An Application of Artificial Neural Network Based Intrusion Detection System

Salih Erdem Erol[1], **Recep Benzer**[2*]

1 Gazi University, **2** National Defense University , ***** Corresponding author, rbenzer@kho.edu.tr

Abstract

Information systems are one of the areas where public and private sector invested the most in recent years. Almost every area of life, including critical infrastructure systems (electricity, water, telecommunications, banking, etc.), is managed by information systems. These developments provide an environment for the rapid increase of cyber-attacks and their application in very different ways. In this study, Intrusion Detection Systems, one of the basic elements for information security, will be evaluated and the results obtained from a sample ANN based IDS application will be analyzed.

Keywords: Information security, Intrusion detection system, ANN, Machine learning.

Citation: Erol, S. E., Benzer, R. (2018, October) *An Application of Artificial Neural Network Based Intrusion Detection System*. Paper presented at the Fifth International Management Information Systems Conference.

Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

IMISC 2018 PAPER

An Application Of Artificial Neural Network Based Intrusion Detection System

Abstract

Information systems are one of the areas where public and private sector invested the most in recent years. Almost every area of life, including critical infrastructure systems (electricity, water, telecommunications, banking, etc.), is managed by information systems. These developments provide an environment for the rapid increase of cyber-attacks and their application in very different ways. In this study, Intrusion Detection Systems, one of the basic elements for information security, will be evaluated and the results obtained from a sample ANN based IDS application will be analyzed.

Keywords Information Security, Intrusion Detection System, ANN, Machine learning

1. Introduction

In today's world Information Technolohy systems are being used in every area such as education, trade, health, communication and public services. And so it is seriously effecting life by making crucial changes in the area that is used.

It is seen that Information Technology systems having much more importance day by day. One of the main reasons is that systems are interconnecting to each other with increasing numbers. Interconnection makes life easier but at the same time it opens a big door to attackers. The sharing, remaining ongoing openness and transmitting of information from the sender to the recipient in a confidential and unified manner without being corrupted, destroyed, altered and seized by others are the basic criteria for ensuring the security of information. In terms of the systems regulating the public life, the importance of these criteria is much clearer.

In this study, firstly, the study on the cyber attack life cycle and network structure was investigated and the studies on the literature about the intrusion detection systems afterwards were examined. In the last part, an application will be performed by using MATLAB Pattern Recognition Tool and evaluations will be made according to the results obtained.

2. Cyber Attack Life Cycle And Basic Network Structure

Cyber attacks changed the shell by increasing importance and quantity of information assets that systems possess. Foreign States, Terrorist Groups, Industrial Cyber Spies and Organized Cybercrime, Hacktivist, Hackers are sources of attack [1]. The difficulty of accessing the system also requires that attackers follow a systematic method. In this context, a 6-stage attack is emerging and implemented as one of the most accurate classifications that define systematic approaches to cyber attacks [2]. The steps of cyber attack life cycle is shown in Figure 1.

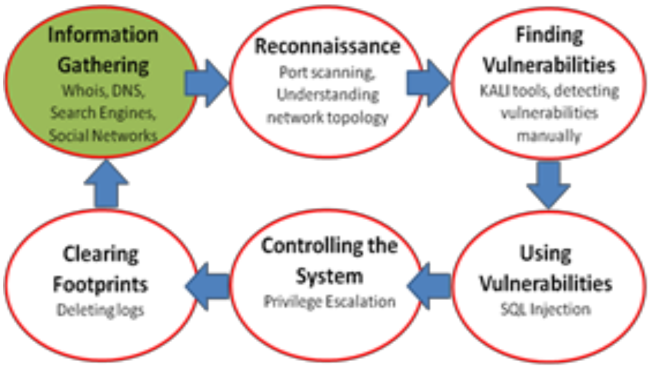
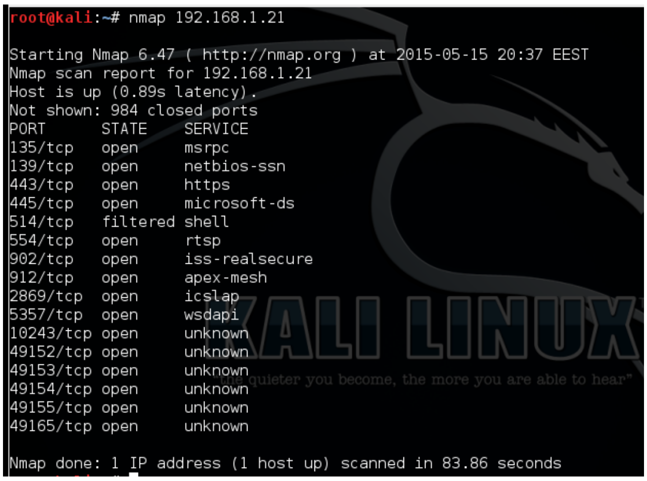


Figure 1. Cyber Attack Life Cycle

The first two steps of the cyber attack life cycle are often intertwined with information gathering and discovery. Attackers use Host Sweep and port scanning methods to infiltrate systems. While hosts are located in a network With TCP Echo, UDP Echo, ICMP Sweep, port scan attacks allow open ports to be identified and to attack using services [3]. With open source tools, port scanning is really easy to apply.



```
root@kali:~# nmap 192.168.1.21

Starting Nmap 6.47 ( http://nmap.org ) at 2015-05-15 20:37 EEST
Nmap scan report for 192.168.1.21
Host is up (0.89s latency).
Not shown: 984 closed ports
PORT      STATE    SERVICE
135/tcp   open     msrpc
139/tcp   open     netbios-ssn
443/tcp   open     https
445/tcp   open     microsoft-ds
514/tcp   filtered shell
554/tcp   open     rtsp
902/tcp   open     iss-realsecure
912/tcp   open     apex-mesh
2869/tcp  open     icslap
5357/tcp  open     wsdapi
10243/tcp open     unknown
49152/tcp open     unknown
49153/tcp open     unknown
49154/tcp open     unknown
49155/tcp open     unknown
49165/tcp open     unknown

Nmap done: 1 IP address (1 host up) scanned in 83.86 seconds
```

Figure 2: NMAP Port Scanning Screen

Three different routes are used in port screening attacks. Browsing all ports of a host is the most commonly used method. Besides, certain ports of different hosts can be scanned, which is used to decide the service to be attacked. The third method is to scan to different ports of different hosts. The NMAP shown in Figure 2 allows all the techniques to be done with a single command.

In the step of scanning for weaknesses, applications are made to find the vulnerability of the system. Vulnerability is defined as weaknesses that can be initiated unintentionally or accidentally or can be intentionally abused [4] or "any component of system that makes it unprotected against an existence threat (lack of security policies and procedures, lack of implementation or not, incomplete or faulty system design and implementation, organizational structure, reasons attributable to managers and employees' knowledge and attitude) [5].

The last step of an attack is the deletion of footprints in the system. An attacker who succeeds to infiltrate a network will want to delete footprints that he or she has left on the system to prevent to be accessed to himself or herself when the damage in the system is discovered.

As well as external attacks, all of these attack steps can be applied to internal attacks, which are defined as insider attacks. It was observed that while some of the insider attacks originated from unsuspecting innocent users who were infected with computer malware, 80% of malicious attacks were made by the technical staff in charge of the system. Investigations show that insider threats are seen more often than external threats [6, 7].

3. An Overview of Intrusion Detection Systems

IDS (Intrusion Detection System) came out as one of the solutions for the protection of the systems against various, rapidly shifting many kinds of attacks. IDS which can be defined as the software and hardware detecting attacks on a system, is divided into NIDS and HIDS. HIDS aims to protect a certain system or network, while NIDS aims to protect a complex network set up by servers or systems [12].

The first model for IDS has been developed by Doroty Derring and many models by other developers offered until today [8].

NIDS use two different methods. First, detect network anomalies by protocol standards. At the time system obtains that protocols don't work properly, intrusion is detected. The second method is about unusual usage of network traffic. For example, if an individual using internet from 15:00 to 17:00 every day but detected in internet after 02:00 am at the middle of night, an attack may be considered. Hybrid methods should be preferred in order to benefit from the advantages of both modes [9].

4. Application of Ann Based Ids

Generally an IDS can be created in 5 steps. In the first step with using packet capture tools packets in the network should be captured. After that packet data that captured should be analyzed and prepared to educate ANN. With data set prepared previous steps used to educate ANN model for the classification of packets. In the last step monitorship done to see results [3].

Different machine learning techniques can be used for training. While designing an IDS rule based approach or machine learning tehcniques such as artificial neural network, data mining, fuzzy logic, Bayesian Networks can be used. Artificial Neural Network (ANN) is a smart classification technique which started to be use for intrusion detection systems. ANN is one of the alternative techniques instead of statistical methods for intrusion detection systems which detect abnormal actions [11].

In literature it is seen that many studies implemented by using DARPA's KDD Cup data set. KDD Cup produced by MIT university and the data set is open to everbody. The data set has criticized mainly in two title, it is an old one and it contains lots of duplicated data rows [14,15]. Because of that reasons in this study a new data set which was produced by New Brunswik University's Information Security Centre of Excellence (ISCX). It is not open to everybody as KDD Cup, but ISCX submit data set to demanders for academic purposes with special permission.

ISCX data set was created by capturing packets for 7 days with a capturing system prepared by them. The data set has 19 data columns. A column contains data such as application name, protocol type, and source port. The 19th data column shows whether the packet row is attack or normal. There are 2.071.653 rows of data in ISCX data set. All data which columns contain is shown in Table 1. Data were submited in xml format and to study and index data database management program such as MYSQL is required.

Data set has 16 different xml packets for 7 days capture. All these packets were opened with MYSQL to integrate as one packet and analyse. The column that shows whether data row is normal or attack is detached from data set and saved as another data set. In detached data set all normal and attack labels were indexed with 0 and 1. After that in main data set all strings were indexed to be able to use in ANN. To do that 176 labels were firstly indexed in a table and according to that list update sentences were written in MYSQL to update all data. After update completed data set has integer values instead of strings. Table 2 shows sample data of protocol name labels after update ("1" instead of "TCP", "2" instead of "UDP" and "3" instead of "ICMP").

It is evaluated that indexing data and detaching data which do not contribute positively to results provides better performance and more success rate. In this study columns shown in Table 3 were detached.

Table 1: ISCX 2012 Data Labels

Column Name
AppName

Table 2: Sample Indexing for Protocol Name Labels

1	TCP
2	UDP
3	ICMP

Table 3: Detached Data Labels

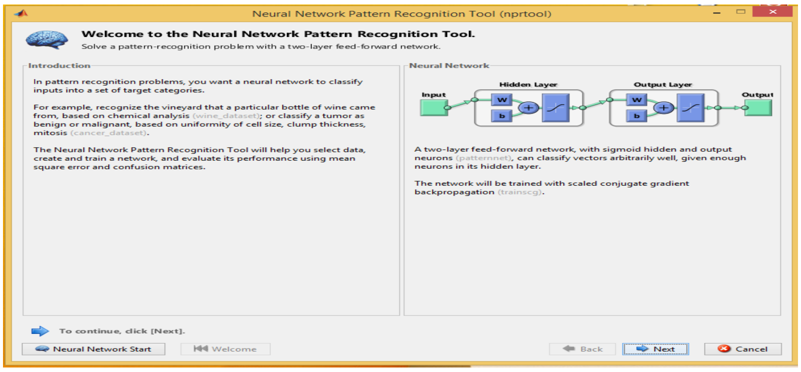
Detached Datas Labels
sourcePayloadAsBase64

totalSourceBytes
totalDestinationBytes
totalDestinationPackets
totalSourcePackets
sourcePayloadAsBase64
destinationPayloadAsBase64
sourcePayloadAsUTF
destinationPayloadAsUTF
Direction
sourceTCPFlagsDescription
destinationTCPFlagsDescription
Source
protocolName
sourcePort
Destination
destinationPort
startDateTime
stopDateTime
Tag

Reasons for these columns to be detached from data set can be summarize in 3 main title;

- Do not contribute positively to the results and classification,
- Not able to be indexed because of meaningless contex and/or
- Distrubition of data is not normal.

After data set prepared the following step is implementing machine learning proccess. Training for machine learning can be done by using different machine learning techniques. In this study to decide whether a packet is an attack or normal an ANN designed and implemented. To create an ANN MATLAB Neural Net Pattern Recognition Tool (NPRTool) was used [13]. NPRTool contains 1 input, 1 hidden layer and 1 output layer. Neuron numbers in layers is decided by designer. To decide how many neuron will give the best result, many implementation should be done with different numbers of neuron and results should be compared.


Neural Network Pattern Recognition Tool (nprtool)

Welcome to the Neural Network Pattern Recognition Tool.

Solve a pattern-recognition problem with a two-layer feed-forward network.

Introduction

In pattern recognition problems, you want a neural network to classify inputs into a set of target categories.

For example, recognize the vineyard that a particular bottle of wine came from, based on chemical analysis (wine_dataset); or classify a tumor as benign or malignant, based on uniformity of cell size, clump thickness, mitosis (cancer_dataset).

The Neural Network Pattern Recognition Tool will help you select data, create and train a network, and evaluate its performance using mean square error and confusion matrices.

Neural Network

A two-layer feed-forward network, with sigmoid hidden and output neurons (patternnet), can classify vectors arbitrarily well, given enough neurons in its hidden layer.

The network will be trained with scaled conjugate gradient backpropagation (trainscg).

To continue, click [Next].

Figure 4: NPRTool Screen View

In this study the best result was gained with 13 input neurons, 10 hidden layers and 1 output neuron. Created ANN is shown in Figure 5.

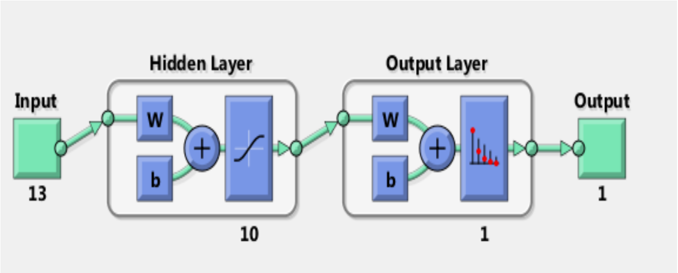
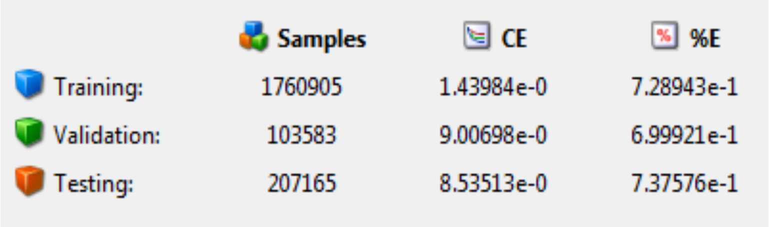


Figure 5: Created ANN

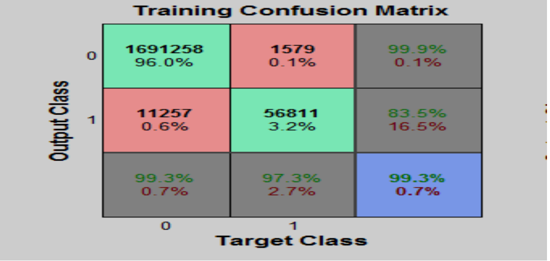
In training process %85 percent of data was detached for training, %5 percent of data detached for validation and %10 percent of data detached for test. Data sample numbers used in training, validation, test processes and error rates with training results is shown in Figure 6.



	Samples	CE	%E
Training:	1760905	1.43984e-0	7.28943e-1
Validation:	103583	9.00698e-0	6.99921e-1
Testing:	207165	8.53513e-0	7.37576e-1

Figure 6: Numbers of Samples Used for Training, Validation, Test with Error Rates and Training Results

After training with 1.760.905 network packets classification is resulted %99.3 percent of accuracy rate. Results are shown in Figure 7.



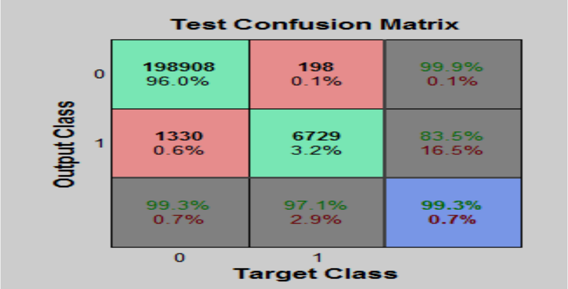
Training Confusion Matrix

Output Class	0	1	
0	1691258 96.0%	1579 0.1%	99.9% 0.1%
1	11257 0.6%	56811 3.2%	83.5% 16.5%
	99.3% 0.7%	97.3% 2.7%	99.3% 0.7%

Target Class

Figure 7: Training Results

Test process was implemented with 207.165 network packets which was detached for test and not used in training process, classification resulted with %99.3 percent of accuracy rate. Only 198 attacks could not detected and only in 1330 packets false alarm produced. Detailed results belonged to test process is shown in Figure 8. In literature it is told that in ANN based intrusion detection systems usuallay false pozitive rate is high. Also in this study we get similar results.



Test Confusion Matrix

Output Class	0	1	
0	198908 96.0%	198 0.1%	99.9% 0.1%
1	1330 0.6%	6729 3.2%	83.5% 16.5%
	99.3% 0.7%	97.1% 2.9%	99.3% 0.7%

Target Class

Figure 8: Test Results

%99.3 percent accuracy rate can be though as it is enough for security, but it is not a true inference for today's world. Small mistakes cost big losses to organizations and people. Because of this reason error rates should be even more less and accuracy rate should be much higher. To get better accuracy rate in IDS which are designed with ANN, 3 main points were shown in this study;

- Duplicated data should be detached
- To decide neuron numbers to get best result in NPRTool too many tries should be performed
- Data columns which do not contribute to classification results or which are not distrubuted normally should be detached from data sets (6 columns were detached) .

5. CONCLUSION

People can reach services easier, faster and cheaper by using information systems effectively in every area of life. However, this allows attackers to reach much more information. To block huge

numbers of attacks, to avoid malwares, evolving and increasing attack types, which targets users and organizations information assets, having smart systems becomes an obligatory. In this study, we proposed an intrusion detection system that based on artificial neural network (ANN) with a new dataset according to DARPA’s KDD Cup.

In this study 6 data columns which do not contribute to the results or distrubition of data do not appropriate were eleminated. Simplificated data set used to educate ANN model and results were analyzed again. Classification accuracy of packets resulted with better success rate.

According to results of this study it is evaluated that well classified and analyzed data set which contains more sample packets will improve the success rate. But at the same time getting more samples the need of the system requirements such as ram, GPU is also reveals.

Next study should be focus on analyzing of data set deeper which used in this study and implementing ANN model created in this study to a real time system to see the results.

KAYNAKLAR

[1] United States Computer Emergency Readiness Team “Control Systems Security Program (CSSP)” http://www.us-cert.gov/control_systems/csthreats.html (2011).

[2] Yiğit, T., Akyıldız, M., A., “Sızma Testleri İçin Bir Model Ağ Üzerinde Siber Saldırı Senaryolarının Değerlendirilmesi”, 2014.

[3] Al-Jarrah, O., Arafat, A., “Network Intrusion Detection System Using Attack Behavior Classification”, 2014, 5th International Conference on Information and Communication Systems (ICICS).

[4] Stoneburner, G., Goguen, A., Feringa, A., “Risk Management Guide for Information Technology Systems”,”Recommendations of the National Institute of Standards and Technology”, Special Publication 800-30, July 2002.

[5] Özbilen, A., “TCP / IP Tabanlı Dağıtık Endüstriyel Denetim Sistemlerinde Güvenlik ve Çözüm Önerileri”, Ankara(2012).

[6] Garuba, M., Liu, C., Fraites, D., “Intrusion Techniques: Comparative Study of Network Intrusion Detection Systems”, 5th International Conference on Information Technology: New Generations, 2008.

[7] Uma, M., Padmavathi, G., “A Survey on Various Cyber Attacks and Their Classification”, International Journal of Network Security, Vol.15, No.5, pg.390-396, Sept. 2013.

[8] Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., Pang-Ning, T., “Data Mining for Network Intrusion Detection”, 2002.

[9] Day, C., Intrusion Prevention and Detection Systems, Computer and Information Security Handbook. DOI:Chapter 26, 2013.

[10] Intrusion Prevention and Detection Systems, Computer and Information Security Handbook (Second Edition), 2013, pg. 485-498.

[11] Cannady, J., “Artificial neural networks for misuse detection”, Proceedings of the 1998 National Information Systems Security Conference (NISSC'98), Arlington, VA, pg. 443-456, 1998.

[12] Przemyslaw, K., Piotr, D., “Intrusion Detection Systems (IDS)” Part 2, Classification, Methods and Techniques. IT FAQ 2003; 1(4) pg. 17-22.

[13] Howard, D., Mark, B., Neural Network Toolbox for Use with MATLAB, User’s Guide Version 3.0, http://home.utad.pt/~psal/Mestrado/ficheiros/nnet.pdf.

[14] McHugh, J., 2000. The 1998 Lincoln Laboratory IDS Evaluation, in: Proceedings of the Third International Workshop on Recent Advances in Intrusion Detection (RAID), Springer-Verlag, London, UK. pg. 145–161.

[15] Brown, C., Cowperthwaite, A., Hijazi, A., Somayaji, A., 2009. Analysis of the 1999 DARPA/Lincoln Laboratory IDS evaluation data with netadhict, in: Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, IEEE Press, Piscataway,NJ, USA. pg. 67–73.