IJARST

International Journal For Advanced Research In Science & Technology

A peer reviewed international journal

www.ijarst.in

ISSN: 2457-0362

NETWORK TRAFFIC ANALYSIS USING MACHINE LEARNING

I.SWAPNA[1], SUVARTHA[2], P.SHRAVANI[3], L.SONY[4], K.MADHAVI[5] , D.KALPANA[6]

[1]Assistant Professor, Dept of CSE, Princeton Institute of Engineering and Technology for Women, Hyderabad, TS, India.

[2,3,4,5,6] UG Students, Dept of CSE, Princeton Institute of Engineering and Technology for Women, Hyderabad, TS, India.

ABSTRACT:

In the world of networking, it sometimes becomes essential to know what types of applications flow through the network for performance of certain tasks. Network traffic classification sees its main usage among ISP’s to analyze the characteristics required to design the network and hence affects the overall performance of a network. There are various techniques adopted to classify network protocols, such as port-based, pay-load based and Machine Learning based, all of them have their own pros and cons. Prominent nowadays is Machine Learning technique due to its vastness in usage in other fields and growing knowledge among researchers of its better accuracy among others when compared. In this paper, we compare two of the basic algorithms, Naïve Bayes and K nearest algorithm results when employed to networking data set extracted from live video feed using Wireshark software. For an implementation of Machine learning algorithm, python sklearn library is used with numpy and pandas library used as helper libraries. Finally, we observe that K nearest algorithm gives more accurate prediction than Naïve Bayes Algorithm, Decision Tree Algorithm and Support Vector Machine.

Keywords: ML, K NN, traffic, wire shark, basic algorithm.

1. INTRODUCTION:

The Network Traffic Classification plays a dynamic role pertaining to threats associated with the emerging technology nowadays. The various machine learning based classification techniques are presented in [1]. It helps internet service providers to manage overall performance of the network by considering the factors associated with certain application protocol. It has its usage in recognizing the unknown network if any tries to intrude the specified traffic lane. By this way we get to study its properties as well. Using the above property of recognizing unknown network, one can also recognize the potential threats that a network can suffer due to certain security attacks. Management of network security and Quality of Service (QoS) is also essential task and can be achieved if we have good techniques to classify network. Blocking or allowance of certain network traffic can also be achieved if we classify our network well. Overall, classification of network helps in overall growth of the network and its efficiency. The second technique that come into existence is Payload based technique in which packets of the associated networks are used to analysis and according to them protocol is identified. This technique is known as Deep Packet Inspection technique due to the fact that it uses packets for analysis. This technique is failed primarily due to the fact that it requires costly hardware installations and this does not work well for the packets that are encrypted. These drawbacks give the way to machine learning technique that has been using nowadays due to its efficiency of the results and similarity with the practical facts. In this technique, labeled classes are turned to model and then are trained and tested to check the correctness using accuracy. The Contributions of the paper are explained as follows. We initially discuss the different techniques and then we employ machine learning techniques to network data set and do a comparative analysis on different algorithm on which one is best suited to analyze the network traffic. We collect the features using Wireshark tool and then we convert this data to csv file format then train and test using Python Libraries which help to predict and further the comparative analysis is carried out. We employ Decision Tree (DT), Naïve Bayes (NB), Knearest neighbor (KNN), and Support Vector Machine (SVM) techniques. We find that KNN Technique outperforms for this application.

2. LITERATURE SURVEY

1) Port Based–All the ports are registered with IANA, by hashing the application protocol with the ports registered from IANA, one can easily identify the traffic in the network. For instance, the standardized port numbers assigned for sending and receiving E-mail are 25(SMTP) and 110(POP3), respectively. Now, these port numbers are universal standard for all the networks all over the globe. There are certain drawbacks associated with this technique. The major drawback felt in this type of classification is detection of only well-known port numbers. Another drawback is in cases of unregistered or dynamic port numbers that are not mentioned by IANA. Below is the table that illustrates the currently registered port numbers with IANA.

This technique also inlays certain drawbacks. The drawback associated with this is that it requires a very expensive hardware in a payload for pattern searching. In an encrypted form of traffic this technique is bound to fail. Finally, while working with new applications, this approach requires updating of new signature patterns [2].

3) Machine Learning Techniques - Machine learning has been used to classify various forms of data. In this project, this technique is employed to classify the application protocols. According to this technique, the machine (learning classifier) is trained, by given certain collection of data, in order to achieve maximum success rate, usually a large amount of data is employed for better training. After the training process, certain sample of data class is supplied to the classifier, so as to check how well the machine is trained, by evaluating the output of the machine and comparing them to original output. There are two types of machine learning techniques: Unsupervised and Supervised Machine Learning.

i) Unsupervised Technique: In this type of technique, A raw dataset is provided to the learning classifier, by raw we mean the data without any predefined labels or tags. This method of machine learning is also known as clustering. This technique divided the dataset into set of clusters with each data entry belonging to a specific cluster and hence can be used to further predict cluster to which future data entry would correspond to. But we cannot use this technique for network classification as it cannot cluster according to predefined class variables.

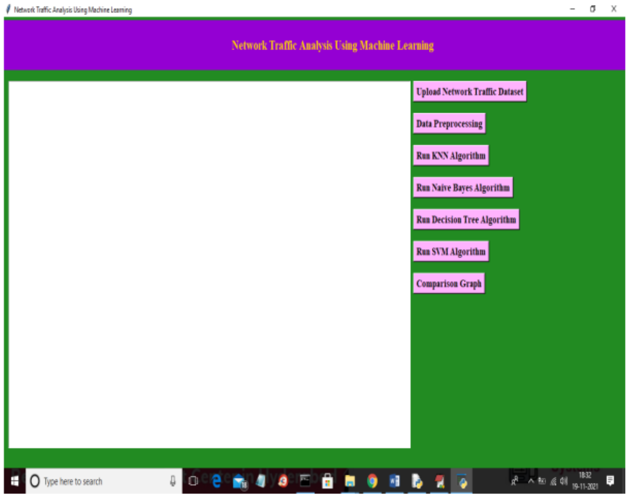
ii) Supervised Technique Yet another classification technique and a useful part of machine learning is Supervised Technique. Presence of a well labeled complete data set is requirement if one wishes to employ any of supervised learning method. The working process of this method is a two steps procedure, first the data sample is trained using the specified labels, and then it is then employed to test in a new data sample [2]. Hence a useful technique is used to classify the network traffic data.

3. METHODOLOGY

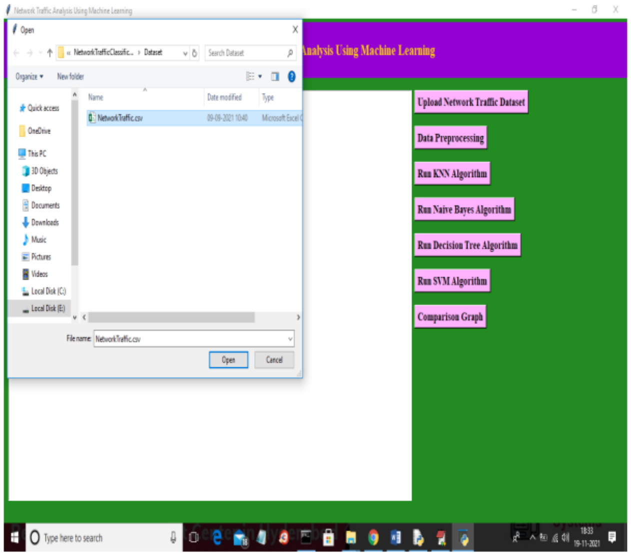
ALGORITHM:

The Convolutional Neural Network gained popularity through its use with image data, and is currently the state of the art for detecting what an image is, or what is contained in the image. CNNs even play an integral role in tasks like automatically generating captions for images. The convolutional neural network (CNN) is a type of multi-layer neural network, which extracts features by combining convolution, pooling, and activation layers. The CNN is widely used in the field of pattern recognition. Many researchers have applied the CNN to traffic sign recognition and detection and have achieved good results.

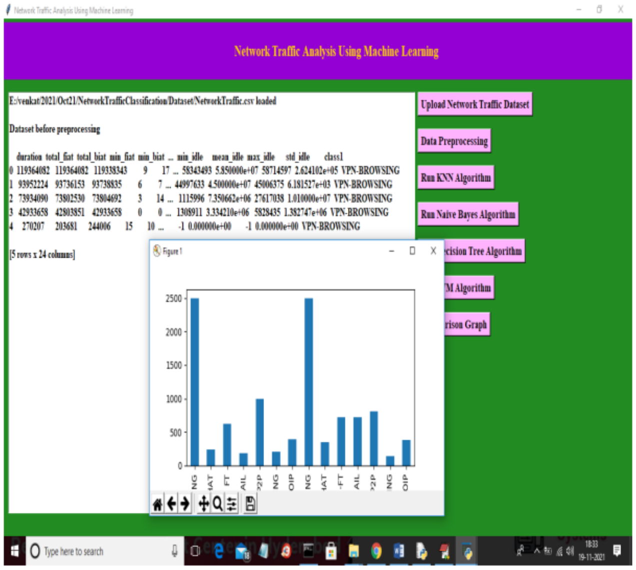
In the detection stage, the traffic signs were classified into two super classes: Circular and triangular traffic signs. In the classification stage, we trained three CNNs for two classification methods. One method trained two CNNs for circular and triangular traffic signs independently. The other method trained one CNN for the overall traffic sign classification. Each of the three CNNs had two convolutional layers, and each of the convolutional layers were followed by a sub sampling layer. They all used a fully connected layer to produce the final classification result. The eight Gabor features of each traffic sign were used as inputs of the three CNNs, with a fixed size of 32 × 32. The first convolutional layer extracted six features for each input with 8 × 6 kernels (size 5 × 5). Additionally, the second convolutional layer extracted 12 features for each input; hence, the second convolutional layer consisted of 6 × 12 kernels with a size of 5 × 5. The 12 feature maps from the second layer were used as feature vector inputs to the fully connected layer, to produce the final classification result.


Network Traffic Analysis Using Machine Learning
Network Traffic Analysis Using Machine Learning
Upload Network Traffic Dataset
Data Preprocessing
Run KNN Algorithm
Run Naive Bayes Algorithm
Run Decision Tree Algorithm
Run SVM Algorithm
Comparison Graph
Type here to search
18:32
19-11-2021

In above screen click on 'upload Network Traffic Dataset' button to upload dataset.



In above screen selecting and uploading 'NetworkTraffic.csv' file and then click on 'Open' button to load dataset and to get below screen.



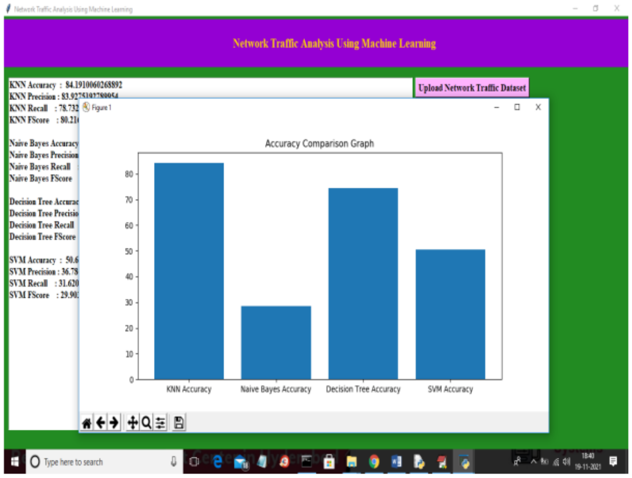
```
E:/venkat/2021/Oct21/NetworkTrafficClassification/Dataset/NetworkTraffic.csv loaded

Dataset before preprocessing

   duration  total_fiat  total_biat  min_fiat  min_biat  ...  min_idle     mean_idle  max_idle       std_idle         class1
0  119364082   119364082   119338343        9        17  ...  58343493  5.850000e+07  58714597  2.624102e+05  VPN-BROWSING
1   93952224    93736153    93738835        6         7  ...  44997633  4.500000e+07  45006375  6.181527e+03  VPN-BROWSING
2   73934090    73802530    73804692        3        14  ...   1115996  7.350662e+06  27617038  1.010000e+07  VPN-BROWSING
3   42933658    42803851    42933658        0         0  ...   1308911  3.334210e+06   5828435  1.382747e+06  VPN-BROWSING
4     270207      203681      244006       15        10  ...        -1  0.000000e+00        -1  0.000000e+00  VPN-BROWSING

[5 rows x 24 columns]
```

In above screen we can see dataset loaded and dataset contains lots of non-numeric values so we need to process it and in graph x-axis we can see traffic type and y-axis represents total records in dataset for that traffic. Now close above graph and then click on 'Data Preprocessing' to clean dataset.



```
KNN Accuracy  :  84.1910060268892
KNN Precision : 83.9275192789954
KNN Recall    : 78.732
KNN FScore    : 80.21

Naive Bayes Accuracy
Naive Bayes Precision
Naive Bayes Recall
Naive Bayes FScore

Decision Tree Accurac
Decision Tree Precisio
Decision Tree Recall
Decision Tree FScore

SVM Accuracy  :  50.6
SVM Precision : 36.78
SVM Recall    : 31.620
SVM FScore    : 29.90
```

In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and in all algorithms KNN shows better result.

CONCLUSION

The Network traffic classification techniques are discussed in this paper to enhance some idea about Machine Learning algorithms for network traffic data. The analysis carried out definitely helps to a new analyst to make the decision about which Machine Learning algorithm is more appropriate for this application. Initially, the network traffic extraction is carried out to evaluate the different Machine Learning algorithm which is trained in later phase. The Machine Learning algorithms are used for managing the performance of network and classification of unknown applications.We then employ four basic Machine Learning algorithms to analyze the protocol. Further, the classifiers using different Machine Learning algorithms are developed to compare the accuracy for this network traffic data. We find that K-nearest neighbor (KNN) algorithm outperforms Naïve Bayes algorithm, Decision Tree and Support Vector Techniques in terms of accuracy which is due to the fact that KNN uses better classification criterion than Naïve Bayes and Decision Tree Algorithm. We find that KNN is most robust among the algorithms: NB, DT, and SVM for out training data set. It is also able to maintain highest mean for accuracy.

REFERANCES

[1] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay. Non-uniform quantized data fusion rule alleviating control channel overhead for cooperative spectrum sensing in cognitive radio networks. in 2017 IEEE 7th International Advance Computing Conference (IACC). 2017. IEEE.

[2] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay, Non-uniform quantized data fusion rule for data rate saving and reducing control channel overhead for cooperative spectrum sensing in cognitive radio networks. Wireless Personal Communications, 2019. 104(2): p. 837-851.

[3] Rueda, A. A survey of traffic characterization techniques in telecommunication networks. in Proceedings of 1996 Canadian Conference on Electrical and Computer Engineering. 1996. IEEE.

[4] Shahbar, K. and A.N. Zincir-Heywood. How far can we push flow analysis to identify encrypted anonymity network traffic? in NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium. 2018. IEEE.

[5] Axelsson, S., Intrusion detection systems: A survey and taxonomy. 2000, Technical report.

[6] Wang, P., Y. Li, and C.K. Reddy, Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 2019. 51(6): p. 110.

[7] Namdev, N., S. Agrawal, and S. Silkari, Recent advancement in machine learning based internet traffic classification. Procedia Computer Science, 2015. 60: p. 784-791.

[8] Cheng, Y., et al., Bridging machine learning and computer network research: a survey. CCF Transactions on Networking, 2019. 1(1- 4): p. 1-15.

[9] Mukkamala, S., G. Janoski, and A. Sung. Intrusion detection: support vector machines and neural networks. in proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), St. Louis, MO. 2002.

[10] Taylor, V.F., et al., Robust smartphone app identification via encrypted network traffic analysis. IEEE Transactions on Information Forensics and Security, 2017. 13(1): p. 63-78.