Development and Assessment of Intrusion Detection System using Machine Learning Algorithm

Vinod Kumar and Om Prakash Sangwan
School of Information & Communication Technology
Gautam Buddha University, Greater Noida
Gautam Budh Nagar, Uttar Pradesh, India

ABSTRACT

In today's world, the internet is an important part of our life. People cannot think of a single moment without the existence of the internet. With the increasing involvement of the internet in our daily life, it is very important to make it secure. Now to make communication system more secure there is a need of Intrusion Detection Systems which can be roughly classified as anomaly-based detection systems and signature-based detection systems. In the paper we presents a simple and robust method for intrusion detection in computer networks based on Principal Component Analysis (PCA) where each network connection is transformed into an input data vector. PCA is used to reduce the high dimensional data vector to low dimensional data vector and then detection is done in less dimensional space with high efficiency and low use of system resources. We have used KDD Cup 99 dataset for experiment and result shown that this approach is promising in terms of detection accuracy. It is also effective to identify most known attacks as well as new attacks. However, a frequent update for both user profiles and attacks databases is crucial to improve the identification rates.

Keywords

Network Security, PCA, NIDS, Kdd Data Set.

1. INTRODUCTION

Intrusion detection systems can be categories as anomaly-based detection systems and signature-based detection systems. Firstly, an intrusion detection system that learns the normal behavior of the system or the network it monitors is called anomaly-based IDS. This system reports an anomaly when the monitored behavior deviates from the normal profile significantly. On the other hand, a signature-based (misuse) detection approach uses information about the known attacks and detects intrusions based on the matches with existing signatures. Both approaches have pros and cons. Anomaly-based detection can detect zero-day or new attacks, but it suffers from a high false-positive rate and signature based detection has low false-positive rate but works only for known attacks.

As we know that this is the era of internet and people are using internets in their daily life work such as in e-commerce, within enterprise and between enterprises. Internet is the medium for communication between two different organizations and for that purpose they uses network to connect. Many organizations network have been broken into by hackers.

Intrusion detection systems were first introduced by James Anderson [5, 6]. The field did not take off until 1987 when Dorothy Denning published an intrusion detection model [1]. Data collection is the first step for most intrusion detection systems. Now days, these data are generally characterized by their elevated volume, which make it difficult to be analyzed. In fact, most current intrusion detection methods cannot process large amounts of audit data for real-time operations and it seems better to have a new information content of user behaviors, emphasizing the significant features.

IDSs detect computer network behavior as normal or abnormal but cannot identify the type of attacks. This model is designed to identify the normal user profile and attack type of profile and it is also able to detect new type of attack.

The paper is organized as follows. In Section 2 we have presented the Intrusion detection systems in some detail about the different techniques used at the present time. In section 3 an overview of the Principal Component Analysis is discussed. Section 4 describes proposed solution. Simulation results are presented in section 5 and it demonstrates that the proposed solution is better in terms of intrusion detection. Finally, in section 6, we have presented conclusion and future work.

2. TYPES OF INTRUSION DETECTION SYSTEM

There are different types of attacks possible. Attacker can harm to a single machine that is host-based or can harm to whole network that is network-based. So considering these scenario intrusion detection systems can be categories in following types:

a. Host-Based IDS

b. Network-Based IDS

c. Network Behavior Analysis

a. Host-Based IDS:

In a host based IDS the host operating system or the application logs in the audit information. These audit information includes events like the use of identification and authentication mechanisms (logins etc.), file opens and program executions, admin activities etc. This audit is then analyzed to detect trails of intrusion. Host-based IDSs can monitor multiple computers simultaneously [2].

Strengths of Host- Based IDS (HIDS):

- They are good to detect inside attack.
- They are good at attack verification.
- They are capable of decrypting the encrypted packets in an incoming traffic.
- It does not require an additional hardware.

Limitations of Host- Based IDS (HIDS):

- Managing HIDS is not easy.
- They are vulnerable to both direct attacks and attacks against host operating system
- They are vulnerable to denial-of-service attacks
- They can increase the performance overhead
- Unselective logging of messages may greatly increase the audit and analysis burdens.

b. Network-Based IDPS:

They reside on computer or any other appliance which is connected to an organization's network and there it looks for signs of attacks. In an organization's network they are installed at specific place from where it can watch the movement of traffic coming in and going out and whenever a predefined condition occurs it takes an action and notifies the appropriate administrator. It yields many more false-positive readings than host-based IDSs [8].

Strengths of Network- Based IDSs (NIDS):

- With NIDSs we can easily monitor a large network.
- They are usually passive and can be easily deployed to an existing networks with no disruption to the normal network operations
- They are not easily detected by an attacker and hence are less susceptible to direct attack.

Limitations of Network-Based IDS (NIDS):

- Due to large network traffic there may be chances that they fail to recognize attacks.
- They fail to analyze packets which are encrypted
- They do not reliably ascertain whether the attack is successful or not, some forms of attack, specifically those involving fragmented packets are not easily distinguished by NIDSs.

c. Network Behaviour Analysis:

Network behavior Analysis (NBA) works similar to Network-Based IDS however the difference between two is that Network-Based IDS are placed at the boundary between two networks and are responsible for monitoring a particular network segments. However, NBA detects for an attack by monitoring network traffic for any unusual flows or sometimes they detect for any policy or rule violation. They use Anomaly-Based methodology.

Strengths of NBA:

- Their detecting efficiency varies with network behavior.
- Since they use Anomaly-Based methodology they are capable of detecting unknown attacks.

Limitations of NBA:

- It takes time to detect an attack due to network traffic due to such a delay attacks such as Denial of Service remain undetected by NBA.
- Since they use Anomaly-Based methodology they are capable of detecting those attacks which have some effects to the network.

3. PRINCIPAL COMPONENT ANALYSIS

Principal component Analysis is a way to identify patterns in data and expressing the data in such a way to highlight their similarity and differences. The main advantage of this is that once you find out patterns in data, you can reduce the dimension and can compress the data without much loss of information. Principal Component Analysis is a technique that is used to reduce the dimension of data for the analysis of large data and for the compression of data. In this approach basically large number of relatively variables is transform into small number of uncorrelated variables by finding by finding a few orthogonal linear combinations of the original variables with the largest variance. The first principal component of the transformation is the linear combination of the original variables with the largest variance; the second principal component is the linear combination of the original variables with the second largest variance and orthogonal to the first principal component and so on. Mostly in large data set first few principal component contribute maximum variance in the original data set, so remaining principal component can be removed with minimal loss of information [4].

4. FLOW GRAPH OF PROPOSED MODEL

Intrusion detection system can be train by labeled network connection as well as with unlabeled network connection. The proposed model is divided into two parts. In first IDS system is train by labeled network connections and in second part unlabeled connections are projected onto the model and tested.

Flow chart for training the model

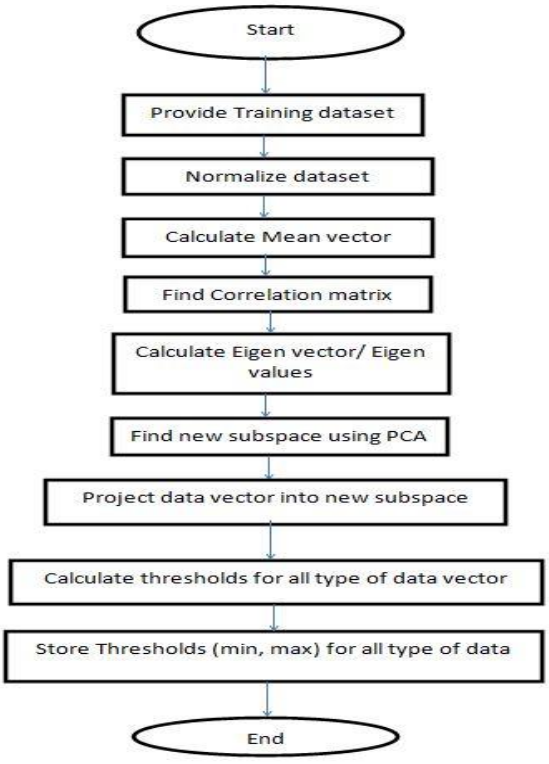

Figure:1 Flow chart for training IDS System

Flow chart for testing:

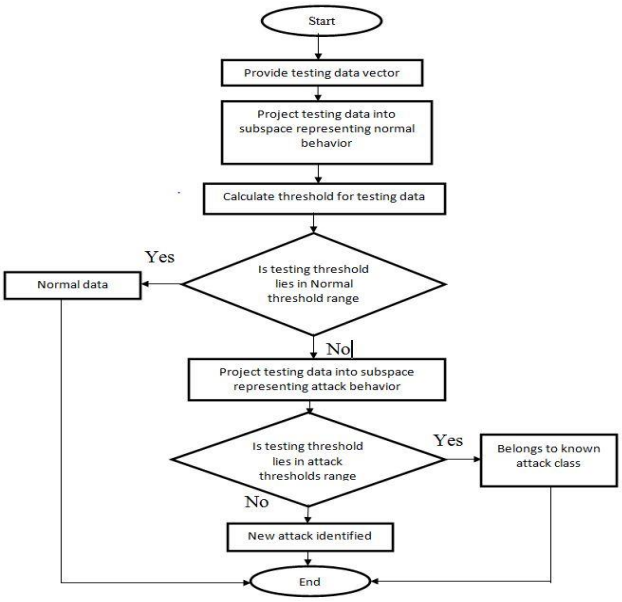

Figure:2 flow chart for testing IDS System

Data collection and analysis:

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 KDD intrusion detection contest uses a version of this dataset, which having a wide variety of intrusions simulated in a military network environment [11]. It has near about 4,900,000 data instances, where each of which is a vector of extracted feature values from a connection record obtained from the raw network data gathered during the simulated intrusions.

Because the dataset was too much large, so for our convenience we choose kddcup.data_10_percent_corrected which has 40950 connections. We can define connection as a sequence of TCP packets to and from some IP addresses. The tcp packets were assembled into connection records using the Bro program modified for use with MADAM/ID [12, 13]. Where each connection is labeled as normal or any specific kind of attack. All labels assumed to be correct.

The simulated attacks belong to one of the following four categories:

- Denial-of-service (DOS)- e.g. a syn flood
- Unauthorized access from a remote machine (R2L) - e.g. password guessing
- Unauthorized access to superuser or root function (U2R) - e.g. buffer overflow attack
- Surveillance and other probing for vulnerabilities (Probing) - e.g. port scanning

There were a total of 24 attack types present in the network connections. And all fell into one of the four categories describe above.

Data Preparation:

First of all, we break the dataset according to their class such that, we put all connections belong to Normal class in single file. Connections belongs to attack classes are kept in separate file like connection belongs to back attack type are kept in a file, connection belongs to Satan are kept in other file and likewise we put all different categories of connections in different files.

Then we break our all dataset into two parts. We used one part for training purpose to our intrusion detection model and second part for evaluation of system. And also we keep some connection untouched that we never used in training duration. At the time of testing we use that connection to check that our system can identify unknown (new attack) attack or not.

5. EXPERIMENTAL RESULTS

To validate our algorithm we had implemented the system into two phases:

Phase 1: In phase 1 we train our Intrusion Detection System for all type of possible attacks individually having large number of connections. Then we test all attacks individually and store there results. Our system was able to identify known attacks as well new attacks.

Table: 1 Testing result for individual type of attack

Attack type	Number in Total	Number for training	Number for testing	IR(%)
Back	2025	1983	102	100
Ipsweep	1194	1101	93	100
Neptune	2627	2500	127	92.91
Nmap	229	200	29	100
portsweep	1040	901	139	100
satan	1534	1433	101	99.01
smurf	233260	233149	111	100
teardrop	977	820	157	100
warezclient	1019	886	133	100
bufferoverflow	29	24	5	100
guesspassword	52	38	14	100
pod	263	200	63	100
ftp_write	7	/	7	70
land	20	/	20	100
load_module	8	/	8	75
multihop	6	/	6	90
rootkit	9	/	9	75

Phase 2: In phase 2 we train our Intrusion Detection System for selected type of possible attacks having large number of connections. We found that it is capable of detecting various attacks either know attacks or unknown attacks in the network and the unknown detection rate was also high.

Table: 2 Testing results for attacks (system train for selected attack classes)

Training data	Number of connections	Testing data type	Number for testing	Identification Rate (%)
Back , ipsweep, neptune, portsweep, smurf, teardop,	240454	phf	3	100
		rootkit	9	45
		perl	2	100
		pod	63	68.25
		teardrop	20	0
		imap	10	20
		loadmodule	8	62.5
		teardrop	7	71.42
		multihop	6	50
		spy	1	100
		back	102	100
		neptune	127	98.24
		teardrop	157	100
		portsweep	139	100
		smurf	111	100
		ipsweep	93	100

6. CONCLUSION

In this paper we have developed an intrusion detection system using principal component analysis to secure network from attacks. We used machine learning technique of dimensionality reduction using principal component analysis. By using PCA we designed a model and implemented it. Our system learns the behavior of connection at training time over training data and at the time of testing it identify known attacks as well as it also identifies new type of attacks.

Extensive experiments are conducted to test our model and to compare with the results of other methods reported in the recent literature. Since in previous studies researcher trained and test their model with selected number of connections according to their convenience but in our study we used testing and training data connection in bulk. In spite of that our model is very much promising in terms of detection accuracy and computational efficiency for real-time intrusion detection in comparison to previous given systems. The model is also effective to identify most individual known attacks as well as new attacks. For the future work, we will develop an online self-adaptive intrusion identification model for updating each individual attack database dynamically and automatically and thus improving the identification rates.

7. REFERENCES

[1] D. E. Denning. 1987. An Intrusion-Detection Model. IEEE transactions on software engineering, Volume : 13 Issue: 2.

[2] Emmanuel Hooper. 2007. An Intelligent Intrusion Detection and Response System Using Hybrid Ward Hierarchical Clustering Analysis, International Conference on Multimedia and Ubiquitous Engineering, in IEEE, 1187-1192.

[3] Guan Xin and Li Yun-jie. 2010. A new Intrusion Prevention Attack System Model based on Immune Principle, International Conference on e-Business and Information System Security (EBISS), in IEEE, 1-4.

[4] I.T. Jolliffe. 2002. Principal Component Analysis, 2nd Edition,Springer-Verlag, NY.

[5] J.P. Anderson. 1972. Computer security technology planning study. Technical Report, ESDTR-73-51, United States Air Force, Electronic Systems Division.

[6] J.P. Anderson. 1980. Computer Security Threat Monitoring and Surveillance. Technical Report, James P. Anderson Company, Fort Washington, Pennsylvania.

[7] Jonathon Shlens. 2009. A Tutorial on Principal Component Analysis. Version 3.01.

[8] R Rangadurai Karthick, Vipul P. Hattiwale and Balaraman Ravindran, 2012. Science Adaptive Network Intrusion Detection System using a Hybrid Approach, Fourth International Conference on Communication Systems and Networks (COMSNETS), in IEEE, pp. 1-7.

[9] Ronald L. Krutz, and Russell Dean Vines. 2010. Cloud Security: A Comprehensive Guide To Secure Cloud Computing, e-book published by Wiley Publishing, Inc., pp. 61-169.

[10] Sodiya, A and Akinwale, A. 2004. A new two - tiered strategy to intrusion detection. Information Management and Computer Security, Volume: 12 Issue: 1, 27-44.

[11] The third international knowledge discovery and data mining tools competition dataset (1999), "KDD99-Cup", available: http://kdi.ics.uci.edu/databases/kddcup99/kddcup99.html

[12] V. Paxson. 1988. Bro: A system for detecting network intruders in real-time, In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX.

[13] W.Lee, S.J. Stolfo, and K. Mok. 1999. Data mining in work flow environments: Experiences in intrusion detection, In Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining (KDD-99).

[14] Zhou, J., Carlson, A and Bishop, M. 2005. Verify Results of Network Intrusion Alerts Using Lightweight Protocol Analysis, Proceedings of the 21st Annual Computer Security and Applications Conference (ACSAC).