DETECTION OF CYBER ATTACK IN NETWORK USING MACHINE LEARNING TECHNIQUE

N. SRINIVASA RAO , BANDARU VEERABABU

Assistant Professor, DEPT. of MCA, SKBR PG COLLEGE, AMALAPURAM, Andhra Pradesh

Email: -naagaasrinu@gmail.com

PG Student of MCA, SKBR PG COLLEGE, AMALAPURAM, Andhra Pradesh

Email:- veerababubandaru55@gmail.com

Abstract: Contrasted with the past, improvements in PC and correspondence innovations have given broad and propelled changes. The use of new innovations gives incredible advantages to people, organizations, and governments, be that as it may, messes some up against them. For instance, the protection of significant data, security of put away information stages, accessibility of information and so forth. Contingent upon these issues, digital fear-based oppression is one of the most significant issues in this day and age. Digital fear, which made a great deal of issues people and establishments, has arrived at a level that could undermine open and nation security by different gatherings, for example, criminal association, proficient people and digital activists. Along these lines, Intrusion Detection Systems (IDS) has been created to maintain a strategic distance from digital assaults. Right now, learning the bolster support vector machine (SVM) calculations were utilized to recognize port sweep endeavors dependent on the new CICIDS2017 dataset with 97.80%, 69.79% precision rates were accomplished individually. Rather than SVM we can introduce some other algorithms like random forest, CNN, ANN where these algorithms can acquire accuracies like SVM – 93.29, CNN – 63.52, Random Forest – 99.93, ANN – 99.11

Index Term: Intrusion Detection Systems (IDS), Criminal Association, Digital Assaults.

I Introduction

The use of new innovations give incredible advantages to people, organizations, and governments, be that as it may, messes some up against them. For instance, the protection of significant data, security of put away information stages, accessibility of information and so forth. Contingent upon these issues, digital fear based oppression is one of the most significant issues in this day and age. Digital fear, which made a great deal of issues people and establishments, has arrived at a level that could undermine open and nation security by different gatherings, for example, criminal association, proficient people and digital activists. Along these lines, Intrusion Detection Systems (IDS) has been created to maintain a strategic distance from digital assaults.

2 Literature survey

Port Scanning is one of the most popular techniques attackers use to discover services that they can exploit to break into systems. All systems that are connected to a LAN or the Internet via a modem run services that listen to well-known and not so well-known ports. By port scanning, the attacker can find the following information about the targeted systems: what services are running, what users own those services, whether anonymous logins are supported, and whether certain network services require authentication. Port scanning is accomplished by sending a message to each port, one at a time. The kind of response received indicates whether the port is used and can be probed for further weaknesses. Port scanners are important to network security technicians because they can reveal possible security vulnerabilities on the targeted system. Just as port scans can be ran against your systems, port scans can be detected and the amount of information about open services can be limited utilizing the proper tools. Every publicly available system has ports that are open and available for use. The object is to limit the exposure of open ports to authorized users and to deny access to the closed ports. Port scanning is a common activity of considerable importance. It is often used by computer attackers to characterize hosts or networks which they are considering hostile activity against. Thus it is useful for system administrators and other network defenders to detect ports cans as possible preliminaries to a more serious attack.

It is also widely used by network defenders to understand and find vulnerabilities in their own networks. Thus it is of considerable interest to attackers to determine whether or not the defenders of a network are port scanning it regularly. However, defenders will not usually wish to hide their port scanning, while attackers will. For definiteness, in the remainder of this paper, we will speak of the attackers scanning the network, and the defenders trying to detect the scan. There are several legal/ethical debates about port scanning which break out regularly on Internet mailing lists and newsgroups.

One concerns whether port scanning of remote networks without permission from the owners is itself a legal and ethical activity. This is presently a grey area in most jurisdictions. However, our experience from following up on unsolicited remote ports cans we detect in practice is that almost all of them turn out to have come from compromised hosts and thus are very likely to be hostile. So we think it reasonable to consider a ports can as at least potentially hostile, and to report it to the administrators of the remote network from whence it came.However it focuses on the technical questions of how to detect ports cans, which are independent of what significance one imbues them with, or how one chooses to respond to them. Also, we are focussed here on the problem of detecting a port scan via a network intrusion detection system (NIDS). We try to take into account some of the more obvious ways an attacker could use to avoid detection, but to remain with an approach that is practical to employ on busy networks. In the remainder of this section, we first define port scanning, give a variety of examples at some length, and discuss ways attackers can try to be stealthy. In the next section, we discuss a variety of prior work on port can detection. Then we present the algorithms that we propose to use, and give some very preliminary data justifying our approach. Finally, we consider possible extensions to this work, along with other applications that might be considered. Throughout, we assume the reader is familiar with Internet protocols, with basic ideas about network intrusion detection and scanning, and with elementary probability theory, information theory, and linear algebra. There are two general purposes that an attacker might have in conducting a port scan: a primary one, and a secondary one. The primary purpose is that of gathering information about the reach ability and status of certain combinations of IP address and port (either TCP or UDP). (We do not directly discuss ICMP scans in this paper, but the ideas can be extended to that case in an obvious way.) The secondary purpose is to flood intrusion detection systems with alerts, with the intention of distracting the network defenders or preventing them from doing their jobs. In this paper, we will mainly be concerned with detecting information gathering port scans, since detecting flood port scans is easy. However, the possibility of being maliciously flooded with information will be an important consideration in our algorithm design. We will use the term scan footprint for the set of port/IP combinations which the attacker is interested in characterizing. It is helpful to conceptually distinguish the footprint of the scan, from the script of the scan, which refers to the time sequence in which the attacker tries to explore the footprint. The footprint is independent of aspects of the script, such as how fast the scan is, whether it is randomized, etc

The attackers used the attack tools such as Nikto, Nessus, and Web Scarab to carry out reconnaissance and attacks automatically. This dataset can be used to test IDS alert rules, but it suffers from the lack of traffic diversity and volume (Sangster et al., 2009). Kyoto (Kyoto University 2009): This dataset has been created through hornpouts, so there is no process for manual labelling and anonymization, but it has limited view of the network traffic because only attacks directed at the honey pots can be observed. It has ten extra features such as IDS Detection, Malware Detection, and Ashula Detection than previous available datasets which are useful in NIDS analyToward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization 109 sis and evaluation. The normal traffic here has been simulated repeatedly during the attacks and producing only DNS and mail traffic data, which is not reflected in real world normal network traffic, so there are no false positives, which are important for minimizing the number of alerts (Song et al., 2011) (M. Sato, 2012) (R. Chitrakar, 2012). Twenty (University of Twenty 2009): This dataset includes three services such as OpenSSH, Apache web server and Profit using auth/indent on port 113 and captured data from a honey pot network by Net flow. There is some simultaneous network traffic such as auth/indent, ICMP, and IRC traffic, which are not completely benign or malicious. Moreover, this dataset contains some unknown and uncorrelated alerts traffic. It is labelled and is more realistic, but the lack of volume and diversity of attacks is obvious (Spratt et al., 2009). UMASS (University of Massachusetts 2011): The dataset includes trace files, which are network packets, and some traces on wireless applications (of Massachusetts Amherst, 2011) (Nehinbe, 2011). It has been generated using a single TCP-based download request attack scenario. The dataset is not useful for testing IDS and IPS techniques due to the lack of variety of traffic and attacks (SwagatikaPrusty and Liberator, 2011). ISCX2012 (University of New Brunswick 2012). This dataset has two profiles, the Alpha-profile which carried out various multi-stage attack scenarios, and the Beta-profile, which is the benign traffic generator and generates realistic network traffic with background noise. It includes network traffic for HTTP, SMTP, SSH, IMAP, POP3, and FTP protocols with full packet payload.

However, it does not represent new network protocols since nearly 70% of today's network traffics are HTTPS and there are no HTTPS traces in this dataset. Moreover, the distribution of the simulated attacks is not based on real world statistics (Ali Shiravi and Ghorbani, 2012). ADFA (University of New South Wales 2013): This dataset includes normal training and validating data and 10 attacks per vector (Creech and Hue, 2013). It contains FTP and SSH password brute force, Java based Meterpreter, Add new Superuser, Linux Meterpreter payload and C100 Webshel attacks. In addition to the lack of attack diversity and variety of attacks, the behaviors of some attacks in this dataset are not well separated from the normal behavior (Xie and Hue, 2013) (Xie et al., 2014)

3. Implementation Study

Blameless Bays and Principal Component Analysis (PCA) were been used with the KDD99 dataset by Alanson and Limit. Similarly, PCA, SVM, and KDD99 were used Chithik and Rabbinic for IDS . In Aljawarneh et all's. Paper, their assessment and examinations were conveyed reliant on the NSL®KDD dataset for their IDS model Composing inspects show that KDD99 dataset is continually used for .There are 41 highlights in KDD99 and it was created in 1999. Consequently, KDD99 is old and doesn't give any data about cutting edge new assault types, example, multi day misuses and so forth. In this manner we utilized a cutting-edge and new CICIDS2017 dataset in our investigation.

DISADVANTAGES OF EXISTING SYSTEM:

- Strict Regulations
- Difficult to work with for non-technical users
- Restrictive to resources
- Constantly needs Patching
- Constantly being attacked

3.1proposed methodology

PROPOSED SYSTEM

Important steps of the algorithm are given in below.

1) Normalization of every dataset.
2) Convert that dataset into the testing and training.
3) Form IDS models with the help of using RF, ANN, CNN and SVM algorithms.
4) Evaluate every model's performances..

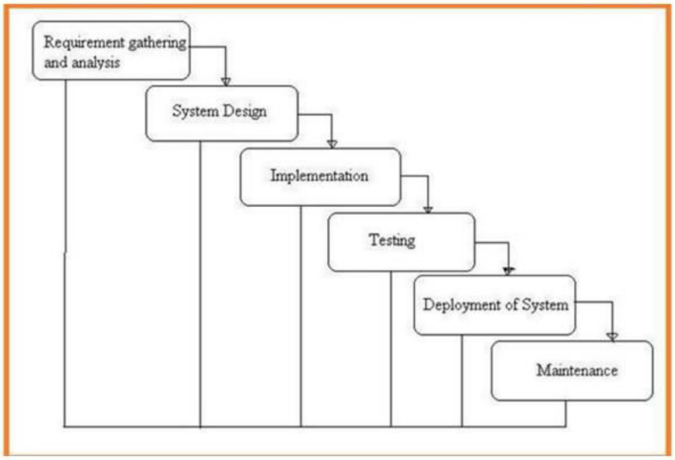


Fig 1: System Architecture

SYSTEM ARCHITECTURE

- Project Requisites Accumulating and Analysis
- Application System Design
- Practical Implementation
- Manual Testing of My Application
- Application Deployment of System Maintenance of the project

3.2 Methodology and Alogrithams

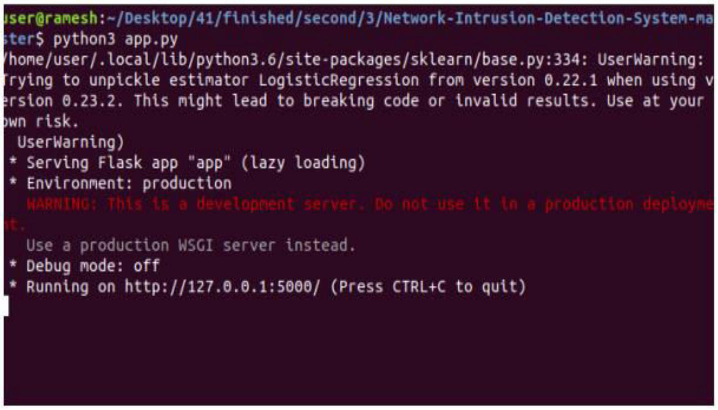
MODULES:

- NLTK
- NUMPY
- PANDAS

MODULES DESCRIPTION:

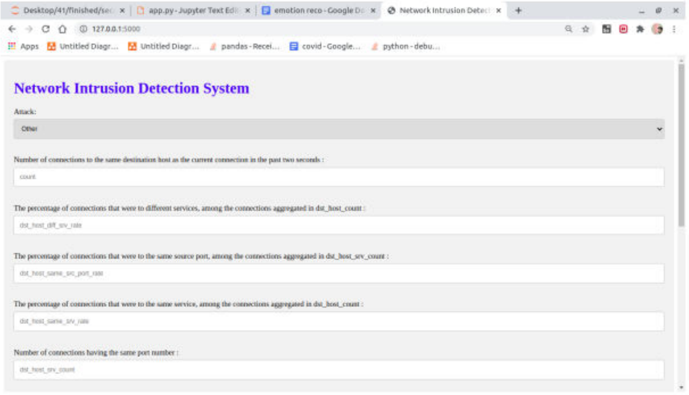
- NLTK: NATURAL LANGUAGE TOOLKIT NLTK is a leading platform for building python programs to work with human language data. It provides easy-to-use interfaces to over 50 Corpora and lexical.
- NUMPY: NUMPY is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fouier transform, and matrices.
- PANDAS: PANDAS is an open source python package that is most widely used for data science/data analysis and machine learning tasks

4 Results and Evolution Metrics



```
user@ramesh:~/Desktop/41/finished/second/3/Network-Intrusion-Detection-System-ma
ster$ python3 app.py
/home/user/.local/lib/python3.6/site-packages/sklearn/base.py:334: UserWarning:
Trying to unpickle estimator LogisticRegression from version 0.22.1 when using v
ersion 0.23.2. This might lead to breaking code or invalid results. Use at your
own risk.
  UserWarning)
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployme
nt.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig2 : Localhost - In Cmd Python App.Py:



Network Intrusion Detection System

Attack:

Other

Number of connections to the same destination host as the current connection in the past two seconds :

count

The percentage of connections that were to different services, among the connections aggregated in dst_host_count :

dst_host_diff_srv_rate

The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count :

dst_host_same_src_port_rate

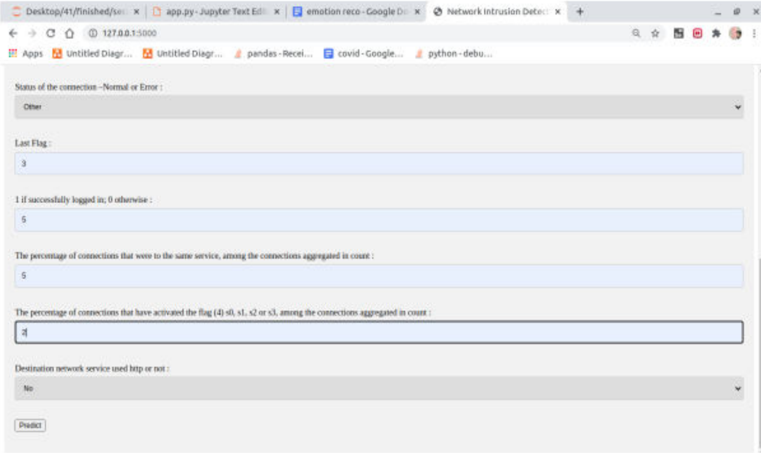
The percentage of connections that were to the same service, among the connections aggregated in dst_host_count :

dst_host_same_srv_rate

Number of connections having the same port number :

dst_host_srv_count

Fig3: input form for difreent parameters



Status of the connection –Normal or Error :

Other

Last Flag :

3

1 if successfully logged in; 0 otherwise :

5

The percentage of connections that were to the same service, among the connections aggregated in count :

5

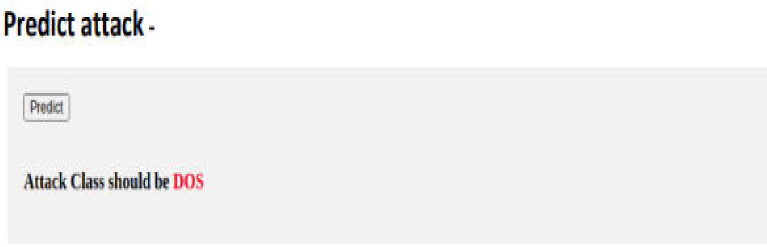
The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count :

2

Destination network service used http or not :

No

Fig 4: Enter the Input form


Predict attack -

Attack Class should be DOS

Fig5: predicted attack

5 .Conclusion

Right now, estimations of help vector machine, ANN, CNN, Random Forest and profound learning calculations dependent on modern CICIDS2017 dataset were introduced relatively. Results show that the profound learning calculation performed fundamentally preferable outcomes over SVM, ANN, RF and CNN. We are going to utilize port sweep endeavours as well as other assault types with AI and profound learning calculations, apache, Hodoop and sparkle innovations together dependent on this dataset later on. All these calculation helps us to detect the cyber attack in network. It happens in the way that when we consider long back years there may be so many attacks happened so when these attacks are recognized then the features at which values these attacks are happening will be stored in some datasets. So by using these datasets we are going to predict whether cyber attack is done or not. These predictions can be done by four algorithms like SVM, ANN, RF, CNN this paper helps to identify which algorithm predicts the best accuracy rates which helps to predict best results to identify the cyber attacks happened or not. In enhancement we will add some ML Algorithms to increase accuracy

6 .References

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, "Bilgi g ¨uvenli ˘gisistemlerindekullanilanarac¸larinincelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," Journal of Computer Security, vol. 10, no. 1-2, pp. 105–136, 2002.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, "Surveillance detection in high bandwidth environments," in DARPA Information Survivability Conference and Exposition, 2003.Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[6] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

[7] N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," in Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on. IEEE, 2015, pp. 25–31.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, "Detection and classification of malicious patterns in network traffic using benford's law," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864–872.

[9] S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detection system using naive bayes and pca algorithm," in Convergence in Technology (I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.

[10] M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component

analysis for ids,” in IEEE International Conference on Communication and Electronics Systems, 2016, pp. 1–5. 79

[11] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization.” in ICISSP, 2018, pp. 108–116.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, “Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm,” in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.

[14] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, “Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark,” IEEE Access, 2018.

[15] P. A. A. Resende and A. C. Drummond, “Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling,” Security and Privacy, vol. 1, no. 4, p. e36, 2018.

[16] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[17] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, “Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct,” Bone marrow transplantation, vol. 49, no. 3, p. 332, 2014.