IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLUSTERING OF INTRUSION DATASET FOR HELPING IN DISCOVERING HARMFUL PROFILE

[1]Zin May Zaw, [2]Nwe Ni San

[1]Department of Computer Engineering and Information Technology

[1]Mandalay Technological University, Mandalay, Myanmar

Abstract: **Nowadays, intrusion is a serious security threat in a network environment. To solve this security requirement, many researchers have argued that data mining can improve the performance of intrusion detection system. Among the datamining techniques, clustering is an important means for intrusion detection. The aim of clustering is to group the two observations into the same cluster if these observations close to each other in the original data space. Because of growing the size of computer networks and developed applications exponentially, the significant increasing of the potential damage that can be caused by launching attacks is becoming obvious. The most important defence tools against the sophisticated and ever-growing network attacks are Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) which can compromise the confidentiality, integrity or availability of resource.**

Index Terms **-** Intrusion Detection System; Cyber Security; Clustering; K-means clustering**.**

I. INTRODUCTION

A big challenge for today's network engineers and researchers is the identification of malevolent activities in a host, which eventually propagate to other hosts over a network. Untrusted programs which forcefully take part in this event of disaster and its popularity called intrusion [8]. Intrusion detection is an important technology not only in business sector as well as an active area of research because it is needed to detect attacks against a computer system. Because of the growth of the usage of computers over network and development in application running on various platforms, Intrusion Detection System (IDS) plays a vital role in detecting anomalies and attacks in the network [1].

The goal of intrusion detection systems is to detect malicious traffic. Therefore, it is needed to monitor all incoming and outgoing traffic. Various methods of detecting are misuse detection against anomaly detection, network-based against host-based systems, passive system against reactive system.

A. Signature Based Detection

This technique identifies and stores signatures of known intrusions and then matches the activities occurring on an information system to these signatures. In order to detect whether the system has been attacked, this requires constant updating of database.

B. Anomaly Based Detection

This technique assumes that an attack will always reflect some deviations from normal system activity. This type of IDS establishes a profile of system's normal activities and then compares activities on the information system to this normal behaviour. When there is a significant difference between the normal behaviour and the observed behaviour, the system signals an intrusion.

II. RELATED WORK

A new intrusion detection dataset was presented in [4]. Due to the lack of adequate dataset, anomaly-based approaches in intrusion detection systems are suffering from accurate deployment, analysis and evaluation. Some intrusion detection datasets such as DARPA98, KDD99, ISC2012, AND ADFA13 have been used in previous intrusion detection approaches. Because of the lack of reliability of previous dataset, the authors of [4] produced a reliable dataset that contains benign and seven common attack network flows, which meets real world criteria and is publicly available.

To identify the relevant, hidden data of interest for the user effectively and with less execution time, data mining concepts ware integrated in [1]. In that system, KDD 99 cup dataset was used. Firstly, the dataset was pre-process and different types of decision tree algorithms (C4.5 and its extensions) were used for the task of detecting intrusion. And then, comparison results for the performance of these algorithms were shown. The work in [1] proposed an intrusion detection system which detects four different types of attacks. The algorithm is trained with labelled KDD dataset and experimented with unlabelled dataset. For showing the results in more understandable and efficient way, Weka tool is used for graphical output.

III. CICIDS 2017 DATASET

CICIDS 2017 dataset contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). Table 1 show the capturing period which started at 9:00 on Monday, July 3rd and continuously ran for an exact duration of 5 days, ending at 17:00 on Friday July 7th. Attacks were subsequently executed during this period [8].

CICIDS 2017 dataset is intended for network security and intrusion detection purpose and it covered a diverse set of attack scenarios. In this dataset, six attack profiles were created based on the last updated list of common attack families and executed them by using related tools and codes.

Table 1.Data Tables in Dataset

Day of Data Collection	File Name	Type of Attacks
Monday	Monday- WorkingHours.pacp_ISCX.csv	Benign
Tuesday	Tuesday-WorkingHours.pcap_ISCX.csv	Benign, FTP-Patator, SSH-Patator
Wednesday	Wednesday-workingHours.pcap_ISCX.csv	Benign, DoS GoldenEye, DoS Hulk, DoS slwoloris, Heartbleed
Thursday	Thursday-WorkingHoursMorning-WebAttacks.pcap_ ISCX.csv	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS
Thursday	Thursday-WorkingHoursAfternoon-Infilteration.pcap_ ISCX.cs	Benign, Infiltration
Friday	Friday-WorkingHoursMorning.pcap_ISCX.csv	Benign, Bot
Friday	Friday-WorkingHours-AfternoonPortScan.pcap_ISCX.csv	Benign, PortScan
Friday	Friday-WorkingHours-AfternoonDDos.pcap_ISCX.csv	Benign, DDoS

A. Brute Force Attack

This is one of the most popular attacks that only cannot be used for password cracking, not also to discover hidden pages and content in a web application.

B. Heartbleed Attack

The "heart" is derived from the heartbeat protocol while the "bleed" indicates leakage of the data, hence the name "Heartbleed". The bug was present in the widely used OpenSSL library implementation for SSL and TLS protocols [3].

C. Botnet

A number of Internet-connected devices are used by a botnet owner to perform various tasks. It can be used to steal data, send spam and allow the attacker access to the device and its connection.

D. DOS Attack and DDOS Attack

A DoS is an attack which is launched to make network's and system's resources unavailable for the legitimate users so that no one else can access it. The main targets of these attacks are web servers, default gateways, personal computers, etc. DoS and its variant, DDoS are possible threats which exhaust the resources to make it unavailable for the legitimate users, thereby, violating one of the security components such as availability [7].

E. Web Attack

Web attacks are also known as internet attacks. Networks usually use different security systems just like intrusion detection system in order to handle the attacks [3].

IV. K-MEANS CLUSTERING ALGORITHM

K-means algorithm is the most commonly used partition clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time [6]. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells [10]. Given a set of observations $(x_1,x_2,..,x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$n) sets $S=\{S_1,S_2,\ldots,S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance) are defined in (1):

$$arg_s min \ \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu i \|^2 \quad (1)$$

where μ_i is the mean of points in S_i .

V. PROPOSED SYSTEM

The system design of proposed clustering on Intrusion Dataset for discovering harmful profile system is illustrated in Figure1.

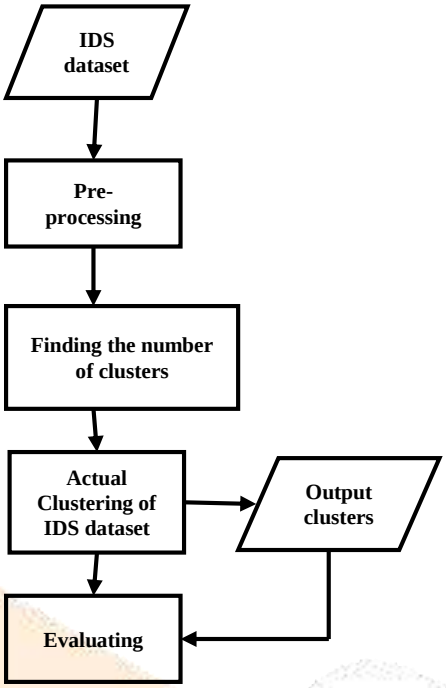

Fig. 1.Proposed Intrusion Dataset Clustering

A. *Dataset Used*

In this system, Thursday. WorkingHours. Morning. WebAttacks. pcap_ISCX data table in CICIDS 2017 (which can be free downloaded from [11]) dataset is used for doing experiment. The following table, Table 2 described the information about the data used for the experiment.

B. *Preprocessing*

The data set contains 170366 observations on 79 attributes including category attribute.

C. *Removing Zero Value Columns*

In the data table, there are attributes which contains only value zeros for every records. These attribute columns are removed as the pre-processing step. After removing these attributes, it remains 68 attributes.

D. *Removing Missing and Not Regular Value Records*

The dataset also contains the “NA” and “NaN” values for some records. For getting the good result, these records are also omitted in the pre-processing step of the system. After cleaning these records, 170231 records were found. In pre-processing steps, the data rows are cleaned not to include missing data and non-regular values. Initially the data set contains “” records. After cleaning the missing values and non-regular values, it remains “” records.

E. *Finding Number of Clusters (k) Using Elbow Method*

In k-means clustering, it is needed to know which is the best number of clusters of “k”.

The elbow method looks at the percentage of variance, explained as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn’t give much better modelling of the data. More, precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information, but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence, the elbow criterion. This elbow cannot always be unambiguously identified.

After cleaning the dataset, the next step is to find the number of clusters for the data by using the Elbow method.

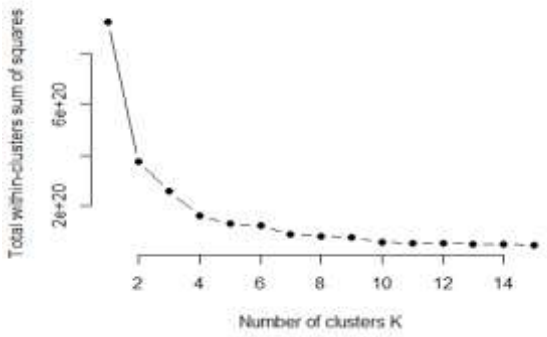

Fig. 2. Total Within Cluster of Sum of Squares

Figure 2 shows in order to find a good value for k, the within groups sum of squares for different values of k can be plot. This metric normally decreased as more groups are added. It is needed to find a point where the decrease in the within groups sum of squares starts decreasing slowly. Therefore, k=4 is the good choice of number of clusters for this dataset.

F. Creating Clusters on the IDS Dataset

When the number of clusters "k" is obtained, it is the step for clustering the dataset. In this experiment, 4 clusters are created for the dataset as shown in Table 3 and the resulted clusters are shown in Figure 3.

Table 2.About data used for experiment

Dataset Name	Dataset Type	No.of Attributes	No.of Observations
Thursday.WorkingHours.Morning.WebAttacks.pcap_ISCX	Multi class	79 (including label)	170366

Table 3. Size of each cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4
155	148012	7264	14800

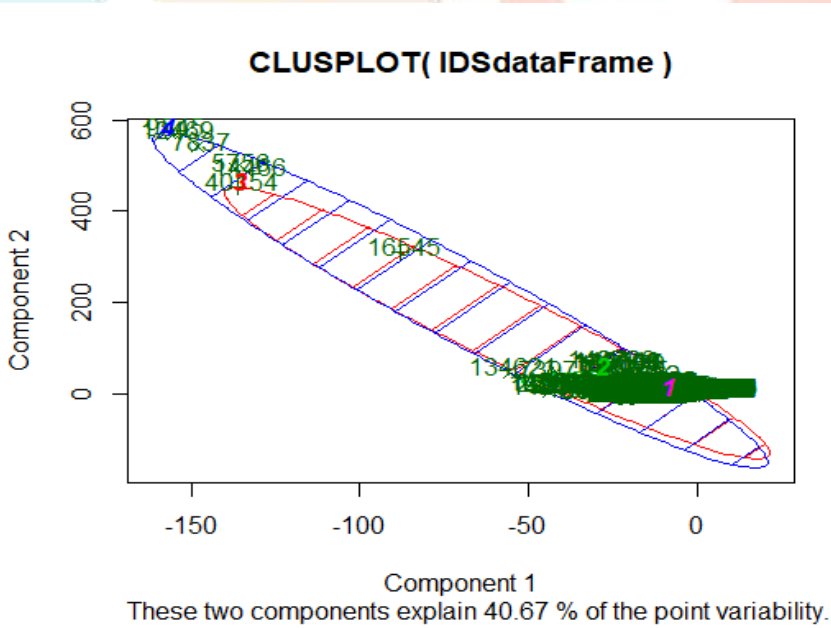

Fig. 3. Resulted Clusters

VI. Performance Evaluation

Purity is an external evaluation criterion for cluster quality: majority class and number of members of the majority class for the clusters. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N, number of observations, as shown in (2).

$$\text{Purity } (\Omega,C) = 1/N \sum_k \max_j |w_k \cap c_j| \quad (2)$$

where $\Omega=\{w_1,w_2,\ldots,w_k\}$ is the set of clusters and $C=\{c_1,c_2,\ldots c_j\}$ is the set of classes [2].

Table 4. Confussion matrix

Cluster	BENIGN	Web Attack Brute Force	Attack Sql Injection	Web Attack XSS
1	155	0	0	0
2	145917	1438	21	636
3	7179	69	0	16
4	14800	0	0	0

According to Table 4, the purity of clustering is (1/170231) x (155+145917+7179+14800), which indicates that the good clustering was obtained. The greater the values of purity indicate the better clustering [9].

VII. CONCLUSION

In this paper, clusters of intrusion are created by using K-means which is implemented in R. The resulting clusters are compared with the true label data set. Moreover, they are also compared with the values of purity calculated clusters which are generated by K-means clusters. The purity of the clusters is measured referencing to the class labels. The higher purity values mean the good cluster. The resulted clustering got the purity value indicates the good clustering result was obtained.

REFERENCES

[1] A. Sawan, et. al. "Intrusion Detection System using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Volume 4, Issue 2, February, 2015.

[2] C. D. Manning, "Introduction to Information Retrieval", Cambridge University Press, 2008.

[3] D. Kalaria, R.Joshi, "Exploiting the OpenSSL Heartbleed Vulnerability", DOI:10.13140/RG.2.1.1557.8489, March, 2019.

[4] I. Sharafaldin, et. al. "Towards Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018) pp 108-118.

[5] M. Arshad, Md. A. Hussain, "A Real-time LAN/WAN and Web Attack Prediction Framework Using Hybrid Machine Learning Model", International Journal of Engineering & Technology, 7(2018), 1128-1136.

[6] M. Deepa, P.Revathy, "Validation of Document Clustering based on Purity and Entropy Measures", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 3, May, 2012.

[7] N. Tripathi, B.M. Mehtre, "DoS and DDoS Attacks: Impact, Analysis and Countermeasures", proceeding of Conference on Advances in Computing, Networking and Security, 2013.

[8] R. Panigrahi, S. Borah, "A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems", International Journal of Engineering & Technology, 1, (3.24) pp 479-481, 2018.

[9] S.C. Sripada, "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering", Indian Journal of Computer Science and Engineering (IJCSE), ISSN- 0976-5166, Volume 2, No. 3, Jun-Jul, 2011.

[10]https://www.tutorialspoint.com/big_data_analytics/k_means_clustering.htm

[11] https://www.unb.ca/cic/datasets/ids-2017.html