IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

Predicting Intruders in DARPA Data Set Using Neural Networks Method

[1] G.Yedukondalu, [2] Dr.J.Anand Chandulal, [3] Dr.M. Srinivasa Rao
[1] Vignan Institute of Technology& Science, Vignan Hills,Deshmukhi, Hyderabad, INDIA
[2] K.L.University, Vijayawada, Andhra Pradesh, INDIA
[3] School of Information Technology, Jawaharlal Technological University Hyderabad, INDIA

Abstract: - **Given a sequence of system calls, we want to predict whether that sequence of system calls is a normal or abnormal action. To do this, we chose traditional machine learning classification algorithm Feedforward Error Back Propagation Algorithm from Python Sci-Kit library and applied on classified data set. In these classified data set each feature is defied as a system call and the feature value is the frequency of that system call.**

Keywords: - Feed forward Error Back Propagation Algorithm; frequency of system call; python sci-kit library; sequence of system calls

I. INTRODUCTION

The Intrusion Detection System monitors the network traffic for malicious activities and issues alerts when such activities are identified. The Intrusion Detection Systems are broadly classified into two types. They are Network Intrusion Detection Systems (NIDS) and Host-Based Intrusion Detection Systems (HIDS). A HIDS that monitors important files of operating system, whereas, NIDS analyses to and pro traffic from the specific computer on which the malicious detection software is installed. Dr.Sanjay Rawat et-al assured that various processes under Unix system are converted into vectors using the frequencies or occurrence of various system calls invoked by those processes under normal condition[1]. A host-based intrusion detection system has the ability to monitor key system files any attempt to overwrite these files whereas Network Intrusion Detection System that analyses incoming network traffic. Harvinder Pal et-al stated that the prior features, intrusions on the system can be detected without any previous learning [5].

II. DATA DESCRIPTION

The data available in the DARPA data set is in the form of sequence of system calls. Each sequence of system calls is process that represents the type of action done by the user on the host system. DARPA data set mainly contains normal actions and attacks. Each text document represents a process which is an action. The DARPA data is shown in Table.1. Xin Xu et-al proposed Reinforcement Learning Approach for Host-Based Intrusion Detection Using Sequences of System Calls due to the complex and dynamic properties of intrusion behaviours, machine learning and data mining methods have been widely employed to optimize the performance of intrusion detection systems (IDSs)[6].

Table.1. Sequence of system calls

setpgrp	ioctl	setpgrp	ioctl
ioctl	setpgrp	ioctl	close
close	close	ioctl	ioctl
close	close	close	close
close	close	close	execve
open	mmap	open	mmap

The DARPA data set is pre-processed as 1-Word term matrix as shown in Table2.

Table2. One Word Sequence Term Matrix of User Interactions

munmap	seteuid	ftpaccess	getaudit	Fork	fcntl	chroot	ioctl	lstat	Setegid	kill	unlink	type_
12	0	0	0	1	4	0	35	0	0	0	0	0
11	0	0	0	1	6	0	141	0	0	0	1	0
5	0	0	0	0	0	0	53	0	0	0	0	0
5	0	0	0	0	0	0	10	1	0	0	0	0
5	0	0	0	0	0	0	52	0	0	0	0	0
8	0	0	0	2	0	0	23	10	0	0	7	0
5	0	0	0	0	0	0	11	2	0	0	0	0
5	0	0	0	0	0	0	9	0	0	0	0	1
3	0	0	0	4	0	0	6	0	0	0	3	1
5	0	0	0	0	0	0	9	0	0	0	0	1
10	2	0	0	8	35	0	262	2	0	0	2	1

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

III. FEED FORWARD ERROR BACK PROPAGATION ALGORITHM –NEURAL NETWORKS METHOD

Back Propagation network is considered to be quintessential Neural Network. Back Propagation is the training or learning algorithm rather than the network itself. To train the network we need to give the output called the Target for a particular input. The input and its corresponding target are called a Training Pair. Once the network is trained, it will provide the desired output for any of the input patterns. The network is first initialized by setting up all its weights to be small random numbers – say between –1 and +1. Next, the input pattern is applied and the output is calculated this is called the forward pass. The calculation gives an output which is completely different to what is expected (the Target), since all the weights are random. We then calculate the Error of each neuron, which is essentially: Target - Actual Output. This error is then used mathematically to change the weights in such a way that the error will get smaller. In other words, the Output of each neuron will get closer to its Target (this part is called the reverse pass). The process is repeated again and again until the error is minimal.

Algorithm:

Step1. Determine the architecture Number of input and output neurons.

Step2. Hidden neurons and layers.

Step3. Initialize all weights and biases to small random values, typically Є [-1, 1], choose a learning rate η.

Step4. Repeat until termination criteria satisfied

Present a training example and propagate it through the network (Forward pass)

Calculate the actual output
- Inputs applied
- Multiplied by weights
- Summed
- squashed by sigmoid activation function
- Output passed to each neuron in next layer

Adapt weights starting from the output layer and working backwards (backward pass). Nishat Mowla et-al followed the Neural Networks for most efficient classification in intrusion detection system [13]. Based on above statement, we used the feedforward Error Back Propagation algorithm for classification in the following section.

IV. METHODOLOGY

The "Feedforward Error Back Propagation Algorithm" is used as prediction model here on – Word term matrix to identify the attacks through host-based intrusion detection system shown in Figure 1. This model consists pre-process, prediction model and accuracy blocks. The sequence of system calls are converted into 1-Word term matrix. The conversion is called pre-processing. The prediction model is applied on 1-Word term matrix to find out intruders in the DARPA data set. The Feedforward Error Back Propagation algorithm is used in the prediction model. Third is accuracy block which will be used to specify the prediction model accuracy.

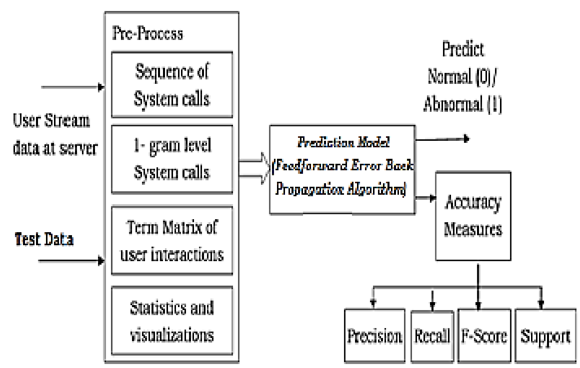


Figure 1. Block diagram of Host based IDS using Feedforward Error Back Propagation Algorithm

The prediction model how correctly it is classifying as either 0 or 1 can be expressed based on the accuracy measures such precision, recall, F-score and support.

The accuracy measures of the prediction model is illustrated below:
precision: When the model predicts 1, how often it is correct?.
The precision is the ratio of true positives(TP) to sum of false positives(FP) and true positives(TP).

Mathematically,
Precision = TP/(FP + TP)
recall:The measure recall is the ratio of true positives(TP) to sum of true positives(TN) and false negatives(FP).
Mathematically,
recall = TP/(TP + FN)

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 12, December 2017

*F-score:*The measure F-score is the 2 multiplied by ratio of the product precision and recall to the sum of precision and recall.

Mathematically,

F-score = 2((precision X recall) / (precision + recall))

Support: It is the sum of false negatives (FN) and true positives (TP).

Mathematically,

Support = FN + TP

You can correlate these measures with the help of experimental results obtained. The occurrences of system calls existed in the Training and Test data set is shown through the bar graph shown in Figure2. In the graph, the number of system calls used on x-axis and system calls are taken on y-axis. This graph shows the command(s) used maximum times and the command(s) used minimum times by the users.

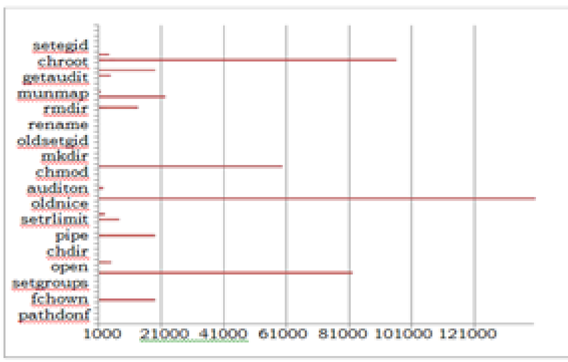


Figure 2. Total occurrences of system calls existed in the Training and Test data set

V. EXPERIMENTAL RESULTS

From the data set, 80% of total data is used for training the algorithm and the remaining 20% of data is used for testing. When we plot a Receiver Operating Characteristic (ROC) curve for Feedforward Error Back Propagation using one word sequence of system calls, the resulting curve is shown in Figure3. In this plot the model has taken False Positive Rate on X-axis and True Positive Rate on Y-axis. Under the area of curve 87% attacks are identified.

Neuralnetwork model results:

NNCoefficients:

[array([0.05582357, 0.12560743, -0.00502897, ..., -0.11708454,
0.26896494, 0.2299846]), array([-0.33487837, 0.16473477, 0.30000363, ..., -0.23860996,
0.40223645, -0.18367941]), array([0.36708108, -0.29406011, -0.12892285, 0.00523353, 0.22152779,
0.57462595]), array([1.95254626])]

[array([[-0.0345535 , 0.11496605, -0.28277785, ..., -0.17119056,
0.23531337, 0.26455328],
[-0.10554316, 0.10878952, 0.21291773, ..., 0.14038361,
-0.12419956, 0.163641],
[-0.21996145, -0.03040878, 0.23113639, ..., -0.04854751,
-0.18146248, 0.01994914],
...,
[-0.18164273, 0.14306382, -0.27870232, ..., -0.27599959,
-0.12469526, -0.27075048],
[0.28173417, 0.04632065, -0.11688396, ..., -0.01035414,
-0.02868439, 0.09756521],
[-0.10179763, 0.10836274, 0.10275406, ..., -0.0716117 ,
0.29780196, 0.18126182]]), array([[-0.16166147,
0.20597144, 0.2696928 , ..., -0.14661424,
-0.29926845, 0.1154414],
[0.27065336, 0.37357268, 0.15926641, ..., 0.35638725,
-0.09398284, 0.42017872],
[0.01603631, -0.27948065, -0.36280424, ..., 0.36666717,
-0.29913234, -0.08964466],
...,
[0.01844053, 0.46155756, 0.36700576, ..., 0.39447428,
0.23770596, -0.14134973],
[-0.36338899, -0.37009184, 0.05946834, ..., 0.3429833 ,
0.38296727, -0.39411064],
[-0.37846462, -0.3688318, -0.14542734, ..., -0.20656528,
0.00192554, 0.12225598]]), array([[0.51508948,
0.39993536, 0.50295277, -0.40743231, 0.45117687,
0.06490792],
[-0.33245345, -0.21520035, -0.40407641, -0.74286892, -0.49454876,
-0.4531981],
[-0.35975179, 0.22637952, 0.0877178 , 0.47579612, -0.30341977,
0.122944],
...,
[-0.24577935, -0.01157045, -0.37367375, -0.00918654, -0.37389813,
0.38246335],
[-0.00381892, 0.54329566, 0.17245819, -0.09624559, 0.55570443,
-0.43981448],
[0.1572731 , 0.19230166, -0.07859043, -0.52356133, -0.00864683,
-0.04149508]]), array([[-0.60574474],
[0.52402576],
[-0.49839802],
[-1.09567739],
[-0.66878146],
[0.56009075]])]

NN_ConfusionMatrix For Train:

[[314 89]

[17 451]]

NN_ConfusionMatrix for test:
[[150 53]]

Table3.Classification_report for train

	precision	recall	f1-score	support
0	0.95	0.78	0.86	403
1	0.84	0.96	0.89	468
avg /total	0.89	0.88	0.88	871

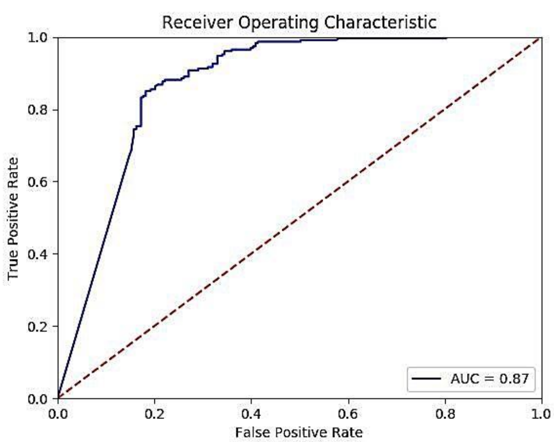
Table4.Classification_report for test

precision recall f1-

	precision	recall	f1-score	support
0	0.86	0.74	0.79	203
1	0.79	0.89	0.84	227
avg / total	0.82	0.82	0.82	430

The classification report for train and test data is shown in Table3 and Table 4 .

Figure.3. ORC curve for one word sequence



VI. CONCLUSIONS

Here, we applied Feedforward Error Back Propagation algorithm on 1-Word term matrix to detect the percentage of intrusions in the DARPA data set. This is experimented on DARPA data set with necessary preprocess steps such as generation of user one word sequence of system calls. These one word system calls are transformed into one word term matrix of size with maximum number of system calls. This data is modelled with Feedforward Error Back Propagation algorithm. The model is tested with test data of size 500 users and accuracy is determined in terms of precision, recall, f-score and support. The ROC curve is showing 87% of accuracy on one word term matrix. In this approach, our model can find 87% of intruders correctly using the DARPA data set.

REFERENCES

[1] Sanjay Rawat, Arun K. Pujari,V. P. Gulati, On the Use of Singular Value Decomposition for a Fast Intrusion DetectionSystem, VODCA 2004 Preliminary Version,pp.1 -13

[2] Solane Duquea, Dr.Mohd, Nizam bin Omar, Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS), Procedia Computer Science 61 (2015) 46 – 51(ELSEVIER)

[3] Richard Zuech*, Taghi M Khoshgoftaar and Randall Wald, Intrusion detection and Big Heterogeneous Data: a Survey, Journal of Big Data (2015)- Springer Open Journal 2:3, DOI 10.1186/s40537-015-0013-4

[4] Mahdi Zamani and Mahnush Movahedi, Machine Learning Techniques for Intrusion Detection {fzamani,movahedi}@cs.unm.edu, Department of Computer Science University of New Mexico, arXiv: 1312. 2177v2 [cs.CR] 9 May 2015.

[5] Harvinder Pal Singh Sasan and Meenakshi Sharma, Intrusion Detection Using Feature Selection and Machine Learning Algorithm with Misuse Detection, International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 1, February 2016

[6] Xin Xu and Tao Xie, A Reinforcement Learning Approach for Host-Based Intrusion Detection Using Sequences of System Calls, Springer-Verlag Berlin Heidelberg 2005, pp. 995 –1003.

[7] Murat OĞUZ, İhsan Ömür BUCAK, A Behavior Based Intrusion Detection System Using Machine Learning Algorithms, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (7) : Issue (2) : 2016,pp.9-24

[8] Sunil Kumar Gautam, Hari Om, Computational neural network regression model for Host based Intrusion Detection System, ScienceDirect Perspectives in Science(2016) 8, 93—95.

[9] Atilla Özgür, Hamit Erdem, The impact of using large training datasetKDD99on classification accuracy, https://doi.org/10.7287/peerj.preprints.283 8v1 | CC BY 4.0 Open Access | publ: 1 Mar 2017

[10] ABUBAKAR, Atiku and PRANGGONO, Bernardi,Machine learning based intrusion detection system for software defined networks, . Machine learning based intrusion detection system for software defined networks. In: Proceedings of the 2017 Eighth International Conference on Emerging Security Technologies (EST). IEEE. (In Press)

[11] Prachi, Usage of Machine Learning for Intrusion Detection in a Network, International Journal of Computer Networks and Applications (IJCNA) DOI: 10.22247 /ijcna /2016 /41278 Volume 3, Issue 6, November – December (2016)

12]. L.P.Dias et-al, Using artificial neural network in intrusion detection systems to computer networks, Published in: Computer Science and Electronic Engineering (CEEC), 2017, INSPEC Accession Number: 17354350, Publisher: IEEE

[13]. Nishat Mowla et-al, Evolving neural network intrusion detection system for MCPS, INSPEC Accession No: 16777265, Publisher:IEEE 2017 [14]. Norbart Adam et-al, Artificial neural network based IDS, INSPEC Accession Number: 16758076, DOI: 10.1109 /SAMI. 2017.7880294, IEEE-March 2017

[15] Ms. Vrushali D Mane , Ms. Abhilasha Sayar , Prof. Sunil Pawar, Anomaly Intrusion Detection System Using Neural Network , IJCSMC, Vol. 2, Issue. 8, August 2013, pg.76 – 81