discussed information was then used to train various ones machine learning algorithms are available in the scikit-lear library. The models used for the purposes of this study are Support Vector Equipment, Decision Tree, K Nearby Neighbors, Random Forest and algorithms of Naive Bayes. I trained models were now used to separate the packages themselves appropriate labels to identify user activity.

E. Performance Measures

Tests of various models were performed at the following steps: Accuracy, Accuracy, Remember, F1-Score and Training Time. A brief description of this functionalitysteps are given below.

1) Accuracy: Accuracy tells us how big we are the prediction is correct.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

2) Accuracy: The sum of good predictions on top of the predicted values available ingood category. It gives us a measure of how good it is the divider is.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

3) Remember: It gives us a measure of perfection a divides into categories. It is listed as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4) F1-score: Provided in the following formula:

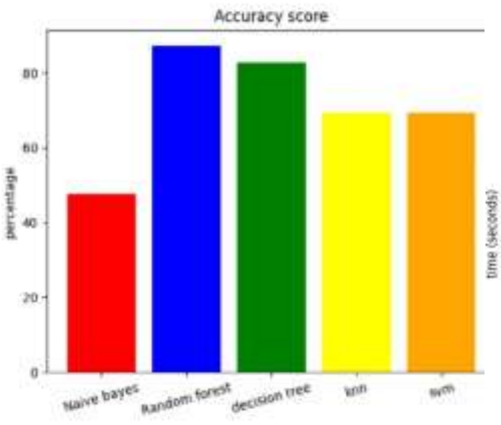
$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

It basically gives us a balance between accuracy and accuracy remember.

5) Training Time: It is the required time for training algorithm.

4.RESULTS AND OSERVATIONS

Accuracy and training time for all algorithms shown in Table IV and illustrated with diagram 2 once Fig. 3 in a row.



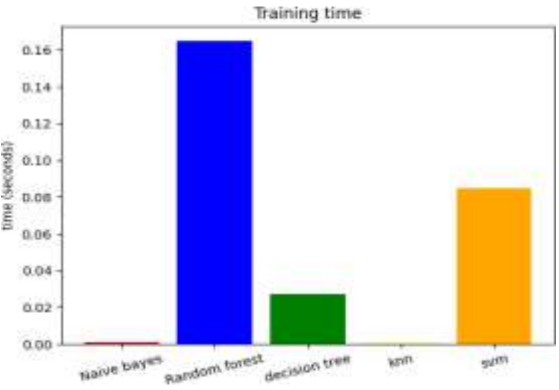


Fig:Accuracy of all algorithms

Fig2:Trainng time of all algorithms

Algorithm	Accuracy (%)	Training Time(s)
Naïve Bayes	47.57	0.0099
Random Forest	87.15	0.17424
Decision Tree	83.00	0.2725

KNN	69.00	0.0005
SVM	69.00	0.08633

From Table IV and Fig. 2, it is clear that the accuracy of The random forest is the highest (i.e. 87.15%) in all uses algorithms. We can see in Table IV and Figure 3 that I the time for random forest training is high and the training time KNN is the lowest of all algorithms. Figure 4, 5 and 6 show accuracy, recall and f1-score values respectively we have listed three different online applications The most accurate algorithms i.e. Random Forest, Decision Tree and SVM. From Fig. 4, 5 and 6 seem to be Random The forest algorithm provides the best accuracy, recall and f1-score values in many online programs compared to others algorithms. The most important factors were Return time limit between arrival, maximum advance time list and split loading rate according to the Random Forest classifier system.

5. CONCLUSION AND THE FUTURE EXTENSION

During this study, a new database was created since packages taken using Wireshark on various sites on 4G network. From the captured packages, 65 vague features were present extracted using a python script by splitting the packets in half various flow. After this, the database was pre-processed, and this The database was used to train 5 ML class dividers. Then these models used to separate packets according to their compatibility user function.

6. REFERENCES

[1] " Wireshark · Go Deep." https://www.wireshark.org/ (accessed Feb. 05,2021).

[2] J. M. Wang, C. L. Qian, C. H. Che, and H. T. He, "Study on process of network traffic classification using machinelearning,"2010, doi:10.1109/ChinaGrid.2010.53.

[3] K. Singh, S. Agrawal, and B. S. Sohi, " A Near Real-time IP Traffic Classification Using MachineLearning,"Int. Intell.Syst.Appl.,vol.5,no.3,2013,doi:10.5815/ijisa.2013.03.09.

[4] D. S. V, " Automatic Spotting of Sceptical Activity with Visualization Using Elastic Cluster for Network Traffic in Educational Campus," J.Ubiquitous Comput.Commun.Technol., vol. 2, no. 2, 2020, doi:10.36548/jucct.2020.2.004.

[5] GitHub - anupamraj1312/Flowmeter: A python script for extracting flow features from a PCAP file." https://github.com/anupamraj1312/Flowmeter (accessed Feb. 07, 2021).

[6] V. Labayen, E. Magaña, D. Morató, and M. Izal, "Online classification of user activities using machine learning on network traffic," Comput. Networks, vol. 181, 2020, doi: 10.1016/j.comnet.2020.107557.

[7] A. (Karunya U. Jamuna and V. (Karunya U. Ewards S.E, "Efficient Flow based Network Traffic Classification using Machine Learning," Int. J. Eng. Res. Appl., vol. 3, no. 2, 2013.

[8] M. Shafiq, X. Yu, and D. Wang, "Network traffic classification using machine learning algorithms," in Advances in Intelligent Systems and Computing, 2018, vol. 686, doi: 10.1007/978-3-319-69096-4_87.

[9] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," 2016, doi: 10.5220/0005740704070414.

[10] B. Yamansavascilar, M. A. Guvensan, A. G. Yavuz, and M. E. Karsligil, "Application identification via network traffic classification," 2017, doi: 10.1109/ICCNC.2017.7876241.