International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195264

A Survey on Intrusion Detection System using Machine Learning and Deep Learning

Hemavati*, Dr. Aparna R
Department of Information Science and Technology, SIT, Tumakuru, Karnataka, India

ABSTRACT

As we know internet of Things (IoT) is one of the fastest growing paradigm which is composed of Internet and different physical devices with different domains or the smart applications like home automation, business automation applications, health and environmental monitoring applications. The dependency on IOT devices is increasing day by day with our daily activities, which leads to most important challenge for security. Since having a better monitoring system for better security is a need. From more than two decades the concept or the frame work called IDS (Intrusion detection system) is playing important role for detecting the attacks in the network. Since the network attacks are not fixed in nature, a new type of attacks are happening on the network applications. There are many traditional IDS techniques are available but they are complex to apply. Since machine learning is one of the important area which is achieving good results in many applications. In this paper we study about the different machine learning techniques used till now and the methodology for the attack detection and the validation strategy. We will also discuss about the performance metrics.

Keywords :- IOT, IDS, deep learning, machine learning.

I. INTRODUCTION

As we know from the history humans are always into finding the new innovations and adopting those into their daily life for the betterment of leading the life. With the hungry of the new innovations the evolution of many new technologies has happened. Considering some of them are sensors, actuators, embedded computing, cloud computing, Internet of things (IoT) and many more. Internet of Things (IoT) transforms the networks communication from device to device based communication and provides an accessibility for the resources through the Internet. It expands the edge of the Internet and from the cloud to fog computing.

When we say network, the internet and the IOT, the cybersecurity come into picture. Cybersecurity is composed of different technologies and processes designed to protect all hardware and software devices. A network security system consists of a network security system and a computer security system. Each of these systems includes firewalls, antivirus software, and intrusion detection systems (IDS).

IDS helps to discover, determine and identify unauthorized system behaviour such as use, copying, modification and destruction [2]. Security breaches include external intrusions and internal intrusions.

II. RELEVANT CONCEPTS

1. Intrusion Detection System

Intrusion detection is the activity of detecting actions that intruders carry out against information systems. These actions, known as intrusions, aim to obtain

unauthorized access to a computer system. Intruders may be external or internal. Internal intruders are users inside the network with some degree of legitimate access that attempt to raise their access privileges to misuse non-authorized privileges. External intruders are users outside the target network trying to gain unauthorized access to system information (Vacca, 2013; Patel et al., 2010). A typical IDS is composed of sensors, an analysis engine, and a reporting system. Sensors are deployed at different network places or hosts. Their task is to collect network or host data such as traffic statistics, packet headers, service requests, operating system calls, and file-system changes. The sensors send the collected data to the analysis engine, which is responsible to investigate the collected data and detect ongoing intrusions. When the analysis engine detects an intrusion, the reporting system generates an alert to the network administrator. IDSs can be classified as Network-based IDS (NIDS) and Host-based IDS (HIDS).

Network-based IDS (NIDS) connects to one or more network segments and monitors network traffic for malicious activities. Host-based IDS (HIDS) is attached to a computer device and monitors malicious activities occurring within the system. Unlike NIDS, the HIDS analyses not only network traffic but also system calls, running processes, file-system changes, inter-process communication, and application logs. IDS hes may also be classified as signature-based, anomaly based or specification based. Since these categories are part of the taxonomy proposed in this paper, more details about them will be provided in [4].

2. IoT

IoT is a concept that gathers all sorts of different applications based on the convergence of smart objects and the Internet, establishing an integration between the physical and the cyber worlds. These applications may range from a simple appliance for a smart home to a sophisticated equipment for an industrial plant. Although IoT applications have very different objectives, they share some common characteristics. Generally speaking, IoT operations include three distinct phases: collection phase, transmission phase, and processing, management and utilization phase (Borgia, 2014).

In the collection phase, the main objective is to collect data of the physical environment. Sensor devices and the corresponding technologies (knowledges) for short range communication are combined to reach this goal. Devices of the collection phase are usually insignificant and resource constrained. Different technologies and communication protocols for this phase are designed to operate at short distances and limited data rates with limited or constrained memory capacity and low energy consumption. Due to these characteristics, collection phase networks often are referred to as LLN (Low-power and Lossy Networks). Solutions for error control, medium access control, routing and addressing in LLNs may be different from those used on the conventional Internet. During the transmission phase it transmits the data collected during the collection phase to applications and, consequently, to users. The technologies such as WiFi, Ethernet, Hybrid Fiber Coaxial (HFC) and Digital Subscriber Line (DSL) are combined with TCP/IP protocols to build a network that interconnects objects and users across longer distances. Accesses are necessary to integrate LLN protocols of the collection phase with conventional Internet protocols employed in the transmission phase. The applications process and collect the data to obtain useful information about the physical environment in the processing, management and utilization phase. These applications may take decisions based on this information, controlling the physical objects to act on the physical environment. This phase also includes a middleware, which is responsible for facilitating the

integration and communication between different physical objects and multi-platform applications.

3. Machine learning

Machine learning is the concept composed of statistical models, algorithms which are used by the computer system for improving the performance on a particular task. The ML algorithms builds a strong mathematical models on a sample data which we call it as training data in order to make decisions or a predictions. Nowadays ML concepts are used in almost all applications and its achieving better results such as intrusion detection, email filtering, computer vision etc.,

III. INTRUSION DETECTION METHODS

There are three main types of network analysis for IDSs: 1. misuse-based, also known as signature-based, 2. anomaly-based, and 3. hybrid.

Misuse-based detection techniques aim to detect already known attacks by using the signatures of these attacks [3]. They are used for known types of attacks without generating a large number of false alarms. However, administrators frequently must manually update the database rules and signatures. New (zero-day) attacks cannot be detected based on misused technologies. Anomaly-based techniques study the normal network and system behaviour and able detect anomalies as deviations from normal behaviour. They are interesting because of their capability to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, therefore making it difficult for attackers to know which activities they can perform undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviours can be categorized as anomalies.

Hybrid detection combines misuse and anomaly detection [4]. It is used to increase the detection rate of known intrusions and to reduce the false positive rate of unknown attacks. Most ML/DL methods are hybrids.

Network security data set

Data constitute the basis of computer network security research. The correct choice and reasonable use of data are the prerequisites for conducting relevant security research. The size of the dataset also affects the training effects of the ML and DL models. Computer network security data can usually be obtained in two ways: 1) directly and 2) using an existing public dataset. Direct access is the use of various means of direct collection of the required cyber data, such as through Win Dump or Wireshark software tools to capture network packets. This approach is highly targeted and suitable for collecting short-term or small amounts of data, but for long-term or large amounts of data, acquisition time and storage costs will escalate. The use of existing network security datasets can save data collection time and increase the efficiency of research by quickly obtaining the various data required for research.

The Features of NSL KDD dataset [2]

1	duration: continuous.
2	protocol_type: symbolic.
3	service: symbolic.
4	flag: symbolic.
5	src_bytes: continuous.
6	dst_bytes: continuous.
7	land: symbolic.
8	wrong_fragment: continuous.
9	urgent: continuous.
10	hot: continuous.
11	num_failed_logins: continuous.

12	logged_in: symbolic.
13	num_compromised: continuous.
14	root_shell: continuous.
15	su_attempted: continuous.
16	num_root: continuous.
17	num_file_creations: continuous.
18	num_shells: continuous.
19	num_access_files: continuous.
20	num_outbound_cmds: continuous.
21	is_host_login: symbolic.
22	is_guest_login: symbolic.
23	count: continuous.
24	srv_count: continuous.
25	serror_rate: continuous.
26	srv_serror_rate: continuous.
27	rerror_rate: continuous.
28	srv_rerror_rate: continuous.
29	same_srv_rate: continuous.
30	diff_srv_rate: continuous.
31	srv_diff_host_rate: continuous.
32	dst_host_count: continuous.
33	dst_host_srv_count: continuous.
34	dst_host_same_srv_rate: continuous.
35	dst_host_diff_srv_rate: continuous.
36	dst_host_same_src_port_rate: continuous.
37	dst_host_srv_diff_host_rate: continuous.
38	dst_host_serror_rate: continuous.
39	dst_host_srv_serror_rate: continuous.
40	dst_host_rerror_rate: continuous.
41	dst_host_srv_rerror_rate: continuous.

Four Major Types of attacks

DOS	Probe	R2L	U2R
back	satan	warezmaster	Rootkit
Neptune	portsweep	warezclient	Butteroverflow
smurf	Ipsweep	Ftpwrite	Loadmodule
tear drop	nmap	guesspassword	Perl
land	Imap		
pod	Multihop		
	phf		
	Spy		

IV. METHODS AND MATERIAL

There are number of approaches based on machine learning, the algorithms such as Support Vector Machine, KNN, Decision tree ANN, Random Forest and so on[4] have been used and achieved success for IDS.

SVM

As we know the SVM is one of the most accurate and robust algorithm in machine learning. The method SVC is basically focus on decision boundaries, which are used to separate the set of instances with respect to class values among two groups. The SVC is mainly used to find out optimal separation hyperplane, during the classification the mapping input vector of features which are located on hyperplane side will fall into one class and the positions fall in to other class.

There are many authors who have used SVM like [7], [8], they got the validation accuracy and classification accuracy as 89.85% and 99.9% respectively.

K-Nearest Neighbour

KNN is basically a classification algorithm based on distance function that measures the similarities between two data points or instances. The standard Euclidian distance d(x,y) function with x,y points is defined as

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Many authors have used KNN algorithm for intrusion detection classification, some of them are listed below.

In [11] initial the data has been pre-processed by PCA to select best features then by using k-means

algorithm they have created clustering centers and labels. These clusters and labels are used to classify the data using KNN. The average accuracy achieved is 90% and they have not considered precision and recall rate.

One more [12], used KNN for IDS on KDD Cup'99 dataset. The improved result of classification accuracy with mixed different algorithms is 98.55%.

Decision tree

Decision trees are a technique from data mining that categorize new pieces of information into a number of predefined categories. Decision trees use a pre-classified dataset to learn to categorize data based on existing trends and patterns. After the tree is created, the logic from the decision tree can be incorporated into a number of different intrusion detection technologies including firewalls and IDS signatures.

IN machine learning the decision tree is considered as a predictive model where is represents the mapping of object attributes and values. Every node in the tree indicates an object and an edge is considered as a possible value. The decision tree will be having single output. By establishing different independent decision trees we can able to handle complex output. There are some commonly used decision trees are C4.5 and CART.

In [9] proposed IDS based on decision tree for NSL KDD dataset. They have selected 14 features by using CFS (correlation feature selection) approach and the result achieved for overall accuracy was 83.7% and 90.3%. FAR was 2.5%, the result could still be improved.

In [10] used C4.5 algorithm and C4.5 decision tree with pruning on KDD cup99 and NSL KDD dataset. Only the discrete values attributes are considered for classification. The result shows that 98.45 and 1.55 as precision and FAR respectively.

Deep learning

Deep learning is one of the important area for the research in artificial intelligence. Its main idea is to form a neural network as imitation to the human brain for analytical learning such as computer vision, image processing, text processing etc.
Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

The differences of machines learning and deep learning lies in the following points
1. Data dependencies
2. Hardware dependencies
3. Feature processing
4. Problem solving method
5. Execution time
6. Interpretability

Both machine learning and deep learning adopts same methodology for processing the data but slightly differs in feature extraction which is automated in deep learning rather than manual in machine learning.

Performance evaluation

$$\text{Acc}=\frac{TP+TN}{TP+TN+FP+FN}$$

TP: True Positive (No.of attacked records classified correclty)
TN: True Negative (No.of normal records classified correctly)
FP: False Positive (No.of records mis classified)

FN: False Negative (No. Of assault Records inaccurately arranged)

Issues, concerns and future research directions

this survey examines that the intrusion detection system based on machine learning and deep learning have got many imbalances shows some problems like there are very limited datasets are available and the methods of extraction is different, the performance metrics for evaluation are not same many of them asses only accuracy and the FAR. Most of the research have only statically tested which are not adopted in real time of the actual network.

The usage of deep learning has become help full in handling large dataset and complex learning but some fine tuning is required in order to experience the interpretability.

These are the some problems may provide future directions for the researchers.

V. CONCLUSION

In this paper we presented the literature review for the intrusion detection system with machine learning methods and little overview of deep learning. Since there is significant growth in the field of machine learning with respect to all computer science field, but it is difficult to select a right method to implement an IDS over other methods. The most important thing is to be considered here the data set. The machine learning and the deep learning methods will work better if we have proper datasets for training and testing the models.

VI. REFERENCES

[1]. S.Revathi and A.Malathi, "A detailed analysis on NSL- KDD dataset using various machine learning techniques for intrusion detection," Int. J. Eng. Res. Technol., vol. 2, pp. 1848 - 1853, Dec. 2013.

[2]. R. R. Reddy, Y. Ramadevi, and K. V. N. Sunitha, "Effective discriminant function for intrusion detection using SVM," in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Sep. 2016, pp. 1148-1153.

[3]. N.Paulauskasand J.Auskalnis, "Analysis of data preprocessing in fluence on intrusion detection using NSL-KDD dataset," in Proc. Open Conf. Elect., Electron. Inf. Sci. (eStream), Apr. 2017, pp. 1-5.

[4]. R. K. Sharma, H. K. Kalita, and P. Borah, "Analysis of machine learning techniques based intrusion detection systems," in Proc. Int. Conf. Adv.Comput., Netw., Inform., 2016, pp. 485-493.

[5]. A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating computer intrusion detection systems: A survey of common practices," ACM Comput. Surv., vol. 48, no. 1, pp. 1-41, 2015.

[6]. E. Viegas, A. O. Santin, A. França, R. Jasinski, V. A. Pedroni, and L. S. Oliveira, "Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems," IEEE Trans. Comput., vol. 66, no. 1, pp. 163-177, Jan. 2017.

[7]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255-260, 2015.

[8]. M. V. Kotpalliwar and R. Wajgi, "Classification of attacks using support vector machine (SVM) on KDDCUP'99 IDS database," in Proc. Int. Conf. Commun. Syst. Netw. Technol., 2015, pp. 987-990.

[9]. D. Moon, H. Im, I. Kim, and J. H. Park, "DTB-IDS: An intrusion detection system based on decision tree using behavior analysis for preventing APT attacks," J. Supercomput., vol. 73, no. 7, pp. 2881-2895, 2017.

[10]. S. Jo, H. Sung, and B. Ahn, "A comparative study on the performance of intrusion detection using decision tree and artificial

neural network models," J. Korea Soc. Digit. Ind. Inf. Manage., vol. 11, no. 4, pp. 33-45, 2015.

[11]. M. Nadeem, O. Marshall, S. Singh, X. Fang, and X. Yuan, "Semi-supervised deep neural network for network intrusion detection," in Proc. KSU Conf. Cybersecur. Educ. Res. Pract., Oct. 2016, pp. 1-13.

[12]. W. Meng, W. Li, and L.-F. Kwok, "Design of intelligent KNN-based alarm filter using knowledge-based alert verification in intrusion detection," Secur. Commun. Netw., vol. 8, no. 18, pp. 3883-3895, 2015.

Cite this article as :

Hemavati, Dr. Aparna R, "A Survey on Intrusion Detection System using Machine Learning and Deep Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 264-270, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195264
Journal URL : http://ijsrcseit.com/CSEIT195264