Elimination of Network Intrusion Using Advance Data Mining Technology

Dhulfiqar Saad Jaafar[1](✉) and Hoshang Kolivand[2]

[1] Ministry of Education, Baghdad, Iraq
ghaffoori15@itu.edu.tr
[2] Department of Computer Science, Liverpool John Moores University, Liverpool L3 3AF, UK
h.kolivand@ljmu.ac.uk

Abstract. Advancements of data mining and machine learning have paved the road for establishing an efficient attack prediction paradigm to protect large scaled networks. In this study, computer network intrusions had been eliminated using smart machine learning algorithm to eliminate network intrusion. Referring a big dataset named KDD computer intrusion dataset which includes large number of connections that diagnosed with several types of attacks; the model is established for predicting the type of attack by learning through this data. Feed forward neural network model is outperformed over the other proposed clustering models in attack prediction accuracy.

Keywords: Network intrusion · Data mining · Machine learning · FFNN · KDD · K-means · DB Scan · Intrusion · Attack

1 Introduction

It has been realized from the literature survey that internet as a public network imposes plenty of challenges on data security. The solutions proposed in the literature involves using of virtual; private network in the standard format to secure the data over such public networks [1]. The advancement of software technology and electronic component manufacturing led to the invention of software defined network. Intrusion detection systems are used to identify irregularities in the flow of data from the host to the client or across the network [3]. Any malicious packet injection can be detected and deleted using an intrusion detection system, which can discover security problems on a network [2]. These systems work by examining the sender's and receiver's digital signatures and certificates. Any infractions of the rules are tracked and reported to the network administrator, who is alerted to the likelihood of a network assault. Packets are sometimes transmitted after decrypting and examining their contents for far more secure transmission [4]. Networks are using data technology such as machine learning algorithms for detection the attack and hence feedback signal can be transmitted to the firewall for blocking the suspicious inward request discovered by the machine learning algorithm.

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 155–161, 2022.
https://doi.org/10.1007/978-3-031-04409-0_15

2 Previous Work

One of the big issues that has been realized from the current approaches of virtual private network is that: using the same paradigm cannot be used to prevent all types of attacks and to secure whole web application. More likely, this is done applying the ordinary virtual private network that securing the two connections through the internet in personal computer into the data security in companies or institution (campus) networks. From the other hand, it was realized that virtual private networks protocols are sensitive for particular attacks or malwares whereas the other malwares or attacks are remained untraceable in virtual private network [5]. Some other tools are proposed for protection purposes by prevention the foreign requests unless it is pre-defined in the routing data. The same can stand with the case of data flooding applications such as ecommerce and social applications. Intermediate gates and software could be used to prevent the malfunctioning requests, the same is realized having negative impact on the network performance from the throughput, time delay, packets drop rates point of views. The main task of virtual private network is imposing the security over the network while the other aspects alike network performance guaranties are not gain their priority level in the research [6]. The number of nodes can be varied from two nodes and can be extended further according to the networking constrains. The security of data is remained disputed even over the virtual private networks which make the last (virtual private network) unable to face this security threat. The term virtual is standing for assumable (imaginary) connections that made as a part of plenty of connections involved in a real (physical) network [9, 11]. Firewalls and other software defined networks are tended to enhance the security and to ensure more privacy to the users' data.

3 Method

3.1 Dataset Description

Our internet network datasets were sourced from the UCI machine learning repository [8]. Intrusion is a famous challenge that can access any network especially those with large number of users such as internet network. Over the internet there are tons of the software defined networks which are linking a plenty of users from different locality in the globe. However, users are getting access into the internet network through smaller networks that acts as relay. So-to-say, routing log data was obtained from open-source database inventory where one thousand inward web requests were recorder. Each request is representing an incoming web connection for a particular network operating over the internet [10, 11].

Each connection is representing a type of attack that aimed for network penetration. The dataset is made by monitoring some features for each connection aiming to classify the said connection according to the type of attack it represents. The last column of the dataset is represented as the target or the classes that made according to understand each attack features.

3.2 Pre-processing Model

The data cells on the dataset are being observed in order to prepare it for the so-called pre-processing. The data is included of large number of entries which represents the connection properties of large number of users (one thousand user). Each user is demonstrated by set of connection features as stated above. The challenges that might be addressed through the pre-processing model can be enlisted as following:

a> data might be included with missing cells that are raised due to several aspects including the storage of bad sector which lead to damage some data and hence it looks alike missing data. All the missing data are usually substituted with question marks, hashes, or even terms alike "missing".

b> missing data are to be recovered using the method of column values averaging which means each column with missing data is to be treated by adding all the columns values together and dividing the result of the summation by the number of elements situated in the column except those which said as missing. Equation 1 demonstrates the missing values substitutionary values; Algorithm 1 shows the model structure of missing values recovery.

Let column "A" witnessing three missing values that denoted as question marks in their particular cells;

Let "S" is the summation of total column elements;

Let "N" is the total count number of column elements;

Let "M" is the total number counts the missing values.

Hence, the missing value will be retained by fulfilling the Equation 1.

$$R(A) = \frac{\sum_{N=1}^{N=x} A(N)}{X} = \frac{S}{X} \tag{1}$$

Algorithm 1: Missing values evaluation and substitution

```
Select the column (Ci), L=Len(A), i=0
While! (A(i)∈[#, &,?, !])
        A(i)=Aver(A)
        i++
   End while
```

From the other hand, data normalization is the last step that was conducted during the per-processing task. The dataset after the decoding will appear with codes stated where no other unknown alphabets or numbers is longer existed. At the end of the encoding stage, data will be ready to be in warded into the machine learning paradigms. However, one more step can be performed for enhancing the performance of training by normalizing the data. Data in each cell over the dataset can be normalized by fining the peak value of every column and then dividing the entire column elements by that value. The same is yielding new appearance of the data which has value of "one" as the maximum value (new peak) and all other values are divisions of this new peak. However, the process of normalization is made to enhance the training performance by minimizing the variance of the overall values in dataset.

4 Intrusion Sensing

Machine learning as well as deep learning approaches were used to perform data clustering tasks in order to predict the intrusion attack. In the first stage of processing, the dataset is made ready for this process as the data can be feed into the algorithm and the algorithm can decided wither this connection is an attack or not. However, four type of classes are being associated with the test dataset as listed in Table 1.

Table 1. Target classes illustrating the number of classes the target vector

Class name	Class code
normal	0
snmpgetattack	1
Xlock	2
Smurf	3

Figure 1 demonstrates the role of the clustering technology performed by both machine learning and deep learning approaches for detecting of the attack.

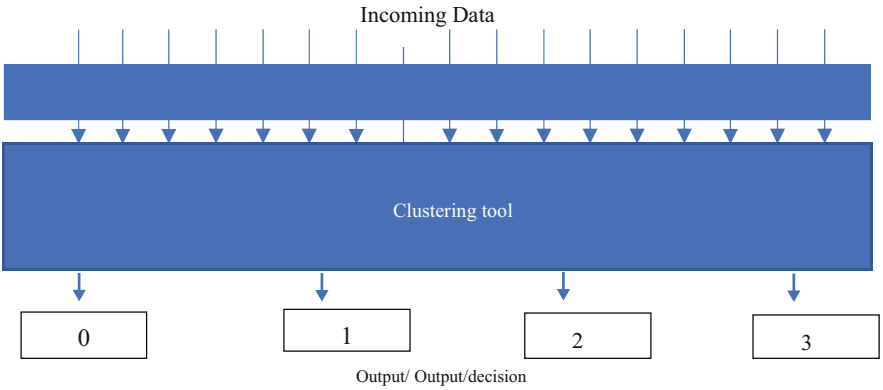

Fig. 1. Decision making in clustering method (general representation).

In order to safe safeguard the network against any malicious attack, a Feed Forward neural Network is utilized to develop a smart attack prevention paradigm. This paradigm may predict occurrence of attack depending of the attitude of each attack before it is actually taking place. Model employing the feed forward neural network for predicting the malicious activities [11, 12].

In order to use this model, the feed forward neural network is being trained using a dataset of network attacks attitudes. Data set is included with large number of connections, those connections included with attack (malicious connection) as well as safe connection [14]. Every connection was diagnosed and accordingly the target column is made to classify the data according to the nature of connection. The attack prevention model is working according to the Algorithm 2.

Algorithm 2: Attack prevention model

Step 1: *network attacks dataset is downloaded from open access data bank and used in the further steps of the system.*

Step 2: *dataset is pre-processed in order to convert any alphabetic entry into numerical entry. From the other hand, all the values (numbers) in the dataset are being normalized in order to reduce the variance between the data cells which may enhance the performance of model training in hereinafter.*

Step 3: *there was no missing values in the dataset entries so-no missing value recovery program was made.*

Step 4: *Feed Forward Neural Network model is used for implementing the model of attack prevention. A prediction process is made firstly by letting the model training using eighty percent of the data.*

Step 5: *after successful training of the model, model is tested using the remained twenty percent of the dataset.*

The performance of the feed forward neural network model is compared with baseline clustering technologies namely: Agglomerative clustering algorithm, DB scan clustering algorithm, K-means clustering and Mini batch K-means clustering. All the mentioned algorithms are made to produce the decision of attack type and performs the operations that similar to the process done by feed forward neural network [15, 16]. FFNN model is first initiated by preforming of fifty iteration with random wright allotment (ordinary FFNN model). Hence after, model is optimized using the weight freezing method which involved selecting the optimal weight coefficients out of the iterated weight in the ordinary model to be the permanent weight of the model. The Freeze FFNN model is eventually observed to be outperformed over the other clustering techniques.

5 Results and Discussion

Each clustering algorithm is made to perform the task of attack clustering so that the attack can be predicted foreach incoming web request before it is actually taking place. The performance of each algorithm is determined using three performance metrics namely: mean absolute error, mean square error and accuracy of clustering. However, the Table 2 below illustrates the performance metrics for each algorithm.

Table 2. Performance metric for the attack clustering

Clustering	MAE	MSE	Accuracy
Agglomerative clustering	1.8653	3.7714	51
DB scan	2.1395	7.7128	13.3
K-means	1.9666	4.2407	46
Mini batch	1.8897	3.8897	50.1
FFNN (Freeze model)	0.9172	2.6142	98

6 Conclusion

Networks are highly susceptible for malware and malicious attacks since it is operating over a public network like internet. The cost of attack prevention machinery and network components is far expensive than network itself in some applications. Hence, the current trends of network protection alike virtual private networks are also can't stand to face the fluctuation on the attacks which are developed a lot due to the software and computer advancement in the current research. In this project, a smart attack prediction approach is proposed in which predicting the attack by learning through the behaviors of each incoming connection. However, big dataset issues which involves a thousand connections, each connection is being featured by several properties such as connection duration, service protocol, number of errors, etc. two approaches of learning are used namely deep learning approach (e.g., Feed Forward Neural Network) and machine learning clustering approaches' (e.g., Agglomerative clustering algorithm, DB scan clustering algorithm, K-means clustering and Mini batch K-means clustering). Results shown that Feed Forward neural network has outperformed for attack clustering and attack prediction with prediction performance of 98 percent. On the other hand, the second-best prediction performance is realized on Agglomerative clustering followed by mini batch, K-means and DB Scan respectively. The other performance metrics are also revealed the same point that is feed forward neural network is the optimum intrusion prediction algorithm.

References

1. Habibzadeh, H., et al.: A survey on cybersecurity, data privacy, and policy issues in cyber-physical system deployments in smart cities. Sustain. Cities. Soc. **50**, 101660 (2019)
2. Clarke, N., Li, F., Furnell, S.: A novel privacy preserving user identification approach for network traffic. Comput. Secur. **70**, 335–350 (2017)
3. Haripriya, L., Jabbar, M.A.: Role of machine learning in intrusion detection system. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 925–929. IEEE (2018)
4. Azwar, H., Murtaz, M., Siddique, M., Rehman, S.: Intrusion detection in secure network for cybersecurity (2018)
5. Panetta, K.: Gartner's Top 10 Security Predictions 2016. Сайт компанії «Gartner» (2016). http://www.gartner.com/smarterwithgartner/top-10security-predictions2016

6. Arshad, J., et al.: A review of performance, energy and privacy of intrusion detection systems for IoT. Electronics **9**(4), 629 (2020)
7. Ande, Ruth, et al. "Internet of Things: Evolution and technologies from a security perspective." Sustainable Cities and Society 54 (2020): 101728
8. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases. University of California, vol. 55. Department of Information and Computer Science, Irvine (1998). http://www.ics.uci.edu/?mlearn/MLRespository.html
9. Riahi, A., Challal, Y., Natalizio, E., Chtourou, Z., Bouabdallah, A.: A systemic approach for IoT security. In: Proceedings of the 2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), Cambridge, MA, USA, 20–23 May 2013
10. Jesus Pacheco, S.H.P.: IoT security framework for smart cyber infrastructures. In: Proceedings of the IEEE International Workshops on Foundations and Applications of Self* Systems, Augsburg, Germany, 12–16 September 2016
11. Dorri, A., et al.: Blockchain for IoT security and privacy: the case study of a smart home. In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE (2017)
12. Yao, X., et al.: A lightweight multicast authentication mechanism for small scale IoT applications. IEEE Sens. J. **13**(10), 3693–3701 (2013)
13. Mamdouh, M., Elrukhsi, M.A., Khattab, A.: Securing the internet of things and wireless sensor networks via machine learning: a survey. In: 2018 International Conference on Computer and Applications (ICCA), pp. 215–218. IEEE (2018)
14. Debar, H., Dacier, M., Wespi, A.: Towards a taxonomy of intrusion-detection systems. Comput. Netw. **31**(8), 805–822 (1999)
15. Meng, G., Liu, Y., Zhang, J., Pokluda, A., Boutaba, R.: Collaborative security: a survey and taxonomy. ACM Comput. Surv. **48**, 1:1-1:42 (2015)
16. Zaidan, M.R.: Power system fault detection, classification and clearance by artificial neural network controller. In: Global Conference for Advancement in Technology (GCAT), Bangalore (2019)