ISSN No:-2456-2165

Exploratory Data Analysis of Network Traffic

G. S. Nagaraja, Kruthi P.R., Shivani Deshpande
Department of Computer Science and Engineering
R V College of Engineering, Affiliated to VTU, Bengaluru, India

Abstract:- We cannot deny that internet and many applications running on internet are growing at an exponential rate. It has become more important for network administrators to have an understanding of the different types of network traffic.

This paper reviews the topic of exploratory data analysis of captured network traffic. There has been an exponential rise in the use of web apps such as social media sites, e-commerce, video streaming services, blogs, e-banking lately. These application has now become the day-to-day mode of communication all over the world and its importance is ever growing.

In this paper, we explore and investigate the data being transmitted by 75 apps and segregate the users of these apps based on network usage and try to determine the most used apps. The output of this exploratory data analysis can be applied in various fields.

Keywords:- *Data preprocessing, data mining, k-means clustering.*

I. INTRODUCTION

Browsing the internet often bring about many common questions about the internet streaming speed, download and upload speed, accessing shared locations, about computers and routers responding slowly. These are answered by network engineer, systems administrator or security engineer. System administrators perform the cumbersome and complicated troubleshooting process the network admin, perform the task of analyzing the network flow. Network traffic analysis involves various tasks of checking the network and to get a good understanding for what is happening in the network. In the process of analysis the data packets are decoded and the network traffic is put in a readable format. This is done to make sure that unknown intrusions has not occurred in the network.

The network analysis is important for making entire network infrastructure secure and strong against attacks by detecting and preventing any malicious activity. Also to detect any anomalies in the network. Network analysis is also done to check the load on web applications to improve applications scalability and help in load balancing.

II. BACKGROUND

Network can be analysed at different levels :

Flow Level 2. Packet Level and 3. Network Level. Below diagrams and tables give a general idea of different types of analysis done and protocols used.

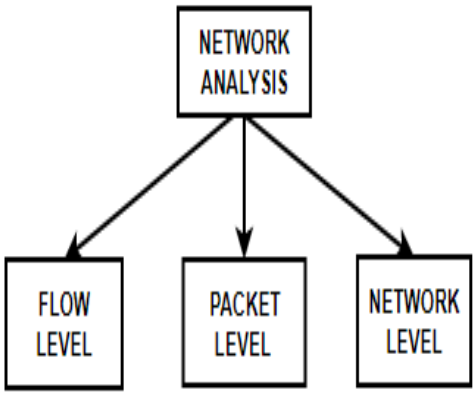

Fig. 1: Classification of network levels.

The network level analysis can be summarized as below:

Level	Protocol	Feature Set
Flow		Duration of the flow, Volume of data, Number of packets per flow.
	TCP	Quantity of keep-alive packets in a flow, Packet inter-arrival time, Number of ports reusing the packets
	IP	Destination IP, IP Geo-location, IP Autonomous system number.
Packet		Length of the packet, mean and variance of the packet length, square of the root mean.
	TCP	Packet Size, Destination port, Number of packets with PUSH bit set, Number of out of order packets.
Network	HTTP	Hostname, current type, referrer, Cookies Agent type.

Table 1: Classification of network protocols

ISSN No:-2456-2165

Network traffic analysis is done at different levels with respect to traffic in the following manner:

A. Source/Destination Analysis
B. Traffic Volumetric Analysis
C. Sequence Analysis
D. Inferential Analysis

The analysis of the above information is summarized as follows:

Analysis	Method	Example
Source/Destination Analysis	The source and the destination of the message is under scrutiny. Message contents are not of importance.	1. Emails sent by crucial government employee. 2. System processing credit card transaction.
Traffic Volumetric Analysis	Number of messages between communicating parties are scrutinized.	1. Communication between different units involved during war situation. 2. Radio silence situation.
Sequence Analysis	Repeatedly occurring communication pattern with almost same size of data.	1. Messages communicated to the members of political opposition group.
Inferential Analysis	Analyzing probably independent events, which constitute the traffic and can forecast upcoming event.	1. Information about emails from event booking company

Table 2: Analysis network information and their methods

III. METHODOLOGY

The dataset used in this paper contains eighty-seven features with each record having the information of an IP flow from the all the devices used in the network such as source and destination IP addresses, ports, inter arrival times. Most of the attributes in this dataset are of numeric type and some are nominal types. There is also a date type for the timestamp.

The main aim of this paper was to explore the dataset and check the apps which had the maximum network traffic. First, we classified the users based on the flow rates, i.e.: the usage of an app by a particular user and segregated them as high, medium and low profile users.

After some data pre-processing and cleaning we were able to find 52 protocol or apps which had the maximum network traffic out of which the first 5 apps took almost 95% of the entire user traffic. Hence these apps were tagged as most used apps.

Below is the methodology adopted for the network traffic analysis:

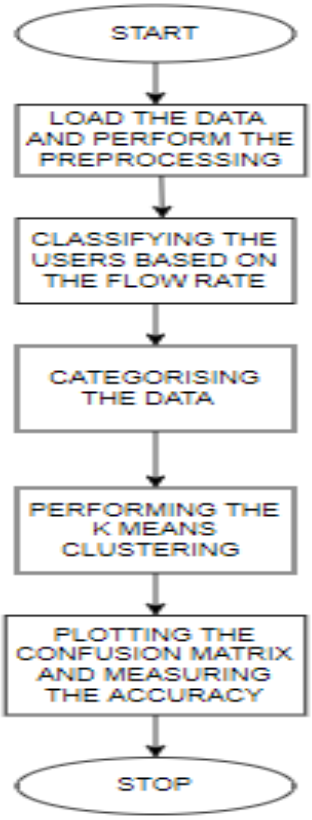

Fig. 2: Methodology for traffic analysis

Various data pre-processing techniques carried out in the work:

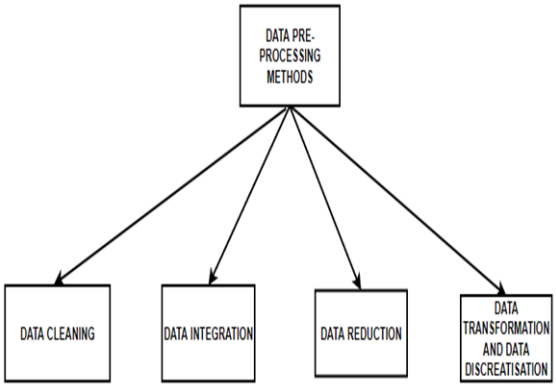

Fig. 3: Various Data Pre-Processing methods

The different network analysis methodologieswhich are available belong to data mining techniques, visualization techniques, machine learning techniques:

ISSN No:-2456-2165

A. Data Mining Techniques

Data mining techniques can be defined as the methods adopted to gain or uncover the significant patterns hidden in the data. The data is preprocessed with different methodologies before the data scientist perform the mining operations. Clustering, association rule mining, regression, classification are some of the majorly used mining techniques.

B. Visualization Techniques

Method to help data scientist gain insightful trends and patterns which might not be directly visible from the data. Dependency on data analyst for output and interpretations. Graphical representation of sets of data is usually termed as data visualisation. Huge or abstract, can help visualising the patterns hidden in the data effectively.

C. Machine Learning Techniques

Machine learning techniques are the methods adopted to be implemented for achieving the task to be accomplished by the machine or any hardware/software device. Bayesian network, fuzzy logic, Hidden Markov Models, neural networks are widely used machine learning methods.

They are further classified as:

- Supervised Learning
- Semi-Supervised Learning: k-means
- Unsupervised Learning

Unsupervised learning can be defined as a toolaiming for the setting in which a set of features are measured on *n* observations.Performing clustering on the observations of a data set, results in partitioning them into clusters such that the data points withineach group are having less inter-cluster distance.Clustering targets to form homogeneous groups among the data points. The methodology followed to obtain the network traffic analysis is k-means clustering algorithm. The data obtained was pre-processed before applying the data mining techniques. Each IP address was treated as an individual vector in the dataset.

D. Algorithm 1:K-Means Clustering

- Randomly assign a number, from 1 to *K*, to each of the observations.
- These serve as initial cluster assignments for the observations.
- Iterate until the cluster assignments stop changing:

a) For each of the *K* clusters, compute the cluster *centroid*. The*k*th cluster centroid is the vector of the *p* feature means for theobservations in the *k*th cluster.

b) Assign each observation to the cluster whose centroid is closest
(where*closest* is defined using Euclidean distance).

Since the input data was an unsupervised data, K-means clustering algorithm was employed to obtain the insights from the data set. The features of a network were considered as a set of related data points. The vectors with relatively less inter-cluster distance would be aggregated into one cluster and the same process would performed for all the different features in the data set.

IV. CONLUSION

Network Analysis is an essential task when network security and visualizing the vulnerabilities of the network is concerned. Knowing the data speeds and the packet source and destination details is very much essential in today's digital word. It supports users in finding the breaches or any anomalies in the network. Optimization of network level usage of application and its relevant bandwidth can be understood with the help of data visualization. In our experiment, K-means clustering Algorithm presented us with a huge number of patterns to analyze the activities and patterns happening on the network. It presented us with the potential web users and their respective websites frequently visited by them.

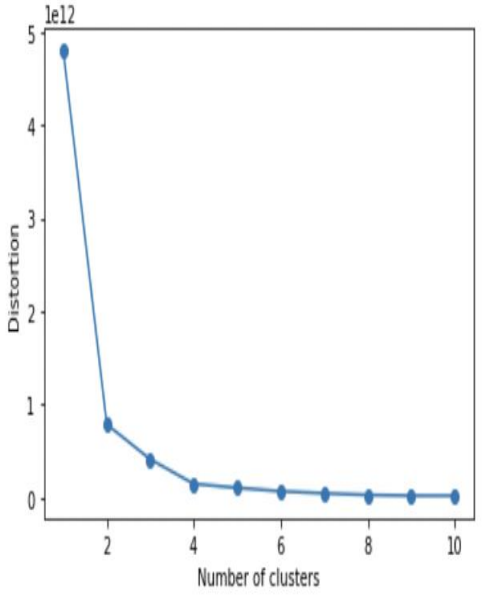

Fig. 4: Clustering formed using K-means clustering method

```
[[ 18895   6938  19304  12114]
 [ 20598    382  16540   1933]
 [108218   7541  86220    473]
 [802501 1665495 331728  68788]]
              precision    recall  f1-score   support

           0       0.02      0.33      0.04     57251
           1       0.00      0.01      0.00     39453
           2       0.19      0.43      0.26    202452
           3       0.83      0.02      0.05   2868512

    accuracy                           0.06   3167668
   macro avg       0.26      0.20      0.09   3167668
weighted avg       0.76      0.06      0.06   3167668
```

Fig. 5: Accuracy measured for K-means process

ISSN No:-2456-2165

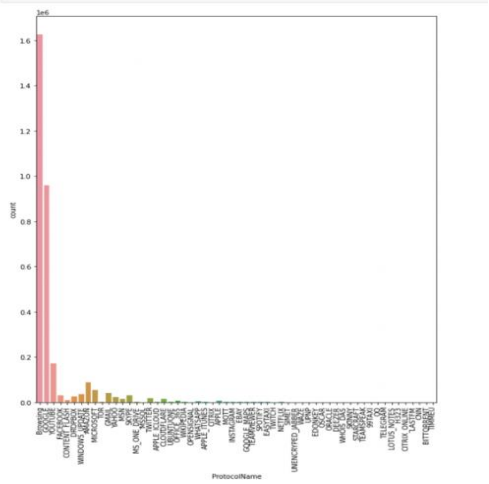

Fig. 6: Graph of the high traffic flow apps

REFERENCES

[1.] Raimir Holanda Filho and José Everardo Bessa Maia ,"Network Traffic Prediction using PCA and K-means", ,2010 IEEE Network Operations and Management Symposium - NOMS 2010 Year: 2010 | Conference Paper | Publisher: IEEE

[2.] Olumide Kayode, Ali Saman Tosun , "Analysis of IoT Traffic using HTTP Proxy", IEEE , ICC 2019 - 2019 IEEE International Conference on Communications (ICC) Year: 2019 | Conference Paper | Publisher: IEEE

[3.] Borja Molina-Coronado;Usue Mori;Alexander Mendiburu;Jose Miguel-Alonso ,"Survey of Network Intrusion Detection Methods From the Perspective of the Knowledge Discovery in Databases Process,IEEE Transactions on Network and Service Management

[4.] Sheikh Muhammad Farjad, Asad Arfeen ,"Cluster Analysis and Statistical Modeling: A Unified Approach for Packet Inspection",2020 International Conference on Cyber Warfare and Security (ICCWS) Year: 2020 | Conference Paper | Publisher: IEEE

[5.] Sheetal Thakare, Anshuman Pund, Dr.M.A.Pund.,"Network Traffic Analysis, Importance, Techniques: A Review", 2018 3rd International Conference on Communication and Electronics Systems (ICCES)

[6.] Nasser, D. Hamad, C. Nasr,"Visualization Methods for Exploratory Data Analysis",2006 2nd International Conference on Information & Communication Technologies.

[7.] Liangqi Chen, Ben Wang, Yinggui Wang, Xiya Wang,"Exploratory Data Analysis on the Usage of COVID-19 Vaccine",Year:2021 | Conference Paper | Publisher: IEEE.

[8.] Pedro M. Santiago del Río;Javier Ramos;Alfredo Salvador;Jorge E. López de Vergara;Javier Aracil;Antonio Cuadra;Mar Cutanda,"Application of Internet traffic characterization to all-optical networks",2010 12th International Conference on Transparent Optical Networks Year: 2010 | Conference Paper | Publisher: IEEE.

[9.] M. T. Asif;J. Dauwels;C. Y. Goh;A. Oran;E. Fathi;M. Xu;M. M. Dhanya;N. Mitrovic;P. Jaillet ,"Unsupervised learning based performance analysis of n-support vector regression for speed prediction of a large road network,2012 15th International IEEE Conference on Intelligent Transportation SystemsYear: 2012 | Conference Paper | Publisher: IEEE "

[10.] Shizeng Lu;Hongliang Yu;Xiaohong Wang;Qiang Zhang;Fanjun Li;Zhao Liu;Fangqian Ning ,"Clustering Method of Raw Meal Composition Based on PCA and Kmeans ", 2018 37th Chinese Control Conference (CCC)