ISSN 2278-3091

Volume 9, No.1.3, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6691.32020.pdf

https://doi.org/10.30534/ijatcse/2020/6691.32020

Network Traffic Profiling Using Data Mining Technique in Campus Environment

Muhammad Azizi Mohd Ariffin[1], Rusmawati Ishak[2], Siti Arpah Ahmad[3], Zolidah Kasiran[4]

[1] Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia,
mazizi@fskm.uitm.edu.my

[2]Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia,
arpah@fskm.uitm.edu.my

[3]Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia, ri4910@gmail.com

[4]Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia,
zolidah@fskm.uitm.edu.my

ABSTRACT

The exponential growth of internet usage poses a challenge in managing the bandwidth and securing the campus network environment. The ability differentiates and profile different type of data traversing in the Internet traffic is essential for ensuring effective bandwidth distribution and safeguarding network security. This work implements unsupervised data mining approach to analyze the network traffic trend and type of traffic in campus network. In this research, Orange tool is used in implementing K-Means Clustering Algorithm to find a network trend pattern of user accessing the Internet and to produce network traffic profiling in high volume traffic. Three clusters have been created based on the network traffic data described as high, medium and low number of hits towards the protocol services and unique IP address. Results shows that 74,869 hits come from DNS (UDP), 40,658 hits from MySQL and 3191 hits from HTTP. These are the high traffic that consume the bandwidth. Data mining process lead to reveal the information gather for profiling purposes and identify type of traffic passing through the campus network. The outcome of this study can be a recommendation of managing or shaping the bandwidth usage and strengthen the security policy of the network.

Key words: Traffic Profiling, Data Mining, K-Means Algorithm, Network Security

1. INTRODUCTION

Internet users are increasing, and majority of the usage is towards Over-the-top (OTT) media service messaging platform such as Whatsapp and Telegram; obtaining information such as Wikipedia; for entertainment such as Youtube and Netflix; and for work such as Google Docs. 0.4 million increase of Internet users is reported in Malaysia between 2015 to 2016[1]. This tremendous increase in Internet users and traffic poses a challenge to network administrator in managing the resources such as network bandwidth distribution and in the same time protecting users' security and privacy. Internet usage mostly are communication carried over TCP/IP protocols that consists of data, voice, video and messaging are carried over TCP/IP protocols [2]. Profiling the network traffic will support the network administration job by giving the pattern of the traffic to differentiate unwanted traffic such as Malware, botnet [3,][4], phishing email DOS attack [5] and data breach[6]. Bandwidth distribution among users could be maintained by the traffic pattern. Network abuse or hog the bandwidth capacity when user run P2P download or video streaming could be stopped [7].

Data mining consists of processes that extract information from a pool of unknown data [8]. The k-means algorithm utilized clustering approach to cluster data such as network traffic into groups of related data without any prior knowledge of those relationships; known as un-supervised learning [9].

There are previous studies that profiled the network traffic using K-means clustering technique [6],[7]. The study of [7] used K-Means algorithm to identify network traffic activities related to teaching and learning by identifying the bandwidth of users' pattern. But the study did not classify network traffic and the scope of the study is not performed in campus network environment. The study of [6] and [10] used K-means clustering for classifying the network log in order to identify internet used behavior. However, the study did not classify the traffic directly as it was performed on the log rather than at network layer packets. Besides that, the work of [11] used K-means algorithm and Euclidean distance on Netflow protocol traffic to determine anomalies caused by UDP flood attack. Other than that, there is an effort from [12],[3] to used data mining technique such as K-Nearest Neighbor to identify anomalies in heart disease and social media. The work of [14] used data mining technique known as Genetic algorithm in order to find a trend in stock market. From the previous study, none of the work used K-means clustering algorithm as a data mining technique to conduct network traffic profiling in campus network environment.

This work implements unsupervised data mining approach to analyze the network traffic trend and type of traffic in campus network. In this research, Orange tool is used in implementing K-means clustering algorithm that could be seen as the most suitable solution [15] to find a network trend pattern of user accessing the Internet and to produce network traffic profiling in high volume traffic. The outcome of this study can be a recommendation for managing or shaping the bandwidth usage and strengthen the security policy as well as help the organization to manage the network and assist the connectivity issue faced by internal and online user in campus environment.

2. REVIEW OF THE LITERATURE

Profiling a network traffic involves finding metadata in a large amount of data [16]

that moved across network at certain time [17]. Example of network traffic profiling usage is to understand Internet users' behavior [18] and to map the level of security treat and identify suspected abuser [10].

Data mining is a process of profiling large data into useful information by using certain techniques. Examples of data mining techniques used are classification, clustering and regression. Classification generally weighs the input data so that the output can be separated into two values. Examples of classification algorithms are neural network, Naive Bayes [19] and K-nearest algorithm [20]. Clustering works by grouping the similarity of the input data. Examples of clustering algorithms are K-means, Mean-Shift and Clustering Mixture Model [21],[22]. Regression technique is used to predict a range of numeric values from a dataset. This work applies K-Mean algorithm on network traffic data.

K-means find the similarity among the data by calculating the means values. It starts by randomly identify points, terms as K. If the K value is 2, it means 2 point is chosen. For each point, the means values between the point and surrounding data is determined. The calculation of the means values might be done using Euclidean distance, Manhattan, Cosine [6].

The are many tools to implement the clustering algorithms. For example, Orange, Rapid Miner, Knime, WEKA, KEEL and R [12] . This works used Orange. It is open source and was developed in 2009 under GNU General Public License. Orange is compatible with Phyton language for machine learning. It consists of machine learning suite, components-based data mining, visual programming front-end, explorative data analysis, Phyton binding and libraries for scripting.

Campus network can be referred to a corporation, government agency or university that interconnected via local area network (LAN) [23]. However, most of the previous studies on investigated network traffic were limited to university campus network only [24-26]. For example, email traffic workload of University of Calgary's has been investigated [24] by comparing the email protocols; IMAPS and SMTP. Fulda University network traffic flow has been used for deep neural networks (DNNs)training for the purpose of improving network utilizing [25]. The flow of data across the network of The Federal University of Technology, Akure (FUTA) has been used to model the queuing model analysis by investigating the TCP and UDP protocol-based packets [26]. To the best of our knowledge, there is no work that use corporation, government agency to investigate the network traffic characteristic. This work used those traffics as a part of dataset, however the name of the organization is not mention due to the privacy issue.

3. METHODOLOGY

This works consists of the three phases. Phase one is data collection and pre-processing. Phase two is datasets preparation. Third phase is to overall implement of the work

3.1 Phase 1- Data Collection and pre-processing

The network traffic data is taken from a campus network. A campus network can be defined as a corporation, government agency or university that interconnected via local area network (LAN) [23]. The dataset is collected from switches and firewall packet and log using Wireshark and WinSCP tool. The data collection is conducted for two days (1st and 2nd May 2018) and produced 235,141 records.

The pre-processing involves the transformation of raw network data to impute missing values and normalize features. Secure Copy Protocol (SCP) is used to transfer the raw data. The Raw data is loaded into spreadsheet using comma- or table limited format for the data cleaning with filter and sort function. During this process the erroneous or missing data is identified.

3.2 Phase 2 -Datasets preparation

Orange is the data mining tool used to cluster the network traffic data into useful information. Phyton language is needed to run the Orange tool successfully [27].

Three datasets are used in this work. The first two datasets are used to validate the Orange tool. The first dataset is Iris dataset; a small dataset that is often used to test machine learning algorithms and visualizations. The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant [28]

. It come with Orange tool. The purpose of using this dataset is to validate the functionality of Orange in the clustering process. The second dataset is KDDCup dataset. This is benchmark network traffic dataset. It is used to test whether Orange can accept network traffic data and cluster it correctly. Both datasets do not undergo the cleaning process since they already in acceptable form to be inserted into Orange tool. The final dataset is the real campus network data set taken from an organization. The aim of the clustering is to identifies type of traffic generated by the users. Recommendation of the bandwidth usage and suggestion of the security policy of the network is derived from the results produced.

3.3 Phase 3 – Overall Implementation

This work used Laptop of Dell Latitude E5470, with Windows10_64-bit, Hard disk of 500TGB, RAM of 16G and CPU of Intel-CoreI7.The Network interface, IP Address and Netmask are required. The data mining tool is Orange 3.13 software with Phyton for scripting language. Network traffic log files collecting required SCP to transfer the data, Notepad++ and Excel to extract the log files. Software monitoring tool used is Wireshark.

The K-means clustering algorithm is used to profile the data from the datasets. Iris dataset used 150 sample dataset, KDDCup dataset and the real campus network datasets used 5000 samples. 300 iteration is used with 10-time re-run processes. The K value is 2.

Fig. 1 illustrates the overall implementation of the work. The work starts with Phyton data mining library preparation, then Orange tool can be launched. Then the datasets are inserted only after it is in the proper format. After that the K-means can be run. Further, the Visually scatted Plot of the results are produced. Profiling analysis is done at the end.

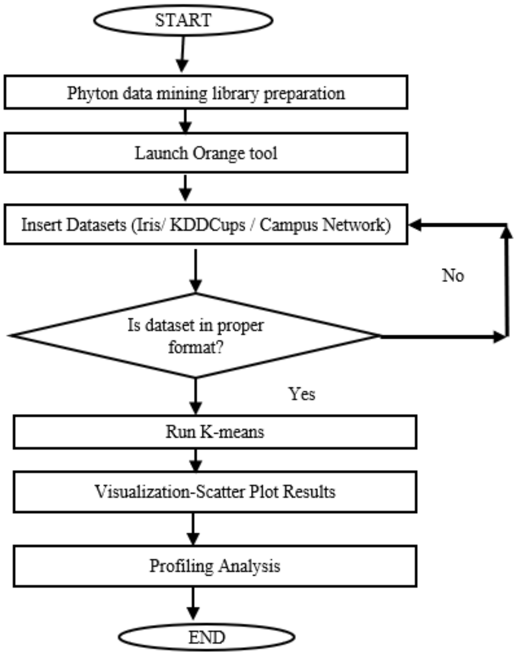

START
Phyton data mining library preparation
Launch Orange tool
Insert Datasets (Iris/ KDDCups / Campus Network)
No
Is dataset in proper format?
Yes
Run K-means
Visualization-Scatter Plot Results
Profiling Analysis
END

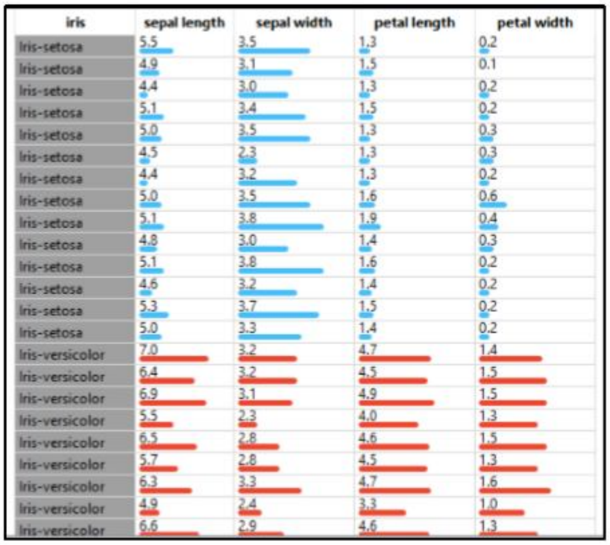
Figure. 1: Overall work implementation

4. RESULTS AND DISCUSSION

The results discussed in the following present the Iris and KDDCup datasets in validating the Orange tool. Section 4.3 illustrates the campus network datasets profiling results, while Section 4.4 is the technical recommendation for the campus network implementation.

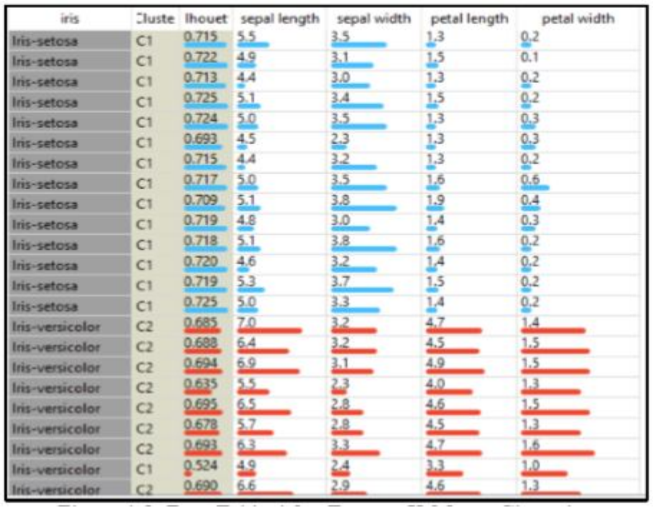
4.1 Dataset Validation using Iris Dataset

The Figure 2 and Figure 3 show the Iris datasets before and after the execution of K-means clustering. Column in Figure 2 consists of iris information, sepal length sepal width, petal length and petal width. Before the K-means clustering, the data is raw, and no clustering information provided. Figure 3 show the implementation of the K-means. It shows that two columns have been added: Cluster and Silhouette. These results show that the tool able to cluster the Iris datasets. Two cluster have been created with its Silhouette values accordingly. The result of visualizing the dataset after k-means with k=2.



iris	sepal length	sepal width	petal length	petal width
Iris-setosa	5.5	3.5	1.3	0.2
Iris-setosa	4.9	3.1	1.5	0.1
Iris-setosa	4.4	3.0	1.3	0.2
Iris-setosa	5.1	3.4	1.5	0.2
Iris-setosa	5.0	3.5	1.3	0.3
Iris-setosa	4.5	2.3	1.3	0.3
Iris-setosa	4.4	3.2	1.3	0.2
Iris-setosa	5.0	3.5	1.6	0.6
Iris-setosa	5.1	3.8	1.9	0.4
Iris-setosa	4.8	3.0	1.4	0.3
Iris-setosa	5.1	3.8	1.6	0.2
Iris-setosa	4.6	3.2	1.4	0.2
Iris-setosa	5.3	3.7	1.5	0.2
Iris-setosa	5.0	3.3	1.4	0.2
Iris-versicolor	7.0	3.2	4.7	1.4
Iris-versicolor	6.4	3.2	4.5	1.5
Iris-versicolor	6.9	3.1	4.9	1.5
Iris-versicolor	5.5	2.3	4.0	1.3
Iris-versicolor	6.5	2.8	4.6	1.5
Iris-versicolor	5.7	2.8	4.5	1.3
Iris-versicolor	6.3	3.3	4.7	1.6
Iris-versicolor	4.9	2.4	3.3	1.0
Iris-versicolor	6.6	2.9	4.6	1.3

Figure 2 : Iris dataset before the K-means clustering

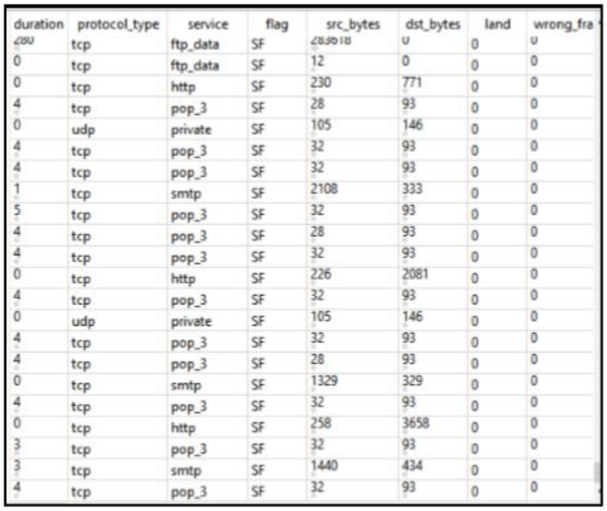


iris	Cluste	lhouet	sepal length	sepal width	petal length	petal width
Iris-setosa	C1	0.715	5.5	3.5	1.3	0.2
Iris-setosa	C1	0.722	4.9	3.1	1.5	0.1
Iris-setosa	C1	0.713	4.4	3.0	1.3	0.2
Iris-setosa	C1	0.725	5.1	3.4	1.5	0.2
Iris-setosa	C1	0.724	5.0	3.5	1.3	0.3
Iris-setosa	C1	0.693	4.5	2.3	1.3	0.3
Iris-setosa	C1	0.715	4.4	3.2	1.3	0.2
Iris-setosa	C1	0.717	5.0	3.5	1.6	0.6
Iris-setosa	C1	0.709	5.1	3.8	1.9	0.4
Iris-setosa	C1	0.719	4.8	3.0	1.4	0.3
Iris-setosa	C1	0.718	5.1	3.8	1.6	0.2
Iris-setosa	C1	0.720	4.6	3.2	1.4	0.2
Iris-setosa	C1	0.719	5.3	3.7	1.5	0.2
Iris-setosa	C1	0.725	5.0	3.3	1.4	0.2
Iris-versicolor	C2	0.685	7.0	3.2	4.7	1.4
Iris-versicolor	C2	0.688	6.4	3.2	4.5	1.5
Iris-versicolor	C2	0.694	6.9	3.1	4.9	1.5
Iris-versicolor	C2	0.635	5.5	2.3	4.0	1.3
Iris-versicolor	C2	0.695	6.5	2.8	4.6	1.5
Iris-versicolor	C2	0.678	5.7	2.8	4.5	1.3
Iris-versicolor	C2	0.693	6.3	3.3	4.7	1.6
Iris-versicolor	C1	0.524	4.9	2.4	3.3	1.0
Iris-versicolor	C2	0.690	6.6	2.9	4.6	1.3

Figure 3 : Iris dataset after the K-means clustering

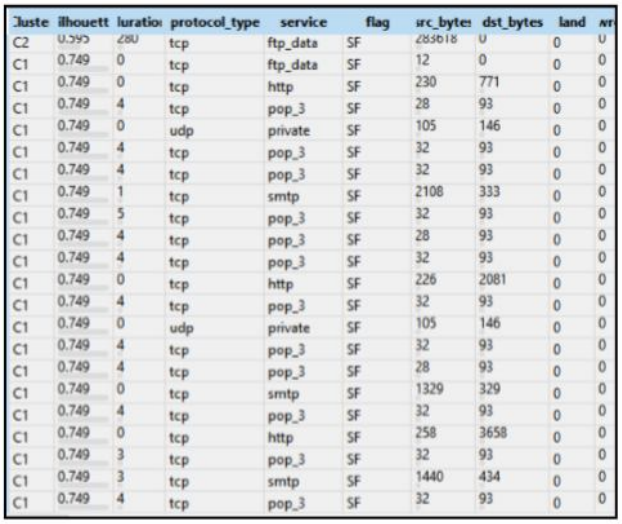
4.2 Dataset Validation using KDDCup Dataset

The Figure 4 and Figure 5 shows the result before and after implementing of K-means on KDDCup dataset. Identical to Iris dataset, Orange tool able to cluster the data by adding the cluster and Silhouette to the table. Based on KDDCup dataset clustering it shows that most common features of the datasets are Service and Protocol type. Figure 6 shows the example of Services of the network traffic are ftp_data, smtp and http. Figure 7 illustrates the KDDCup clustering on protocol type.



duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fra
280	tcp	ftp_data	SF	283618	0	0	0
0	tcp	ftp_data	SF	12	0	0	0
0	tcp	http	SF	230	771	0	0
4	tcp	pop_3	SF	28	93	0	0
0	udp	private	SF	105	146	0	0
4	tcp	pop_3	SF	32	93	0	0
4	tcp	pop_3	SF	32	93	0	0
1	tcp	smtp	SF	2108	333	0	0
5	tcp	pop_3	SF	32	93	0	0
4	tcp	pop_3	SF	28	93	0	0
4	tcp	pop_3	SF	32	93	0	0
0	tcp	http	SF	226	2081	0	0
4	tcp	pop_3	SF	32	93	0	0
0	udp	private	SF	105	146	0	0
4	tcp	pop_3	SF	32	93	0	0
4	tcp	pop_3	SF	28	93	0	0
0	tcp	smtp	SF	1329	329	0	0
4	tcp	pop_3	SF	32	93	0	0
0	tcp	http	SF	258	3658	0	0
3	tcp	pop_3	SF	32	93	0	0
3	tcp	smtp	SF	1440	434	0	0
4	tcp	pop_3	SF	32	93	0	0

Figure 4: KDDCup dataset before K-means clustering using Orange tool.



Cluste	ilhouett	luratio	protocol_type	service	flag	src_byte	dst_bytes	land	wr
C2	0.595	280	tcp	ftp_data	SF	283618	0	0	0
C1	0.749	0	tcp	ftp_data	SF	12	0	0	0
C1	0.749	0	tcp	http	SF	230	771	0	0
C1	0.749	4	tcp	pop_3	SF	28	93	0	0
C1	0.749	0	udp	private	SF	105	146	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	1	tcp	smtp	SF	2108	333	0	0
C1	0.749	5	tcp	pop_3	SF	32	93	0	0
C1	0.749	4	tcp	pop_3	SF	28	93	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	0	tcp	http	SF	226	2081	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	0	udp	private	SF	105	146	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	4	tcp	pop_3	SF	28	93	0	0
C1	0.749	0	tcp	smtp	SF	1329	329	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0
C1	0.749	0	tcp	http	SF	258	3658	0	0
C1	0.749	3	tcp	pop_3	SF	32	93	0	0
C1	0.749	3	tcp	smtp	SF	1440	434	0	0
C1	0.749	4	tcp	pop_3	SF	32	93	0	0

Figure 5: KDDCup dataset after K-means clustering using Orange tool.

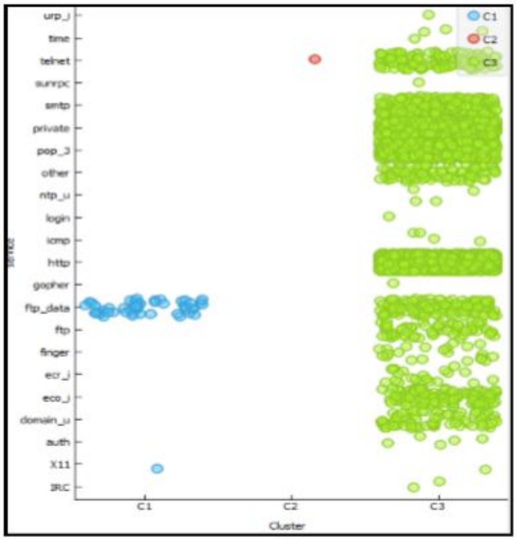


Figure 6: KDDCup clustering by K-means on services.

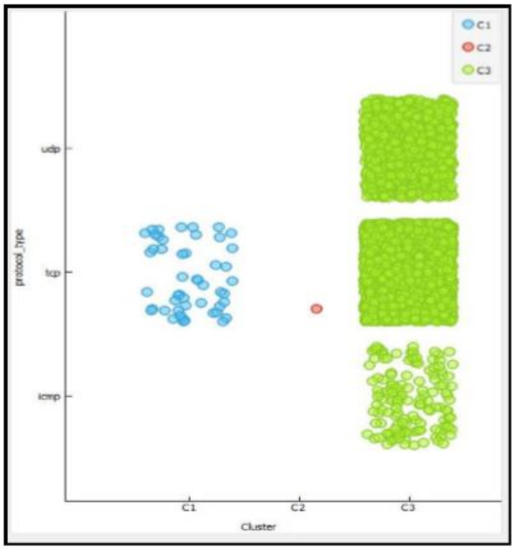
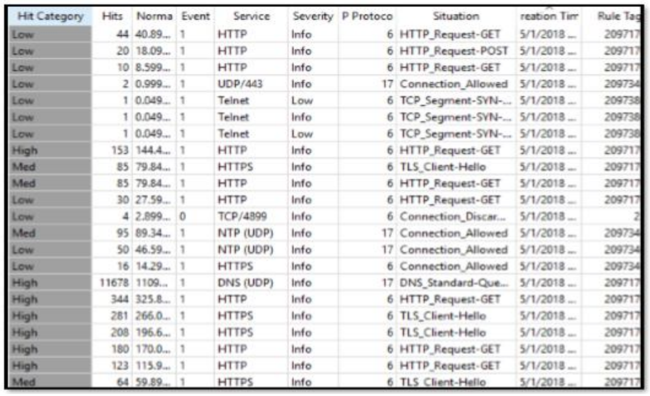


Figure 7: KDDCup clustering by K-means on protocol_type.

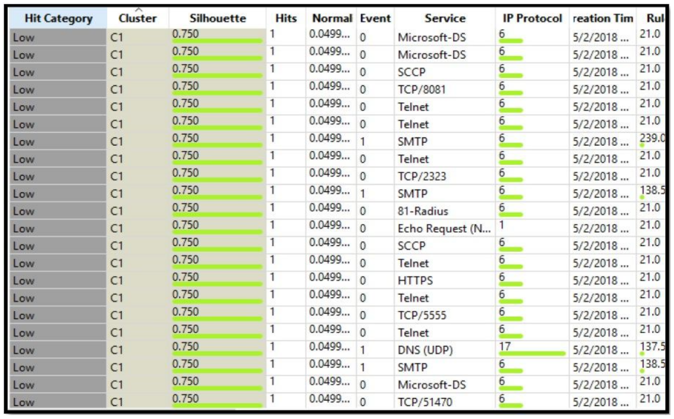
4.3 Campus Network Datasets profiling results

Figure 8 shows the campus network data traffic before the clustering. The data has hit category as low, medium (med) and high that refer to access towards the network resources.



Hit Category	Hits	Norma	Event	Service	Severity	P Protoco	Situation	reation Tim	Rule Tag
Low	44	40.89...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Low	20	18.09...	1	HTTP	Info	6	HTTP_Request-POST	5/1/2018 ...	209717
Low	10	8.599...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Low	2	0.999...	1	UDP/443	Info	17	Connection_Allowed	5/1/2018 ...	209734
Low	1	0.049...	1	Telnet	Low	6	TCP_Segment-SYN-...	5/1/2018 ...	209738
Low	1	0.049...	1	Telnet	Info	6	TCP_Segment-SYN-...	5/1/2018 ...	209738
Low	1	0.049...	1	Telnet	Low	6	TCP_Segment-SYN-...	5/1/2018 ...	209738
High	153	144.4...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Med	85	79.84...	1	HTTPS	Info	6	TLS_Client-Hello	5/1/2018 ...	209717
Med	85	79.84...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Low	30	27.59...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Low	4	2.899...	0	TCP/4899	Info	6	Connection_Discar...	5/1/2018 ...	2
Med	95	89.34...	1	NTP (UDP)	Info	17	Connection_Allowed	5/1/2018 ...	209734
Low	50	46.59...	1	NTP (UDP)	Info	17	Connection_Allowed	5/1/2018 ...	209734
Low	16	14.29...	1	HTTPS	Info	6	Connection_Allowed	5/1/2018 ...	209734
High	11678	1109...	1	DNS (UDP)	Info	17	DNS_Standard-Que...	5/1/2018 ...	209717
High	344	325.8...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
High	281	266.0...	1	HTTPS	Info	6	TLS_Client-Hello	5/1/2018 ...	209717
High	208	196.6...	1	HTTPS	Info	6	TLS_Client-Hello	5/1/2018 ...	209717
High	180	170.0...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
High	123	115.9...	1	HTTP	Info	6	HTTP_Request-GET	5/1/2018 ...	209717
Med	64	59.89...	1	HTTPS	Info	6	TLS_Client-Hello	5/1/2018 ...	209717

Figure 8: Campus Network traffic before clustering



Hit Category	Cluster	Silhouette	Hits	Normal	Event	Service	IP Protocol	reation Tim	Rul
Low	C1	0.750	1	0.0499...	0	Microsoft-DS	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Microsoft-DS	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	SCCP	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	TCP/8081	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	1	SMTP	6	5/2/2018 ...	239.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	TCP/2323	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	1	SMTP	6	5/2/2018 ...	138.5
Low	C1	0.750	1	0.0499...	0	81-Radius	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Echo Request (N...	1	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	SCCP	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	HTTPS	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	TCP/5555	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	Telnet	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	1	DNS (UDP)	17	5/2/2018 ...	137.5
Low	C1	0.750	1	0.0499...	1	SMTP	6	5/2/2018 ...	138.5
Low	C1	0.750	1	0.0499...	0	Microsoft-DS	6	5/2/2018 ...	21.0
Low	C1	0.750	1	0.0499...	0	TCP/51470	6	5/2/2018 ...	21.0

Figure 9: Campus Network traffic after clustering

Figure 9 is the network traffic data after clustering. Based on the figure it shows that, Orange tool able to the clustering by clustering the low hit traffic to cluster C1.

Figure 10 illustrates further clustering towards network services. Three clusters have been applied. Result illustrates that cluster 1 and cluster 2 shares the same services such as DNS(UDP). As for cluster 3, it only consists; MySQL and DNS(UDP).

4.4 Technical Recommendation

Technical recommendation is based on the Protocol type. The recommendation is based on the 2 days that the traffic data has been collected from the campus network. The two day were public holiday which was a Labor Day holiday.

A. Domain Name System (DNS) User Datagram Protocol (UDP)

DNS(UDP) at port 53, should be allowed. This recommendation is based on the records' results that shows one to one IP address communication that hits 54,879 high volumes in concurrent sessions in two (2) days that shows in cluster 3 out of 74,869 totals of the service hits.

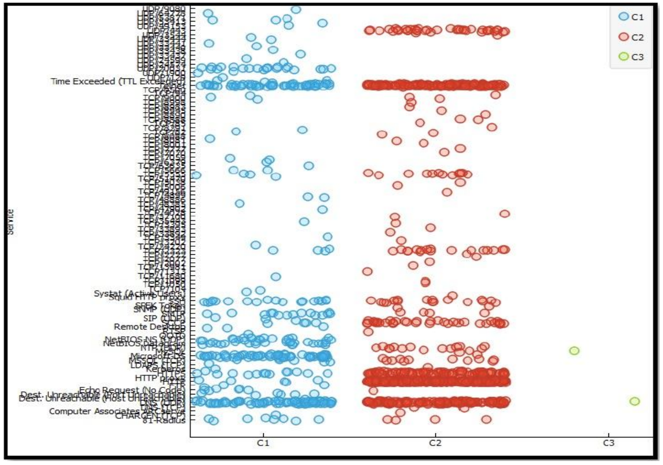


Figure 10: Campus Network traffic Clustering towards Network services

B. MySQL

MySQL at port 3306 should be terminated. An attempt has been done to this service. It has been hit by 40,658 hits. It is advisable to create a default rule that restrict the connection except for the authorized access only.

C. SSH

SSH is at port 22 should be terminated. Total attempt is 363 from different public IP. 112 record come from a single IP address. It is advisable to set an IP range for the connection and restricted for administrative job only.

D. HTTP

HTTP is at port 80. This port should be restricted from same source IP based on range connection. Results show that an attack has come from public IP address with 3191 hits.

E. NTP

NTP is at port 123. Attempts to update this server has been recorded (302 hits). It is advisable to allow only for local area network only.

F. Telnet

Telnet is at port 23 and it is not secure. It is advisable to terminate the connection.

5. CONCLUSION

This work has been successfully profiled a network traffic data on a campus network by using K-means algorithm. The profiling is based on clustering approach. As a conclusion, the increasing use of Internet among users has posed a challenge to campus network administrator in ensuring fair distribution of bandwidth and protecting the security of its users. Therefore, by having the capability to profile the internet traffic into different type will provide better visualization on the campus network usage. Future work would be comparing different type of clustering approaches on the network traffic data.

ACKNOWLEDGEMENT

This paper would like to acknowledge Industry, Community & Alumni Network Office of Fakulti Sains Komputer dan Matematik UiTM Shah Alam for providing us with the necessary fund, needed for this research. We also want thanks our fellow members of Computer Technology and Networking Department for giving us support and advice in completing this research.

REFERENCES

1. Malaysian Communications and Multimedia Commission, "Internet Users Survey 2017 Statistical Brief Number Twenty-One", 2017, ISSN 1823-2523.
2. Malaysian Communications and Multimedia Commission, "Internet Users Survey 2018 Statistical Brief Number Twenty-Three", 2018, ISSN 1823-2523.
3. Mizuno, S., Hatada, M., Mori, T., & Goto, S. **BotDetector: A robust and scalable approach toward detecting malware-infected devices**. In *Proc IEEE International Conference on Communications (ICC),* 2017 https://doi.org/10.1109/ICC.2017.7997372
4. Anwar, S., Zolkipli, M., & Zain, J. **Android Botnets: A Serious Threat to Android Devices.** *Pertanika Journal Of Science & Technology*, Vol 26 no 1, 2018.
5. Mohd Yusof, M., Mohd Ali, F., & Darus, M. **Classification Algorithm Against Different Types of DDoS Attacks Using Hybrid Approach**. *International Journal Of Innovations In Engineering And Technology (IJIET),* Vol 14 no 3, 2019.
6. Muhammad Zulfadhilah, Imam Riadi and Yudi Prayudi L**og Classification using K-Means Clustering for Identify Internet User Behaviors**. *International Journal of Computer Applications Vol* 154 no 3 pp34-39, 2016.
7. Purnawansyah and Haviluddin. **K-Means clustering implementation in network traffic activities**, in *Proc of International Conference on Computational Intelligence and Cybernetics*, Makassar, 2016, pp. 51-54. https://doi.org/10.1109/CyberneticsCom.2016.7892566
8. T. Mauritsius, A.S.Braza and Fransisca**, Bank Marketing Data Mining using CRISP-DM Approach.** *International Journal of Advanced Trends in Computer Science and Engineering IJATCSE*, vol.8 no 5, pp.2322 – 2329, 2019 https://doi.org/10.30534/ijatcse/2019/71852019
9. M.Syamala and N.J.Nalini, **A Deep Analysis on Aspect based Sentiment Text Classification Approaches**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE),* vol.8 no 5, pp. 1795-1801, 2019. https://doi.org/10.30534/ijatcse/2019/01852019
10. Muhammad Zulfadhilah,Yudi Prayudi and Imam Riadi, **Cyber Profiling using Log Analysis and K-Means Clustering**, *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 7, No. 7, 2016.
11. D. S. Terzi, R. Terzi and S. Sagiroglu, **Big data analytics for network anomaly detection from netflow data**, in *Proc International Conference on Computer Science and Engineering (UBMK)*, Antalya, pp. 592-597. 2017.
12. S. Kodati and Dr. R. Vivekanandam, **Analysis of Heart Disease using in Data Mining Tools Orange and Weka**,*Software & Data Engineering Global Journal of Computer Science and Technology*, Vol 18 no 1 Version 1.0, 2018.
13. S. Gole. **A survey of Big Data in social media using data mining techniques**, *in Proc IEEE Int. Conf. Advanced. Comput. Commun. Syst.*, pp. 5–10, 2015.
14. Tawarish, M., & Satyanarayana, K. **An enabling technique analysis in Data Mining for Stock Market trend by Approaching Genetic Algorithm**. *International Journal Of Advanced Trends In Computer Science And Engineering*, Vol 8 no 1, pp 27-33, 2019. https://doi.org/10.30534/ijatcse/2019/06812019
15. D. S. Terzi, R. Terzi and S. Sagiroglu, **Big data analytics for network anomaly detection from netflow data**, *International Conference on Computer Science and Engineering (UBMK*), Antalya, 2017, pp. 592-597.
16. Z. Abedjan, L. Golab and F. Naumann, **Data profiling,**In Proc *IEEE 32nd International Conference on Data Engineering (ICDE)*, Helsinki, 2016, pp. 1432-1435.
17. Ruan, Z., Miao, Y., Pan, L., Xiang, Y., & Zhang, J. (2018). **Big network traffic data visualization**. *Multimedia Tools and Applications*, Vol 77 no 9, pp 1459-11487. https://doi.org/10.1007/s11042-017-5495-y
18. Sasa M, Suleyman U Anel T, **Geographic Profiling for serial cybercrime investigation**", *International Journal of Digital Investigation*, Vol 28, pp 176-182, 2019.
19. Mucahid M.S, Ali Y, **Performance Analysis of ANN and Naïve Bayes Classification Algorithm for Data Classification**, *International journal of Intelligent System and Application in Engineering*, Vol 7 no 2, pp 88-9. 2019
20. Agrawal R. **Integrated Parallel K-Nearest Neighbor Algorithm**. In: *Satapathy S., Bhateja V., Das S. (eds) Smart Intelligent Computing and Applications. Smart Innovation, Systems and Technologies*, vol 104. Springer, Singapore, 2019
21. M. Yang and K. P. Sinaga, **A Feature-Reduction Multi-View k-Means Clustering Algorithm**, *IEEE Access*, vol. 7, pp. 114472-114486, 2019. https://doi.org/10.1109/ACCESS.2019.2934179
22. Gaël B, Tarn D, Mustapha L, Hanane A, Christophe C, **A Distribution Approximate thr Nearest Neighbors Algorithm for Efficient Large Scale mean Shift Clustering** , *International Journal of Parallel and Distributed Computing*, Vol 134, pp 128-139, 2019.
23. A.Marwan. (2019). Cisco Community, Campus Network design Guideline, https://community.cisco.com/t5/networking-documents/campus-network-design-guideline/ta-p/3140160
24. M. Karamollahi and C. Williamson, **Characterization of IMAPS Email Traffic**, *in Proc IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, Rennes, FR, 2019, pp. 214-220.

25. C. Hardegen, B. Pfülb, S. Rieger, A. Gepperth and S. Reißmann, **Flow-based Throughput Prediction using Deep Learning and Real-World Network Traffic**," In *Proc of 15th International Conference on Network and Service Management (CNSM),* Halifax, NS, Canada, 2019, pp. 1-9.
26. S.Oluwadarea and O.C.Agbonifo. **Network Traffic Analysis using Queuing Model and Regression Technique**, *Journal of Information*, vol 5, pp16-26, 2019. https://doi.org/10.18488/journal.104.2019.51.16.26
27. Analysis of Data Using Data Mining tool Orange (2017)
28. Iris Dataset, https://archive.ics.uci.edu/ml/datasets/Iris. Retrieved on 17th Feb. 2020.