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 present 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.
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

![Flow chart for training the model](image-url)
Flow chart for testing:

![Flow chart for testing](image)

**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 has a wide variety of intrusions simulated in a military network environment [11]. It has nearly 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 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. All fell into one of the four categories described above.

**Data Preparation:**

First of all, we broke the dataset according to their class such that, we put all connections belong to Normal class in single file. Connections belong to attack classes are kept in separate files 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 up 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.

**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.
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


