TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

Host-Based Intrusion Detection and Prevention System based on Machine Learning Algorithms

S. Maheswari, K. Arunesh

Department of Computer Science, Sri S Ramasamy Naidu Memorial College, Sattur, Virudhunagar, Tamilnadu, India.

Article Info
Volume 82
Page Number: 9412 - 9417
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 10 February 2020*

Abstract:
An intrusion detection system gathers, analyzes packets and generates an alert which reports the security violations to the system analyst. Host-based Intrusion Detection and Prevention System (HIDPS) track intrusions from the host side and works for intrusion detection and prevention. IDS face many challenges regarding performance accuracy, speedup and time consumption. Due to the complexity of the network, more numbers of alerts are generated which becomes unmanageable by the system analyst. Network Intrusion Detection often faces challenges in constructing classifiers that could handle the distribution of attack categories in KDDCup 99 datasets. In the implementation of HIDPS, there are different techniques used. The main aim of this paper exhibits a mechanism for HIDPS. This paper also computes the algorithmic complexity of various techniques used by the Host-based Intrusion Detection System. The time complexity is calculated to show the run time taken by each algorithm and space complexity concentrates on the space and the auxiliary space taken by the intrusion detection process while executing. Review on different host-based intrusion detection techniques is made and comparisons between these methods are done based on the Space complexity and the time complexity. The results can be utilized to select apt IDS for the required application.

Keywords: *HIDS, Intrusion, Signature, Anomaly, Malicious.*

I. INTRODUCTION

An Intrusion Detection System (IDS) keeps track of the network or system and alerts the system analyst if any malicious activities and access violations occur [1]. An intrusion may be any unauthorized attempt to access the private data for which the intruder does not have access rights. The core objective of this system is to avoid intruder from accessing the protected data. IDS continuously analysis the network traffic and detects the intrusions. IDS are derived by two types: One is a Network-based Intrusion Detection System (NIDS) and another one Host-based Intrusion Detection System (HIDS). The NIDS monitors several hosts or devices interconnected by a network and inform the analyst if any attack is identified. NIDS includes plenty of sensors to monitor and examine every packet and frame in real-time traffic to detect the intrusions. These sensors deployed in two types: Inline mode and Passive mode. In Inline mode, a sensor is deployed in inline networks so that the network traffics are monitored the data pass through it. The inline sensor is also able to block the attack as soon as it occurs [2]. In Passive mode, Sensor are not deployed actual network or direct pathway of data pass through, instead of the actual network they placed in a copy of the actual network traffic; original traffic does pass through the sensor.

HIDS monitors and analyzes individual hosts or devices, on which intrusion detection system runs. Initially, HIDS is deployed only to monitor but now most HIDS have the ability to prevent malicious activities and access violations on the host system by configuring the baseline of the system. Two

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

techniques are used to detect intrusion in IDS: Anomaly-based IDS and Signature-based IDS. Anomaly-based IDS monitors the network for intrusions based on the behaviour of the system, if it does not match with the behaviours specified by the analyst, alerts will be generated. Signature-based IDS monitors the network packets and compares them with the signatures or patterns of previous attacks which are stored in the database. There are many successful and effective technologies available to detect intrusions [3]. However, a common problem of IDS is the large numbers of alerts they generate false-positive alerts. Many methods have been proposed to minimize these alerts.

In the current development of technology, there are many devices; it ranges from Smartphone to supercomputer. Any device needs IDS, Implementation on the major factor that affects the device is how much space an intrusion detection system requires to run. The space complexity computes the system space used by various host-based intrusion detection system techniques. The earlier research on space complexity states, a computation is one in which, at all times, the memory state of the computation at any point of time can be reconstructed [4]. The time complexity shows the computational time taken by a program for a given input. Calculating complexity of an algorithm helps to know which algorithm is suitable for which device. For most objectives of Intrusion Detection and Prevention System is ensuring the availability, credibility and integrity of secure data systems. This can be achieved by tracking malicious activities, intrusions and attacks, Tries to prevent the computer system and the resources from such similar incidents.

INTRUSION DETECTION AND PREVENTION SYSTEM

Intruders attempt to gain the access rights of privileged users and exploit the system [5]. External intruders are unauthorized users and Internal intruders who have permission to access the system but not for the entire system, this is they have a limited privilege on the system. Intrusion is an activity that attempts for breach integrity, credibility and availability of a system. Classification of IDPS is shown in the following figure,

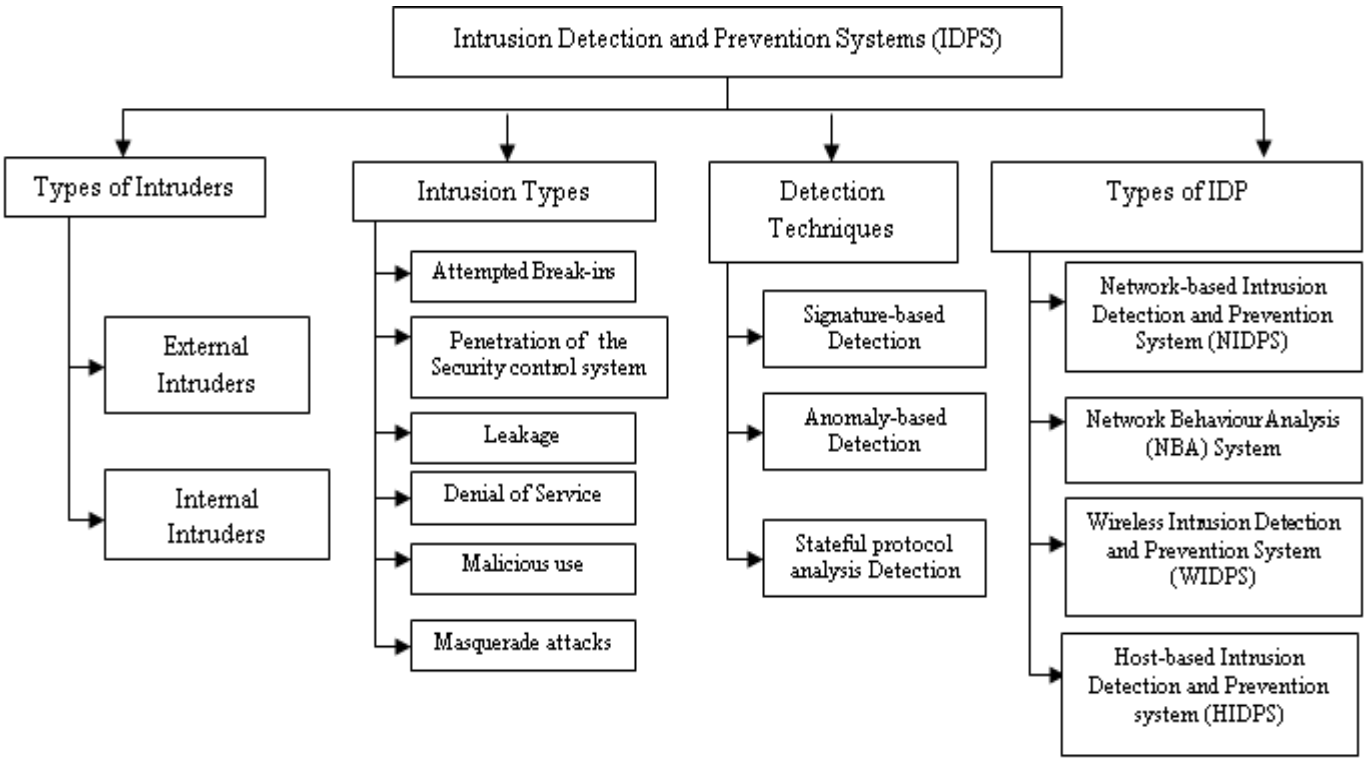

Figure 1. Classification of Intrusion Detection and Prevention Systems

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

A. Approaches to HIDS

1. File System Approaches are one of the commonly used approaches for host-based intruder detection. According to the file system approach, it can store most attack traces on OS and this storage is permanent. Hence it turns out to be an effective and easier one [6]. There are many systems proposed using the file system approach technique. In FWRAP attempted to minimize the mimicry attack and proposed a method that attempts to detect by monitoring the trace of events on the permanent store.

2. System call approach is another method used in HIDS; the main thing is to apply the approach on the kernel of the system, using the kernel level calls in order to reflect the essential activities hidden in all major programming languages, which helps to understand the processing anomaly and also the behaviour [7]. The normal system calls are, login, fork, read, write, execute. These are the system calls used for file access.

3. Attack Graph technique is a method where an attack graph is drawn based on the graph the intrusions are predicted [8]. Generally, there are two vertices in the graph. Among the two, one represents the privileges of the attacker could obtain by an outbreak of the vulnerabilities in the system. Another represents the step which leads to the privilege. Arcs connecting these vertices show the logical connection between the steps of the attacks.

4. Genetic algorithm is another approach that infuses the biological concept behind the genes and the genetic technology [9]. This concept generally uses the danger theory and the dendritic cell theories, it is already used the immunity system. Algorithms presented in the more conceptual, numerical simulations, all the concepts mainly work based on the three agents Ag agent, DC agent and TC agents, and they coordinate and communicate regarding the intrusions in the system.

PROPOSED MODEL FOR HIDPS

Host-based Intrusion Detection and Prevention Systems monitor a variety of host's eventual activity to find any malicious code and intrusion attack in the host such as PC, Mail Servers, Data Servers, Application Servers and File Servers. When malicious code and unexpected behaviors are detected in a system the HIDPS is executed and prevented. The proposed HIDPS model components are discussed as follows,

1. Data Pre-Processing –Data is filtered and segmented.

2. Features Extraction – Packets are decomposed.

3. Selection of Features – features vectors are chosen as input for ML algorithms.

4. Misuse Detection Engine – Process the data to find and compare with previously known attacks.

5. Anomaly Detection Engine – Process the data to compare for normal behavior.

6. Knowledge-based Database – Misuse attacks and alerts are kept in this database that is required by Misuse detection engine.

7. Behavior-based Database – User's Behavior are stored in this database that is required by Anomaly detection engine (ADE).

8. Counter Measure – Blocking and preventing the detected attacks

9. Launch Action – Display warning and tracing the attacks and intruders

10. System Administrator – takes the appropriate action based on the launch action activities

II. RESULTS AND DISCUSSIONS

A) The Distribution of Attack Categories

The standard KDDCup (knowledge discovery and data mining competition) is widely used now a days [5]. The KDDCup 99 dataset consist of 4,94,020 records in training datasets and 3,11,029 records of test datase. The proposed methodology considers

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

12% of the datasets because of the JVM's heap size. The dataset consist of 24 attack types, and additionally 14 types of attacks are included in the test dataset. The attacks fall into four categories of attacks, namely,

1. Denial of Service (DoS)

2. Probing (Probe)

3. Remote to Local (R2L)

4. User to Local (U2L)

Table-I: The Distribution of Various Attack Categories

Attack Categories	Training Dataset Records	Testing Dataset Records
Normal	15,091	17,928
DoS	20,054	14,383
Probe	5,345	20,008
R2L	1,066	2,887
U2R	47	67

B) Performance Evaluation

i) Confusion Matrix

		Predicted Class	
		Normal	Attack
Actual Class	Normal	TP	FN
	Attack	FP	TN

True Positive (TP) – Normal correctly

Classified a Normal

False Negative (FN) – Normal wrongly

Classified as Attack

False Positive (FP) – Attack wrongly

Classified as Normal

True Negative (TN) – Attack correctly

Classified as attack

ii) Detection Rate

The detection rate is defined as the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances present in the test set.

TP

Detection Rate = TP/TP+FP

The proposed work aimed at obtaining the improved detection rate of rare category which was dominated by non-rare categories. The number of records in the dataset where partitioned in orders to reduce the dominating effect of non-rare categories but the Probe category were maintained.

In the following Table 2, the detection rate of a single classifier (NBC) and multiple classifiers (NBC-J48) is shown. The detection rate of Probe category is increased by 2.5%. Upon dichotomizing the dataset, the detection rate of rare attack categories both R2L and U2R is increased by 2.51% and 1.78 % respectively. At the same time the detection rate of Normal and Dos attack category is also improved accordingly. The chart Figure 2 represents the detection rate of Normal and four categories of attack which was obtained from both the single classifier (NBC) and multiple Classifiers (NBC-J48).

Table-II: Detection and False Positive Rate of Single and Cascaded Classifiers

Attack Categories	NBC	NBC-J48
Normal	96.58	98.17
DoS	82.17	91.15
Probe	88.42	90.99
R2L	46.55	49.06
U2R	86.30	88.08

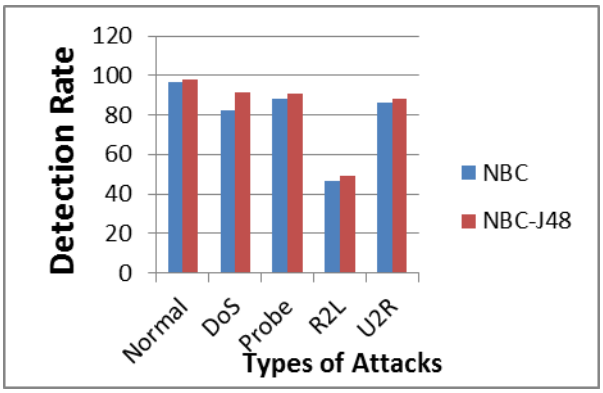

Figure 2. Detection Rate of NBC and NBC-J48

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

ANALYSIS

In the calculation of time complexity and space complexity, the average case has been considered and made a comparison on table 3. Generally, space complexity is the total space taken by any algorithm to compute a given input. The space complexity is mainly based on the input size. The calculation of space complexity includes space and auxiliary space. The time complexity decides how much running time the algorithm takes to run for a given input. The file system approach takes the training set data and finally calculates the false positive and the detection rate [10]. The time complexity is constant since there are no looping structures involved here. The system calls [11] approach mainly concentrates on the traces of the system calls. The system call traces are given as input for further processing. It calculates the threshold value, with that the result is classified either as anomalous or normal. The time complexity of system call approach is quadratic, as the nested looping structure is used for the calculation of threshold through. In the attack graph method [12], the time complexity is a linear one as the comparisons need to be done for the N number of inputs used. The last technique is the genetic algorithm the input is the signal pair and the output signals for the detection of intrusion are generated. The time complexity is calculated to be linear for the genetic algorithm. The information collected and the complexity calculated clearly shows, how the various approaches followed earlier works. From this analysis and comparison, some knowledge gaps are identified. They are as follows:

i. User activity can also be included as a part of IDS processing.

ii. Files can be prioritized for easy detection and for classification too.

iii. CPU overhead can be reduced by triggering the IDS system, rather than running it all the time when the CPU overload reaches the threshold value.

iv. IDS for detecting automated intrusion (Bots) can be included.

The following table explains the technique used and the algorithmic complexity computed for each technique. The criteria considered for the complexity calculation are also listed in the table. By comparing these values, the suitable algorithm for the device can be implemented.

Table-III: Algorithm Complexity Analysis

S.No	Technique used	Time Complexity	Criteria	Space Complexity
1	File System Approach	$O(1)^8$	Based on the input, the condition is checked and the detection rate is calculated if the non-sliding or sliding values.	O(n)
2	System call traces approach	$O(n^2)$	The call traces are been nested looped	O(n)
3	Genetic Algorithm	O(n)	Based on the agent value the output signals are derived	O(1)
4	Attack Graph procedure	O(n)	Since it takes n number of values as the input and it has to go on a linear way by checking each from the attack graph derived	O(1)

III. CONCLUSION

Intrusion Detection systems are useful to prevent the systems/networks from malicious activities. The survey emphasis the use of alert management techniques to minimize the alerts generated by IDS and to improve the detection accuracy. The proposed model in this paper is expecting high security, performance and accuracy.

January - February 2020
ISSN: 0193 - 4120 Page No. 9412 - 9417

Generally, anomaly detection techniques detect attacks such as DOS (Denial-of-Service), probe, User-to-Root, Root-to-Local etc., In signature-based detection techniques, classifier-based and filter-based approaches have been discussed. Also, attack scenario techniques that predict the future intrusions have been reviewed. The summary of intrusion detection techniques is given in table 3 which helps understand the time and space complexity of various algorithms. From this, a conclusion can be made on the algorithm before implementing the HIDS on any device. The time complexity gives the run time complexity of the algorithm and space complexity computes the memory space required for the given input. Any of malpractices and unusual behaviours of internal or external intrusions attacks can be detected and prevented from the systems by HIDPS.

The proposed work aimed at building a network intrusion detection system which improves detection rate of rare categories. Most of the classification approaches where the training dataset involves all attack categories, could not perform well because of the astounding effects of the dominating categories. To overcome this problem the proposed method partitioned the training dataset and classifiers were prepared to handle the rare and other categories. The proposed classifiers NBC and J48 which were coupled to form a multiple classifier. The proposed work has several facilities. (1) The impact of dominant categories alleviated, thus improving the detection rates of the rare categories. (2) The partitioned training dataset has a smaller size, therefore incurs low computational cost for processing and learning by the classifiers. Building an IDS system with the help of the proposed is and superior and uncomplicated.

REFERENCES

1. Chari SN, Cheng P-C., "BlueBox: a policy-driven host-based intrusion detection system. ACM Transactions on Information and System Security", vol. 6(2), 2003, pp. 173-200, May 2003.
2. Tysen Leckie and Alec Yasinsac, "Metadata for anomaly-based security protocol attack deduction", IEEE Trans. Knowl. Data Eng., vol(16), 2004, pp.1157–1168.
3. G. P. Spathoulas and S. K. Katsikas, "Reducing false positives in intrusion detection systems," Published by Elsevier Ltd. Computer Security, vol (29), 2009, pp. 35-44.
4. D. Bolzoni, S. Etalle, and P. H. Hartel, "Panacea: Automating attack classification for anomaly-based network intrusion detection systems," in RAID '09: Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection. Berlin, Heidelberg: Springer Verlag, 2009, pp. 1–20.
5. Hung-Jen Liao et al., "Intrusion detection system: A comprehensive review", Elsevier, Journal of Network and Computer Applications 36, 2013, pp. 16–24.
6. https://en.wikipedia.org/wiki/Intrusion_detection_system#Signaturebased_IDS
7. Bhuyan M, Bhattacharyya D,and Kalita J, "Network anomaly detection: methods, systems and tools". IEEE Communication Surveys and Tutorials Early Access, vol (1), 2013, pp. 1–34.
8. R. Mitchell and I.R. Chen, "Behavior rule based intrusion detection systems for safety critical smart grid applications", IEEE Trans. Smart Grid, vol (4), 2013, pp. 1254–1263.
9. E.-S. M. El-Alfy and F. N. Al-Obeidat, "A multicriterion fuzzy classification method with greedy attribute selection for anomaly based intrusion detection," Published by Elsevier, Proceedia Computer Science, vol (34), 2014, pp. 55–62.
10. Wang et al., "Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networks", Published by Elsevier on Knowledge-based system, vol (70), 2014, pp.103–117.
11. Jabez J and Dr. B. Muthukumar, "Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach", Published by Elsevier, Computer Science, vol (48), 2015, pp. 338 – 346.
12. Fernandes Jr et al., "Autonomous Profile-based Anomaly Detection System Using Principal Component Analysis and Flow Analysis". Applied Soft Computing, vol (34), 2015, pp. 513-525.