INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 2057-5688

Network Security Model for Intrusion Detection with Bigdata Spark Using MLib

Submitted by

EADA SRINIVASA REDDY (17W91D5808) (Mtech)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MALLA REDDY INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Autonomous Institution - UGC, Govt. of India)

Hyderabad, TS, India.

Under the Esteemed Guidance of

MR. SVSV PRASAD SANABOINA B.Tech, M.Tech, (Ph.D)

ASST PROFESSOR, CSE

MALLA REDDY INSTITUTE OF ENGINEERING AND TECHNOLOGY

Hyderabad, TS,

Abstract:

An intrusion detection device is employed to detect various kinds of malicious behavior that could affect the security and reliability of an electronic system. It is able to detect attacks on vulnerable networks as well as data driven attacks on software, host-based attacks, such as privilege escalation, unauthorized logins, access to sensitive files as well as malware. It can operate in the context of host or the network level by using signature-based or misuse-based detection, and anomaly detection. Typically, attacks that are unable to be detected by a network-based intrusion detection systems can be detected by an intrusion detection system that is hosted and the reverse is true. At every level, attacks are detected using an intrusion detection techniques, such as anomaly detection or misuse detection. In the case of misuse detection, it can detect only known threats with high detection accuracy , whereas anomaly detection can identify both known and undiscovered attacks with a an extremely high false positive rates. To overcome the weaknesses of the individual system of intrusion detection, this research paper proposes a new data mining-based multi-layered intrusion detection.

ISSN : 2057-5688

In this study an intelligent hybrid architecture is suggested to combine detection methods as well as various levels for intrusion detection systems. To achieve this methods of data mining like clustering and classification algorithms have been suggested and implemented to aid in the selection of features as well as misuse detection and detection of anomalies. In this research , system for hybrid intrusion detection were suggested as a combination of misuse-anomaly detection and network-host. They've also been developed to ensure a the highest detection rate while reducing number of false positives. To select features, techniques for data pre-processing that are based in Genetic Algorithm, Support Vector Machine and Discriminant Analysis are used in the sense of generating a pertinent collection of attributes taken from the KDD CUP 1999 Dataset. The hybrid intrusion detection system was evaluated and verified by comparing it to the current intrusion detection systems, and it has been found that the new hybrid intrusion detection method has greater detection and lower false positive rate.

.

I Introduction

With the Internet being a major factor in constant communication however, its efficiency can decrease because of the intrusions. An intrusion is a process that negatively impacts the target system. An intrusion (Heady and co. 1990) can affect the integrity, confidentiality and accessibility of the resources of the system being attacked. A computer's security may be at risk when an attack occurs. Intrusions are classified as host intrusions as well as network intrusions. Host intrusions are attempts that are not authorized to gain access, manipulate and alter, or even destroy information, or render a system unstable or inaccessible. They can be caused by manipulating system calls, modifications to files and privilege escalation. They also include unauthorized logins, and access to sensitive files as well as malware (viruses trojan horses, viruses, and worms) that alter the status of the system.

Network intrusions are those attacks that occur by incoming packets within the network, which carry out harmful activities, such as Denial of Service (DoS) attacks, or even attempting to hack into computers. An DoS attack is a method to block a

ISSN : 2057-5688

computer's resources inaccessible to the intended users. There are a variety of attacks, including Land and Ping Of Death (POD), Flood attacks and others. The indicators of intrusions are unexpected results when executing different requests from users, as well as slow performance of the system unexpected system crashes, modifications to kernel data structures, and unusually slow performance of networks (opening files or opening websites). Methods for preventing intrusions like authenticating users (e.g. using biometrics or passwords) as well as avoiding programming errors and information security (e.g. encryption) have been employed to secure computer systems as the first option for defense. Security measures alone are not enough because as systems grow more complicated there will always be vulnerability that can be exploited that can be exploited due to the design or programming mistakes, or other "socially engineered" methods of penetration.

For instance, even after it was first discovered years ago an vulnerability "buffer overflow" persists in new software programs because of programming mistakes. The rules that combine convenience with strict control over the system and access to information are also making it difficult for a system operating to be 100% safe. The intrusion detection (Lee and Stolfo 1998) is, therefore, an essential another security feature to safeguard computers. The term "intrusion detection" refers to the practice of monitoring networks or computers to detect unauthorized access or activity, as well as file modification. The main components of intrusion detection include: the resources to be protected within the system of target, i.e., user accounts as well as system kernels, file systems as well as models that describe how "normal" and "legitimate" behaviour of the resources they are referring to; methods that analyze the system's actions with established models and detect those that have been identified as "abnormal" as well as "intrusive".

Hybrid intrusion detection systems

Intrusion Detection Systems is defined as the technique used to detect threats and the position of the IDS within the network. IDS can perform anomaly detection or misuse detection and is either a network-based or host-based systems. The result can be found in four broad categories, namely misuse-host, misuse-network and anomaly-host, and

ISSN : 2057-5688

anomaly network (Bashah and colleagues, 2005). Certain IDSs incorporate characteristics from the four categories (usually employing both anomaly detection and misuse) and are referred to as hybrid systems. It is vital to identify the main difference between anomaly detection as well as methods for detecting misuse (Ghosh and Schwartzbard 1999). The primary benefit of misuse detection methods is that well-known attacks are detected quite reliably and with low false positive rates. Since specific attack sequences can be stored in systems for detecting misuse which makes it easy to pinpoint exactly what attacks, or potential attacks, the system currently encountering. If the logs do not contain the signature of an attack the alarm will not be signalled. This means that there is a false-positive rate that will be decreased to close to zero.

II. LITERATURE SURVEY

There are numerous research and development trends related to Intrusion Detection System using deep machine learning and deep learning methods. The work involved will be explained in the following way:

The study, [5, introduced the hybrid machine learning technology (decision tree, and support Vector Machine Algorithms) to enhance the precision the machine learning system could offer. Decision Tree algorithm can be employed to categorize the various attacks of various kinds. The support vector machine (SVM) algorithm used to categorize data that is normal. The model was built using NSL-KDD Dataset. The accuracy of the method is 96.4 percent. In [6,] researchers used an algorithm known as the gene algorithm (GA) in conjunction alongside the support vector machines (SVM) to detect intrusion packets. SVM together with GA are used with particular features. For classification and regression issues researchers use SVM. The data utilized in the study was KDD Cup 1999. the detection accuracy was 97.3 percent.

Researchers of [7] devised an algorithm to detect networks using an algorithm called the Recurrent Neural Networks algorithm with NSL-KDD as a data set. The results of the research paper are broken down into binary classifications, with 83.28

ISSN : 2057-5688

percent accuracy, and a multiclass classification which has 81.29 percentage accuracy. The method developed in [8] relies on the convolutional neural network to identify any intrusions into the network. The datasets were generated using the KDD Cup 1999 dataset, two-dimensionalization was performed on test data to determine the accuracy of this model. The detection rate of the algorithm is 97.7 percent. in [9], researchers employed Artificial Neural Network for network intrusion detection using the KDD Cup 1999 dataset .in the process of preprocessing the system used Principal Component Analysis (PCA) to decrease in the quantity of attributes, and also the min/max formulas for normalized data.

The paper discusses Artificial Neural Network Architecture this paper uses Feed Forward Neural Network (FFNN) and Levenberg-Marquardt (LM) Backpropagation as well as mean squared error (MSE) as loss functions. The system's accuracy was 97.97%.In the proposed system, it was a deep neural network constructed by using NSL-KDD data .they suggested using label-encoder and min-max normalization for processing and auto-encoders to build the deep-learning layers. This model was constructed by five different categories .the most accurate detection of the five category was the dos attack. reaching 97.7 percent and 89.8 percent for probe and.

III. PROPOSED METHODOLOGY

The intrusion detection system that is proposed is able to detect attacks using an algorithm called deep neural networks which employs anomaly detection methods without accessing any data within its the payload. To ensure that there is no privacy breach, this system is constructed using datasets through an order of actions. Datasets are the most important element that is used to develop machine-learning algorithms to detect suspicious threats to train machine learning algorithms to identify suspicious. However, the findings from this study suggests that a lot of researchers still use an outdated data set, KDDCup99 and NSL-KDD (a version of the KDD00 dataset) that have been widely criticized for being ineffective and outdated for the present infrastructure for networks. This dataset was developed in 1999, making it nearly two

ISSN : 2057-5688

decades old. Rapid advancement and the evolution of Information Technology, such as the cloud, social media as well as the Internet of Things are changing the way we live network infrastructure landscape. These shifts are impact of changing the nature of threats and attacks itself.

Thus, a lot of research findings which show high accuracy are being considered to be overstated due to the fact that the data being used is not representative of the present threat or the infrastructure.

The KDDCup99 dataset is a well-known dataset used for this years Third International Knowledge Discovery and Data Mining Tools Competition. Each connection is described using 41 attributes (38 continuous or discrete numerical attributes, as well as three symbolic attributes). Each is classified as normal or an attack of a particular kind. These attacks fall in one of four categories: Probe DoS, U2R and R2L. These categories are listed below. Probe: This kind of attack collects details about the system targeted prior to initiating the actual attack. DoS (pronounced "dos"): Denial of Service (DoS) The kind of attack can cause the non-availability of resources on the network to legitimate requests , either by using up all bandwidth or overburdening the computational resources.

Users to Root (U2R) in this particular instance in this instance, the attacker has access an account that is used to login regularly for the user. The attacker is able to exploit weaknesses in security to get access to source of the system.

Remote to Local (R2L) In this case, the attacker does not have an account on remote computers. Instead, they send an email message to the remote computer via a network, and exploits weaknesses for local access by pretending to be an user of the machine.

ISSN : 2057-5688

The dataset NSL-KDD was developed in 2009, and is actually an upgraded version of the KDDCup99 dataset. NSLKDD is a bid to improve KDDCup99 dataset by removing duplicate records, like the inconsistent the number of instances as well as different attack classes [22]. However, it is taken over the main limitation of the dataset.

KDDCup99 has its negatives. The first is that the data was developed in 1999 making use of Solaris as the Solaris operating system to collect diverse data due to its user-friendly nature. But, there are some important differences between the operating systems that aren't even close to Solaris. In the current era of Ubuntu, Windows and MAC, Solaris has almost no market share.

The traffic collector employed for KDD datasets, TCPdump is extremely likely to be overwhelmed and stop sending packets due to an excessive traffic load. In addition, there's some confusion regarding the attack distributions in the datasets. Based on an analysis of attacks, Probe isn't an attack unless the number of iterations exceeds an amount that is a predetermined threshold and label inconsistency is observed.

The third reason is that the rise of new technologies like cloud computing social media, cloud computing along with the Internet of Things has changed the infrastructure of networks in a dramatic way. These changes could bring about new types of threats.

.

IV DISCUSSION:

ISSN : 2057-5688

Protocol type
tcp
tcp
udp
icmp
tcp
icmp

Protocol type tcp	Protocol type udp	Protocol type icmp
1	0	0
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

Fig 1: One hot encoder for the model

		Predicted		total
		Attacks	Normal	
actual	Attacks	1 177 312	207	1 177 519
	Normal	108	291 903	292 011
total		1 177 420	292 110	

Fig 2: Active and passive cases

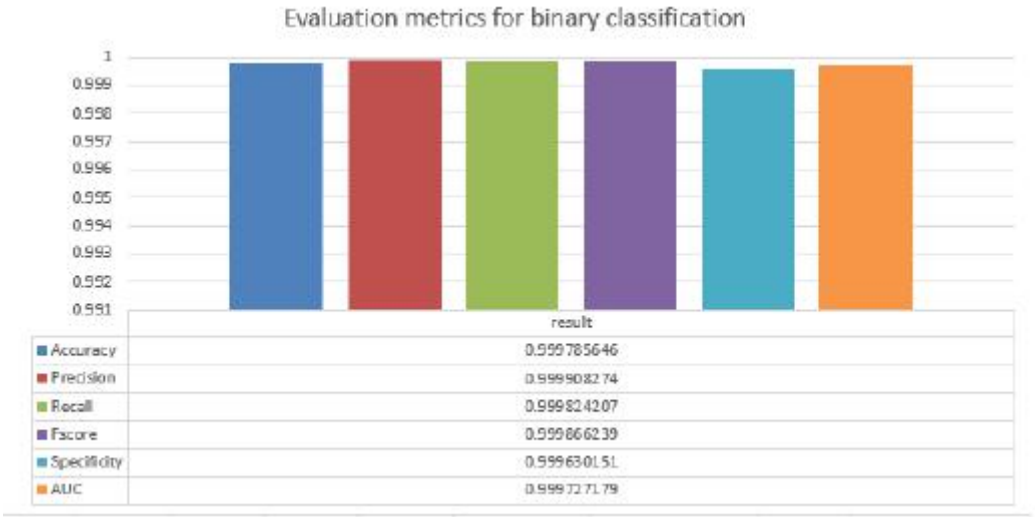

Fig 3: Evaluation metrics for the results

ISSN : 2057-5688

		Predicted					total
		DOS	Probe	R2L	U2R	normal	
actual	DOS	1 165 359	1	0	0	13	1 165 373
	Probe	5	12293	1	0	100	12 399
	R2L	1	0	269	0	79	349
	U2R	0	0	0	0	9	9
	normal	42	7	50	0	291 391	291 490
total		1 165 407	12 301	320	0	291 592	

Fig 4: CLassification on multi class with confusion matrix

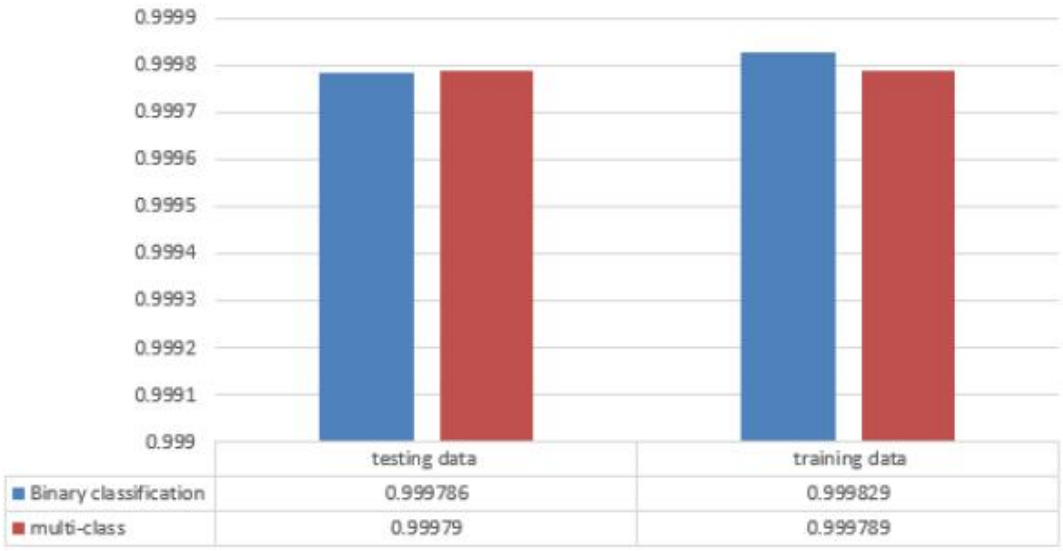

Fig 5: Accuracy of multi-class and binary classification

Soft computing methods are receiving much attention from researchers working in IDS. This is due to the fact that this method is simple to implement and can often yield better results than a one algorithm. A well-balanced mix of many algorithms is the most effective option. Researchers are mostly focused upon the categorization of IDS which can be useful in determining the nature of intrusions. However, it could cause problems in identifying suspicious intrusions, which could contain new or modified intrusion attacks. In order to create an improved IDS it is recommended that a The clustering algorithm is a possibility in the near future for further the future development. KDDCup99 as well as its variant NSL-KDD dataset are two of the most popular datasets even though they're more than 20 years old. The constant change in the data could result in stagnant growth in IDS because of intrusion attacks are

ISSN : 2057-5688

constantly evolving with the latest technologies and user habits. In the end, this will lead to the demise of IDS as an overall cyber security tool.

V CONCLUSION

This paper propose the two types of models (multi-class as well as Binary classification) . Then, we suggested in these models make use of deep learning techniques in the detection of network attack instead of making use of machine learning signatures or rules. In this research, we have demonstrated multi-class classification that was discovered by analyzing KDD cup 99. KDD cup 99, and other data sets. We have shown that the models that are supervised, which are DNN capable in decoding and classifying attacks with high accuracy (99.98 percent) and this was made based on the examination of network packets and connections parameters that do not contain information about the payload packet. Furthermore, the precision of dos attacks detected was high and reached 99.99 percent.

VI REFERENCES

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network anomaly detection: methods, systems and tools,” Ieee communications surveys & tutorials, vol. 16, no. 1, pp. 303–336, 2013.

[2] K. Finnerty, S. Fullick, H. Motha, J. N. Shah, M. Button, and V. Wang, “Cyber security breaches survey 2019,” 2019.

[3] Z. E. Huma, S. Latif, J. Ahmad, Z. Idrees, A. Ibrar, Z. Zou, F. Alqahtani, and F. Baothman, “A hybrid deep random neural network for cyberattack detection in the industrial internet of things,” IEEE Access, vol. 9, pp. 55 595–55 605, 2021.

[4] D. E. Denning, “An intrusion-detection model,” IEEE Transactions on software engineering, no. 2, pp. 222–232, 1987.

ISSN : 2057-5688

[5] G. G. Liu, "Intrusion detection systems," in Applied Mechanics and Materials, vol. 596. Trans Tech Publ, 2014, pp. 852–855.

[6] C. Chio and D. Freeman, Machine Learning and Security: Protecting Systems with Data and Algorithms. " O'Reilly Media, Inc.", 2018.

[7] A. Ali, S. Shaukat, M. Tayyab, M. A. Khan, J. S. Khan, J. Ahmad et al., "Network intrusion detection leveraging machine learning and feature selection," in 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET). IEEE, 2020, pp. 49–53.

[8] S. Shaukat, A. Ali, A. Batool, F. Alqahtani, J. S. Khan, J. Ahmad et al., "Intrusion detection and attack classification leveraging machine learning technique," in 2020 14th International Conference on Innovations in Information Technology (IIT). IEEE, 2020, pp. 198–202.

[9] M. A. khan, M. A. Khan, S. Latif, A. A. Shah, M. U. Rehman, W. Boulila, M. Driss, and J. Ahmad, "Voting classifier-based intrusion detection for iot networks," in 2021 2nd International Conference of Advance Computing and Informatics (ICACIN). Springer, 2021.

[10] L. N. Tidjon, M. Frappier, and A. Mammar, "Intrusion detection systems: A cross-domain overview," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3639–3681, 2019.

[11] A. Shenfield, D. Day, and A. Ayesh, "Intelligent intrusion detection systems using artificial neural networks," ICT Express, vol. 4, no. 2, pp. 95–99, 2018.

[12] M. Rege and R. B. K. Mbah, "Machine learning for cyber defense and attack," DATA ANALYTICS 2018, p. 83, 2018.

[13] S. Latif, Z. Zou, Z. Idrees, and J. Ahmad, "A novel attack detection scheme for the industrial internet of things using a lightweight random neural network," IEEE Access, vol. 8, pp. 89 337–89 350, 2020.

[14] K. Keshari, Top 10 Applications of Machine Learning: Machine Leaning Applications in Daily Life, 2020.

[15] R. Sober, Data Breach Response Times: Trends and Tips, 2020.

ISSN : 2057-5688

[16] A. Shafique, J. Ahmed, W. Boulila, H. Ghandorh, J. Ahmad, and M. U. Rehman, "Detecting the security level of various cryptosystems using machine learning models," algorithms, vol. 1, p. 5, 2021.