International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com ǁ Volume 02 - Issue 01 ǁ January 2017 ǁ PP. 28-33

Comparative Analysis of Deep Neural Network with different algorithms for Detecting Intrusions

Kanupriya Arora[1], Ritu Chauhan[2]
[1]*(Computer science and engineering/ ITM University, India)*
[2]*(Computer science and engineering/ ITM University, India)*

Abstract: Security on the web is a crucial issue and the intrusion recognition is one of the real research issues for business and individual systems. The intrusion detection system is an application to monitor the unauthorized access to different devices. In this paper different data mining algorithms are analyzed to find out the efficiencies and for classification of system activities. Different data mining algorithms, like linear regression, k-mean clustering, SVM and Deep neural network are deeply studied. Among all the various methodology we observed deep neural network algorithm works very efficiently and effectively.
Keywords: Linear Regression; K-Means clustering;SVM;Deep Neural Network**.**

I. INTRODUCTION

An Intrusion Detection System (IDS) is a application that screen systems or framework exercises for all the pernicious exercises and unapproved access to gadgets. IDS arrive in an assortment of "flavors" and it goes for recognizing suspicious activity in numerous ways [1].

A few sorts of assaults on correspondence system are vulnerabilities, SYN flooding, Distributed Denial-of-Service (DDOs), surfing and the rundown goes on. Interruption identification implies to the identification of noxious movement (assaults, break-ins, infiltrations and different types of PC mishandle) in a PC related frameworks or in the correspondence systems. IDS distinguish gatecrasher's activities that undermine the confidentiality, accessibility, integrity and honesty of assets [4].

There are fundamentally two sorts of IDS: host-based (HIDS) and network based (NIDS). HIDS dwells on a specific host and searches for signs of assaults on that specific host while an NIDS analyzes the activity packets progressively or near genuine time and endeavor to identify the interruption patters in the system movement [4]. The accessible assets and the general dangers to the association settles on the decision of which IDS to be utilized.

The attacks classes show in the NSL-KDD information set are assembled into four classifications [2] :

a) DOS: Denial of a service is an assault classification, which drains the victim□ s assets consequently making it not able to handle authorized requests – e.g. syn flooding. Pertinent elements: "source bytes" and "rate of error packets"

b) Probing: Surveillance and other examining attack□ s goal is to pick up data aboutthe remote victim e.g. port examining. Pertinent elements: "length of association" and "source bytes"

c) U2R: unapproved access to super local client (root) benefits is an assault sort, by which an attack dog uses a ordinary record to login into a victim framework and tries to pick up root/controller benefits by abusing a few powerlessness in the victim e.g. buffer overflow assaults. Pertinent components: "number of record creation" and "number of shell prompts invoked,"

d) R2L: unapproved access from a remote machine, the attack dog interrupt into a remote machine and increases local access of the victim machine. E.g. secret key speculating Pertinent components: Network level elements – "term of association" and "administration asked for" and have level highlights - "number of fizzled login endeavors"

The investigation of the NSL-KDD information set [2] is made by utilizing different bunching calculations accessible in the WEKA information mining instrument. The NSL-KDD information set is dissected and ordered into four unique groups delineating the four normal diverse sorts of assaults. A top to bottom expository review is made on the test and preparing information set. To utilize the NSL-KDD, the information is set to uncover the most helpless convention that is often utilized gatecrashers to dispatch organize based interruptions.

Correspondence framework assumes an unavoidable part in
normal man's day by day life. PC systems are adequately utilized for business information preparing, training also ,learning, across the board information procurement also and stimulation. The PC arrange convention stack that is being used today was created with an intention to make it straightforward and easy to understand. The adaptability of the convention has made it defenseless against the assaults propelled by the gatecrashers. This makes the prerequisite for the PC systems to be constantly observed and secured. The checking procedure is robotized by an intrusion detection system (IDS) [2]. The IDS can be made of blend of equipment and software.

II. LITERATURE SURVEY

Dikshant Gupta, Suhani Singhal, Shamita Malik and Archana Singh [1] proposed the technique of NIDS (Network Intrusion Detection System). The objective of planning NIDS is to secure the information's classification and honesty. Our venture concentrates on these issues with the help of Data Mining. This examination paper incorporates the usage of various information mining calculations counting Linear regression and K-Means clustering to consequently produce the tenets for group arrange exercises. A near examination of these strategies to identify interruptions has additionally been made. To take in the examples of the assaults, NSL-KDD dataset has been utilized.

L.Dhanabal and Dr.S.P.Shantharajah [2] dissected the NSL-KDD information set and used to concentrate the adequacy of the different order calculations in identifying the oddities in the system activity designs. We have additionally investigated the relationship of the conventions accessible in the normally utilized system convention stack with the assaults utilized by interlopers to create peculiar system activity. The examination is done utilizing characterization calculations accessible in the information mining instrument WEKA. The review has uncovered numerous truths about the holding between the conventions and system assaults.

Sasanka Potluri and Christian Diedrich [4] principally centers taking care of the enormous datasets with quickening agent stage and finding the unpredictable relations on the input dataset for distinguishing diverse assault sorts. Quickened DNN based IDS is produced for this reason. Trial comes about demonstrated that the proposed calculation could distinguish the relations between the info information. It demonstrates that the DNN based IDS is dependable and productive in interruption identification for distinguishing the particular assault classes with required number of tests for preparing (see DoS and Probe assaults in Table 3) and was not able adequately classify the assault sorts (see R2L and U2R in Table 3) with least number of tests for preparing. We can likewise watch that the identification exactness's were dependable on NSL-KDD dataset by summing up the assault classes to less sorts. Because of absence of adequate information for preparing the U2R and R2L were not very much distinguished and this lessens the generally speaking discovery precision of the DNN based IDS.

Tuan A Tang, Lotfi Mhamdi, Des McLernon, Syed Ali Raza Zaidi and Mounir Ghogho [5] apply a profound learning approach for stream based abnormality discovery in a SDN environment. They manufacture a Deep Neural Network (DNN) display for an interruption recognition framework and prepare the model with the NSLKDD Dataset. In this work, they simply utilize six essential elements (that can be effortlessly acquired in a SDN situation) taken from the fortyone elements of NSL-KDD Dataset. Through investigations, we affirm that the profound learning approach demonstrates solid potential to be utilized for stream based irregularity discovery in SDN situations

III. Dataset description

To check the effectiveness and practicality of the proposed IDS framework, we have utilized NSL-KDD dataset.

It is a fresh out of the box new form of KDDcup99 dataset. NSLKDD dataset have a few favorable circumstances over KDDcup99 dataset. It has tackled a portion of the natural issues of the KDDcup99, which is considered as the standard benchmark implied for interruption identification appraisal. The preparation dataset of NSL-KDD is like KDDcup99 which comprise of around 4,900,000 single association vectors each of which contain 41 highlights and are marked as either typical on the other hand assault sort, with precisely one particular assault sort.

Sr.No.	Feature Name
1	Duration
2	Protocol_type
3	Service
4	Flag
5	Src_bytes
6	Dst_bytes
7	Land
8	Wrong_fragment
9	Urgent
10	Hot
11	Num_failed_logins
12	Logged_in
13	Num_compromised
14	Root_shell
15	Su_attempted
16	Num_root
17	Num_file_creations
18	Num_shells
19	Num_acess_files
20	Num_outboun_cms
21	Is_host_login
22	Is_guest_login
23	Count
24	Srv_count
25	Serror_rate
26	Srv_serror_rate
27	Rerror_rate
28	Srv_rerror_rate
29	Same_srv_rate
30	Diff_srv_rate
31	Srv_iff_host_rate
32	Dst_host_count
33	Dst_host_srv_count
34	Dst_host_same_srv_rate
35	Dst_host_diff_srv_rate
36	Dst_host_same_src_port_rate
37	Dst_host_srv_diff_host_rate
38	Dst_host_serror_rate
39	Dst_host_srv_serror_rate
40	Dst_host_rerror_rate
41	Dst_host_srv_rerror_rate
42	Normal or Attack

FEATURES OF NSL-KDD CUP'99 DATASET[1]

Because of the accompanying reasons, NSL-KDD hasturn out to be more prevalent dataset than KDD glass 99 dataset for interruption location purpose[1].

1. Repetitive records from the preparing set are disposed of.

2. Copy records from the test sets are evacuated to upgrade the interruption discovery execution.

3. Utilization of NSL-KDD dataset for characterization gives aprecise assessment of various learning systems.

4. It is moderate to utilize NSL-KDD dataset for the test reason furthermore it comprises of sensible quantities of examples both in the preparation set too as in the testing set.

IV. DIFFERENT APPLICATIONS OF ALGORITHM IN IDS

i) Linear Regression[1]:

1. Study of linear relationship between an output variable and different input features.
2. Here hypothesis form[6] will be

$$h\theta(x) = \theta 0 + \theta 1x \quad (1)$$

Here we have two parameters determined by cost function (θ0, θ1) and one variable x .

3. Now we will change θ0 and θ1 where (x) is close to y.
4. Gradient Descent Algorithm

$$J(\theta 0, \theta 1)=(1/2m)*\sum_{i=1}^{m}(h(x^i)-y^i)^2 \quad (2)$$

4. To minimize the error and it also suited to big data (when n is massive) we use Gradient Descent Algorithm .

ii) K-Means Clustering

K-means is one of the most straightforward unsupervised learning calculations that take care of the outstanding bunching issue. The strategy takes after a basic and simple approach to arrange a given information set through a specific number of groups (accept k bunches) settled from the earlier. The primary thought is to characterize k centroids, one for every bunch. These centroids ought to be put slyly as a result of various area causes distinctive outcome. Thus, the better decision is to place them however much as could reasonably be expected far from each other.

The calculation is made out of the accompanying strides[7]:

- Put K points into the space spoke to by the objects that are being grouped. These points speak to beginning gathering centroids .
- Assign each object to the group that has the closest centroid
- At the point when the sum total of what objects have been doled out, recalculate the places of the K centroids

Rehash Steps 2 and 3 until the centroids do not move anymore. This creates a detachment of the items into gatherings from which the metric to be minimized can be ascertained.

iii) SVM

Support vector machine consists of three stages[10] :

- Preprocessing:It is used to convert the non-numerical values to the numerical values e.g TCP-1,UDP-2, and ICMP-3 .Also it converts the attack types into the numerical values .
- Training: Here we are preparing SVM for training on different normal data and attacks .As we already know that we have 41 features which fall into these two categories .
- Testing: To measure out the performance we will apply the testing .

Support vector machines, or SVMs[11], are learning machines that plot the preparation vectors in high-dimensional highlight space, marking every vector by its class. SVMs characterize information by deciding an arrangement of support vectors, which are individuals from the arrangement of training inputs .

Due to following reason we use SVM in IDS .

- Speed
- Scalability

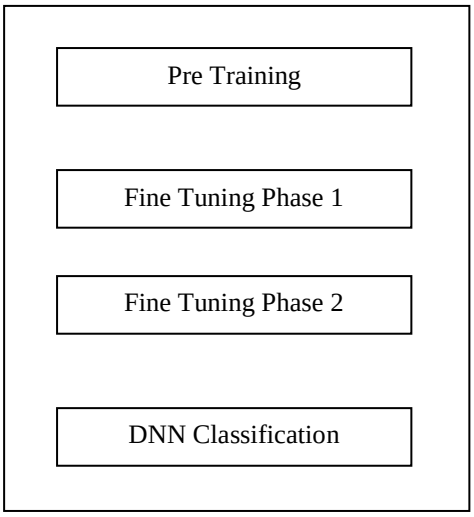
The SVMs depend on the possibility of auxiliary hazard minimization, which minimizes the speculation mistake, i.e. genuine mistake on inconspicuous cases. The quantity of free parameters utilized as a part of the SVMs relies on upon the edge that isolates the information focuses however not on the quantity of info highlights, consequently SVMs don't require a lessening in the number of elements keeping in mind the end goal to abstain from overfitting.

iv) Deep Neural Network

DNN's learn in various leveled layers of representation from data sources so as to perform important order undertaking . A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between inputs and its outputs[4] . Each layers can learn features at a diverse level of abstraction.

In order to make effective training of each layer first of all we will choose to train one layer at a time . This is accomplished by training a unique kind of system known as auto-encoders for each wanted hidden layers[12]. Here pre training of the DNN is divided into different steps.

Deep Neural Network



DNN Based Workflow[4]

a. Pre training : Pretraining is a way to avoid network initialization closer to local minima. As the learned initialization from pretraining gives an idea to network what to look for the network should initialize closer to a global minima.During pretraining you give all of your training without it's label to the data. And you try to minimize the reconstruction error of the data .We used it as a feature extractor and as a fined tuned network or to train model on a bigger data . Here we are providing the training to the first auto encoder after that we will provide training to the other auto encoder but before giving training to second we will extract feature from the first one.
b. Fine tuning phase 1 : To classify the attacks and normal data ,here supervised training is done with the features extracted from the auto encoder .
c. Fine tuning phase 2 :In this phase backpropagation is performed on all hidden layers ,it is used to compute the necessary corrections ,the backpropagation algorithm is stopped when the value of the error function has become sufficiently small ,therefore it is more useful in IDS .
d. Due to all these factors time needed for training each of the layer decreases and efficiency to attack detection increases .

V. CONCLUSION AND FUTURE SCOPE

Deep Neural Network is the best algorithm for detecting attacks in intrusion detection system .Algorithm based on data mining shows different accuracies depending on the attack type. This analysis shows that NSL-KDD dataset is a best appropriate data set to simulate and test the performance of IDS (Intrution Detection System). In future, Dimensionality Reduction using principal component analysis can be used to improve the time and visualization.

References

[1] Dikshant Gupta, Suhani Singhal, Shamita Malik and Archana Singh, "Network Intrusion Detection SystemUsing various data mining techniques," RAINS-2016, pp.1-6,April 06-07, 2016.
[2] L.Dhanabal1, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms" International Journal of Advanced Research in Computer and Communication Engineering,Vol. 4, Issue 6, June 2015
[3] Tavallaee , Mahbod , " A detailed analysis of the KDD CUP 99 data set ." Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009 .

[4] Sasanka Potluri, Christian Diedrich, "Accelerated deep neural networks for enhanced Intrusion Detection System" 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA),Sept 6-9, 2016.

[5] Tuan A Tang,Lotfi Mhamdi , Des McLeron, Syed Ali Raza Zaidi and Mounir Ghogo, " Deep Learning Approach for Network Intrusion Detection in Software Defined Networking", International Conference on Wireless Networks and Mobile Communations(WINCOM)-2016,p 258-263,2016.

[6] http://www.holehouse.org/mlclass/04_Linear_Regression_with_multiple_variables.html .

[7] Dabas , Poonam , and Rashmi Chaudhary . " Survey of Network Intrusion Detection Using K-Mean Algorithm ." International Journal of Advanced Research in Computer Science and Software Engineering 3 . 3 (2013): 507 – 511 .

[8] Solanki , Miss Meghana , and Mrs Vidya Dhamdhere . " Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier ."

[9] KDD Cup 1999, Ocotber 2007, [online] Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[10] S Mukkamala, G Janowski, A H. Sung, "Intrusion Detection Using Neural Networks and Support Vector Machines", *Proceedings of IEEE International Joint Conference on Neural Networks 2002*, pp. 1702-1707, May 2002.

[11] S Mukkamala, G Janowski, A H. Sung, " Identifying important features for intrusion detection using support vector machines and neural networks" , Symposium on application and the internet ,2003,proceeding,pp 209-216, 2003 .

[12] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle,"Greedy Layer Wise Training of Deep Networks," Adv. Neural Inf. Process. Syst., vol. 19, p. 153, 2007.