Towards the Development of an Efficient Intrusion Detection System

Samarjeet Borah
Dept. of Computer Science & Engineering
Sikkim Manipal Institute of Technology
Majitar, East Sikkim

Anindita Chakraborty
Dept. of Computer Science & Engineering
Sikkim Manipal Institute of Technology
Majitar, East Sikkim

ABSTRACT

Intrusion is a set of related activities which is performed to provide unauthorized activities such as access to the useful information, file modification etc. It is a set of any actions that attempt to compromise the integrity, confidentiality, or availability of a computer resource. Intrusion Detection Systems (IDS) are used to monitor and detect the probable attempts of such types. An IDS collects system and network activity related data. These data may contain network attacks against vulnerable services, data driven attacks on applications, host based attacks etc. There are several IDSs in literature proposed using various computational techniques such as statistical methods, artificial intelligence, data mining etc. Among these, data mining based methods are comparatively more successful in detecting unknown attack patterns. This paper reviews some remarkable works from the literature along with the basic concepts of intrusion detection. It also includes some suggestions for developing an efficient IDS based on the analysis carried out

Keywords

Intrusion detection system (IDS), Neural networks, Self organizing map (SOM).

1. INTRODUCTION

The Intrusion Detection System (IDS) monitors the events that are occurring in the system or in the network and analyzes them for intrusions. The intrusion detection system is a combination of hardware and software which is used for performing intrusion detection. The intrusion detection is a process of accumulating knowledge related to intrusion which is being performed in the process of monitoring the events and analyzing them for signs of intrusion. The alarms are raised when there is a possible intrusion. The intrusion detection system can be used as a countermeasure which is used to preserve the data integrity from various attacks. Intrusion detection system collects the information from various system and network resources and analyzes this data in order to detect the any attack or intrusion in the system. In addition the intrusion detection system helps the system managers to handle the monitoring the audit and assessment of their system and networks, which is an important part of security management.

1.1 Types of IDS

Intrusion detection systems can be broadly classified into two classes by the scope of protection they are providing. It is also based on the locality of their working environment. These are:

1.1.1 Host Based Intrusion Detection System:

The Host Based Intrusion Detection system detects intrusion on a single computer and performs intrusion based on system calls, kernel and firewall and system logs. They are aimed at collecting information about activities on a particular system or hosts. They are sometimes referred to as sensors and every system consists of a single individual sensor. These sensors collect the data about the various events taking place in the system which is being monitored. This data is recorded by the operating system mechanism called audit trails. Host based sensors keeps track of the behavior of individual user which help in catching an attack while it is being performed or stop a potential attack before it attacks the system. The host based systems are very versatile and have the potential to work in environment that is encrypted as well as over a switched network topology [11].

1.1.2 Network Based Intrusion Detection System:

The Network Based Intrusion Detection system collects the information from the network itself. The intrusion detection system checks for the attack by checking the contents and the header of every packet that is moving over the network. The network sensors uses a method a called packet sniffing where the signature of the attackers stored in the database are compared with the traffic that is being captured which allows the sensors to identify hostile attacks. The network attacks include four main categories: DoS, Probe, U2R, R2L. The network based system monitors the traffic over the specific network and is independent of the operating system that it is installed. The Network Intrusion Detection works on large scale. It monitors and examines the network traffic and based on this determines the traffic as normal or abnormal traffic. The traffic monitoring is performed at firewall, hub and switch etc [11].

2. INTRUSION DETECTION TECHNIQUES

The various intrusion detection systems mentioned in the previous section mainly apply two different techniques for detection of possible intrusions. These are:

2.1 Signature based Detection

The patterns of all the recognizable attacks are described and stored in the intrusion detection system for identifying an intrusion. It is also known as misuse detection. The alarms are generated based on specific attacks signatures. Rule based (or misuse based) detection tries to search for a specific patterns in the data so that it is able to detect the known intrusion effectively.

2.2 Anomaly based Detection

Anomaly detection analyzes and reports the unusual behavioral patterns that are present in computing systems. Here a model of normal system behavior is build from the data observed and then it is distinguished if any significant deviations or exceptions are observed from this model. It implicitly assumes that any deviations from the normal behavior are anomalous. It has the capability of detecting new or unknown behavior. The anomaly based detection generates false positive alarms as it considers every pattern that does not match with normal as anomalous although it may not be. The anomalies can be detected either by using supervised or unsupervised approaches.

2.2.1 Supervised Anomaly Detection:

In supervised anomaly detection the normal behavior model of the system or network is established by training with purely normal dataset. The normal behavior models can be used to classify the new network connections. The system will generate an alarm if the connections are categorized as abnormal. In real practice training of a normal data is not easy and is time consuming. Some of the widely used supervised anomaly detection techniques are k-nearest neighbor algorithm, multi layer perception.

2.2.2 Unsupervised Anomaly Detection:

In unsupervised anomaly detection no prior training is provided. There is no training data or these models may be trained on unlabeled data and try to find intrusions present in the system. Here it is able to detect unknown intrusions without having prior knowledge about it. Some of the widely used supervised anomaly detection techniques are Self Organizing Maps (SOM) [9], Clustering [10].

2.2.2.1 Intrusion detection using Self Organizing Map (SOM)

The self organizing map is neural network model which is used for analyzing and visualizing high dimensional data. The self organizing maps are data visualization technique which reduces the dimensions of data by producing a map of 1-2 dimensions array of neurons. It is a competitive learning network. The audit data is trained using self organizing map. During data training the self organizing map defines a mapping of high dimensional data space into a regular 2-D space. Each neuron represented in the map (suppose i) is associated with n dimensional reference vector where n is the dimension of the input vector. The reference vectors together form the codebook. The neurons of the map are connected to the adjacent neuron by the neighbor relation. In SOM algorithm, the topology and number of neurons remains fixed in the beginning.

In construction of self organizing map the first step is initializing weight. Therefore we assign random numbers to each weight but it should be kept in mind that the magnitude of weight should be small. The next step is similarity matching. With the use of Euclidean minimum distance formulae we calculate the distance from the training data set to the weight vectors. Thus we select any random data from given set and their attributes. These data samples are considered as input at each iteration cycle. From the randomly selected data the weight having shortest distance from data is declared as winner. After the wining vector is found the next step of learning is updation. The weight vectors of the winning neuron along with its neighbors are adjusted towards the input vector. The above process is repeated again and again for fixed period of time or until no changes is seen as in [9].

2.2.2.2 Intrusion detection using Clustering:

Clustering is a process of partitioning the data into meaningful groups or clusters so that all that objects that are present within the cluster have the same characteristics and the objects that belong to other clusters have dissimilar characteristics. It is an unsupervised classification. A good clustering method will be able to produce high quality clusters with higher intra-cluster similarity and lower inter-cluster similarity. There are many types of clustering algorithms such as partition based, hierarchical clustering, model based, density based etc.

The self organizing map is subjected to unsupervised anomaly detection. The similar clusters are grouped into one same class and dissimilar classes are grouped into another class. This detection generally clusters the test dataset into groups of similar instances out of which some may be normal data and some intrusion. Many clustering technique are used. In hierarchical clustering method the clusters are created either by top to bottom approach or bottom to top approach. Other example is K-Means clustering which decreases the overhead of performing the detection over whole datasets. Another method for clustering is fuzzy K-Means clustering where the membership values of each data point corresponding to that cluster are calculated as in [10].

3. LITERATURE SURVEY

Several woks are available in literature on host based intrusion detection systems. Some of the techniques are discussed in the section.

Nadya et al. proposed a clustering algorithm as mentioned in [1] where the features are selected and their maximum and minimum values are found, upper and lower limit are calculated. If the values of the features are within the upper and lower limit a table is formed of true and false. The features are compared with one another in the table and grouped in a cluster. This process continues till there are less than ten percent of the total data present which are grouped in one cluster. Then Euclidean distance is calculated and density is found out. The K-Means clustering algorithm is used in detecting the abnormal data present. Every cluster in K-Means is given a label and percentage of abnormal connections is calculated and the anomaly or normal data is found out.

The system has a benefits- it gives better detection rate for DOS and R2L attacks.

The system has some problems- since we group less than ten percent of the total dataset as one cluster some normal and abnormal data may get mixed or be grouped in same cluster resulting in lower detection rate, the system has low detection rate for Probe and U2R attacks

Kopelo Letou et al. propose a model for host based intrusion detection prevention system as in [2]. The system consists of various components- data preprocessing, feature extraction, feature selection, misuse detection engine, anomaly detection engine, knowledge based database, behavior based database, counter measure, launch action, system administrator. The misuse detection uses C4.5 detection technique and the anomaly detection uses support vector machine algorithm.

The system has a benefit- it expects in providing high security, performance and accuracy.

Manish Kumar et al. as in [3] proposed a model to minimize false positive alert. The Snort IDS was used in the evaluation. The system was fed with sets defining the alarms generated by Snort, set containing alarms generated be partial or exact matching of the signature, set containing alarms generated due to exact matching, set of alarms partially matched. It was found that the minimization of false positive alert can be possible if the set containing the partially matched alarms is reduced to zero. Thus the system by taking the set of alarms minimizes the false positive alert generation.

The system has a benefit- it considers the attack generated using spoofed IP address.

The system has a demerit- the system is not fully able to detect the attacks so that it is not able to reach the final stage. Kusum Bharti et al. proposed a model for intrusion detection using clustering as mentioned in [4] which have high precision and recall rate which is used as performance metric. Recall is used to determine how many misclassifications are made and precision is used for determining how many are correctly classified. True positive, true negative, false positive, false negative are used as measures for measuring the performance. The new methods are introduced for clustering to class mapping which improves the accuracy of the proposed model for all type of attacks.

The system has some benefits - this model gives better results than the K-means clustering over KDD cup 1999 datasets for all types of attacks (Dos, Probe, U2R, R2L).

The system has some problems such as for normal class; K-means clustering gives better results than the proposed model as the number of clusters increases precision of K- means clustering increases.

Vivek et al. proposed a system using standard DARPA dataset as referred in [5]. The SOM is trained by finding the Best Matching Unit (BMU) and updating the codebook vectors and increasing the learning rate. This process continues till the training ends. The system also calculates the mapping precision and the topology preservation. The system is able to detect the intrusion present and thus groups them into three namely normal, intrusion and possible intrusion.

The system has some benefit- it is simple and easy algorithm, topological clustering, works on non linear dataset, excellent dimensional reduction, since SOM is a very powerful mechanism it is able to detect any anomalous intrusion. But the algorithm has a higher time complexity.

Liberios Vokorokos et al. present an intrusion detection system and design architecture of intrusion detection based on neural network SOM referred in [6]. Here a core of the designed architecture represents the neural network Self Organizing Map which classifies the monitored user behavior and also determines the possible intrusion of the monitored computer system. Input to the network represents the multidimensional input vector which describes the actual system state and the activities taken on the controlled computer system. Neural network output is value which represents the possible state of system intrusion.

The main advantage of the system is that it has a sensor which is responsible for detecting intrusion. It is termed as the core element.

The shortcoming of the system is as large numbers of variations are present in data, it is necessary to normalize every input vector.

Peter Lichodzijewski et al. developed a host based intrusion detection system using SOM as referred in [7]. This system uses two methods – the implicit method uses a FIFO or shift register in which any extra event causes the content present to shift along one position and the explicit method provides a timestamp for each event. The implicit method for representing time is found to provide much better separation between user types. The training of SOM is done by dividing it in two levels. In first level dimensions of the map and training period were considered. The output of the first level is given as the input to the second level. In second level nodes were clustered using a clustering algorithm reducing the amount of information that was given to second level SOM.

Some of the benefits of the technique are that the system uses only the session information of the user of a host to detect potential intruders, Self Organizing Map is trained under an implicit coding of data that are demonstrated and provide much clear identification of abnormal behaviors.

Some of the shortcoming of the technique is that when the data set was input to the second level map of the SOM training the neurons were excited by suspicious behaviors patterns which were not clear as the feature in the map was not sharp.

Albert J. Hoglund et al. proposed a host based anomaly detection technique using SOM in [8]. This technique monitors network host users. The system is able to detect something suspicious and it gradually adapts in the user profiles, but it also produces some false positive or false alarms i.e. anomalies which are not intrusions.

The system has many features-adopts to changing data, detects a feature or variable that is out of its reach, detects abnormal feature or variable combinations, can handle different kinds of distribution, the method is multivariate, the method also reports the most abnormal features or variables

The system has several disadvantages- as it is an anomaly detection system, normal changes in the user behavior may lead to false positive, the longer the period of reference of training of user specific Self Organizing Map is required the longer the system adapts to the change in user profiles.

4. ANALYSIS

Several issues were found from the literature survey, which are not being able to address by the developed techniques properly. Some of these are:

- It is seen that, normal user behavior may change sometimes. But, the intrusion detection techniques are not being able to differentiate between attacks and the modified user behavior.
- Feature selection approaches employed varies from techniques to techniques. Some of the techniques stick to some user specified features i.e. year, month, day, hour, minute, second, user name, connection type etc. But these features are only applicable to some specific attacks. It will not provide a standard technique for various types of attack detection.
- Selection of proper classification method (supervised/ unsupervised) may also lead to improper results.
- Generally real world datasets are full of noisy and duplicate values. In this regard, an appropriate normalization technique is always necessary prior to attack detection. But it is seen that most of the techniques are avoiding this step.

4.1 Problem Found

From the literature survey it is found that several techniques have been developed to resolve the issues of host based intrusion detection systems. Some of the major problems found are:

4.1.1 Lack of Relevant Features:

During construction of the map by SOM the nodes were not able to correspond any malicious behavior whether present or not due to the lack of features.

4.1.2 Slower SOM Training Process:

During training if the period of reference used for the user specific SOM is longer, the system adaption to changes slows down.

4.1.3 Generation of False Alarm:

As the system is anomaly based, any slight deviation from the normal user behavior lead to the false alarm generation which lead to an increase in false positive alarm rate.

Table 1. Comparative Study of the Methods

Approach	Nadya et al	Kopelo et al	Manish et al	Kusum et al	Vivek et al	Liberios et al	Peter et al	Albert et al
Time Complexity	Worst as cluster initialization is a lengthy process	Best as it chooses for Misuse Anomaly algorithm	Best as it already generates set of matched and partially matched signatures	Worst as no. of cluster increases the precision rate decreases and recall rate increases	Worst as the no. of neurons affects the performance of algorithm	Worst as normalization of every i/p vector is large time consuming	Worst as the session information of every user the nodes of the SOM are not sharp	Best as the system has automatic anomaly detection component and SOM test for anomaly
SOM Training	None	None	None	None	Best as the SOM is trained by finding the BMU and updating the codebook vectors	Worst as Normalization is done for every user and threshold value is calculated	Worst as the Characteristics Identification of normal user in target host	Best as it has a separate algorithm for SOM training
Input	Best as 13 features out of 42 to reduce the noise, dataset complexity, time complexity	Worst as the dataset extracted selected to misuse detection anomaly detection algorithm	Best as the set of alarms generated due to partially or exact matching signatures is considered	Worst as System log information of every users is considered	Best as the features of dataset are selected using Filter and Wrapper selecting method	Worst as System log information of every user is taken into account	Worst as the whole KDD 1999 dataset is taken	Worst as the Session information of every user is considered
Output	Best as K-Means clustering provide better detection rate	Worst as the system provide minimal protection towards DOS, Probe U2R, R2L	Worst as the system less false positive alert minimization	Best as the clustering method increases accuracy	Best as detection leads to three possible state- normal, intrusion and possible intrusion	Best as states formed defined as intrusion, possible intrusion and normal	Worst as the SOM map of neurons contains suspicious behaviour and features are not sharp	Best as detection of variable out of normal range
Clustering Algorithm	Best as K-means is used for detection	None	None	Best as K-Means is used	Worst as batch training algorithm is used	None	None	None

5. A PROPOSED SCHEME

As the host based systems are vulnerable to attacks, it is necessary to build a system that is able to detect any new type of attacks before it causes any harm to the system. To achieve this, a solution may be monitoring the event logs frequently in hosts. The knowledge of how attacks are being performed, what files are being attacked, along with behavior, location, types of attacks performed are important.

Based on the current scenario an efficient host based intrusion detection technique with better detection rates can be developed by using the following scheme:

- Perform data pre-processing for noise removal and uniformity.
- Perform training of self organizing map according to the features
- Perform unsupervised classification and detection of attacks
- Develop an alert generation process

The above mentioned scheme can be implemented by setting the following objectives:

5.1 Data Preprocessing

The raw data contains many unwanted things. First the raw data is cleaned by removing noise and incomplete data which makes intrusion detection difficult. Data can be converted to unit-less values for smooth computation

5.2 Feature Extraction and Selection

The features are extracted based on various processes such as Principal Component Analysis (PCA) etc. Lastly the features are selected according to the optimal need and it removes the duplicate copies. The features selected may vary from different dataset to dataset. For e.g. the features of event logging may not match with traffic logging.

5.3 Training

The training of the Self Organizing Map is done according to selected features. During training phase some of the techniques used for features such as PCA are done and based on this features extraction the testing are done on them for detection. The features which shows the best results or for which the detection rates increases are considered in for testing.

5.4 Classification

The sample data are then subjected to a process of unsupervised classification i.e. clustering. The clustering algorithm depends on the type of SOM taken into consideration.

5.5 Detection

The sample data is then compared with the trained data for detection of any abnormal behavior or malicious activity if present.

The self organizing map and the clustering can be integrated as follows:

- Use self organizing map is to extract and visually display topological structure of high dimensional input data.
- Apply clustering to partition the input data into groups.
- Integrate clustering and SOM using a 2-level based approach.

In the first approach the SOM is trained. The training of SOM takes place by randomly choosing any data vector ‘x’. Distance from ‘x’ and all prototype vectors are computed and the Best Matching Unit (BMU) or winner is declared, who has the prototype vector closet to “x”. The prototype vectors are updated and the BMU and its neighbors are moved closer to the input vector and input space. This allows different neurons to be trained for different typed data.

The second level SOM takes the output of the first level SOM as the input to the second SOM layer. The no. of neurons on the second map is equal to the desired number of clusters. The second layer of SOM is comparable as clustering of SOM using various clustering techniques/ algorithms. For e.g. Unified matrix or U-matrix is used to visualize the distance between map units and its neighbor and the value is calculated. The merging process of the neighboring neurons are based on various criteria such as intra distance cluster, inter distance cluster etc. Therefore the assessment of the integration of SOM and clustering is the optimal partition the input data.

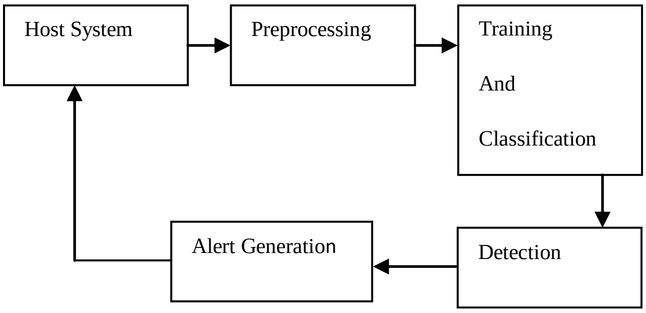

Fig 1: Scheme for a Host Based Intrusion Detection System

6. CONCLUSION

The self organizing map is an extremely powerful tool for the automatic mathematical characterization of acceptable system. Anomaly detection can be done by separately installing different and individual SOM on every network node which has to be monitored. The SOM is a form of unsupervised learning which uses it own intelligence to create a map of near or similar data in order to detect any intrusion present in the network. SOM uses different forms of color to differentiate from the normal to abnormal situations. For e.g. if the nodes are close to each other and are represented by gray color they are similar but if any one of them are black which represent dissimilarity and thus are detected. It is neural network method for the analysis and visualization of high-dimensional data which maps the nonlinear statistical relationships between high dimensional measurement data into simple geometric relationships. The mapping roughly preserves the most important topological and metric relationships of the original data elements and inherently clusters the data. Thus the use of SOM helps us in detecting the intrusion present in the network and if present raising an alarm in order to get into the knowledge of the valid user. The scheme proposed in this paper is under development. Hope results will be found as expected.

7. REFERENCES

[1] Nadya El Moussaid, Ahmed Toumanari, Maryam Elazhari, Intrusion Detection Based On Clustering Algorithm, International Journal of Electronics and Computer Science Engineering, Volume-2, Issue 3, ISSN-2277-1956, 2013

[2] Kopelo Letou, Dhruwajita Devi, Y. Jayanta Singh, Host Based Intrusion and Prevention System (HIDPS), International Journal of Computer Applications (0975 – 8887) Volume 69– No.26, 2013.

[3] Manish Kumar, Dr. M. Hanumanthappa, Dr. T. V. Suresh Kumar, Intrusion Detection System- False Positive Alert Reduction Technique, ACEEE Int. J. on Network Security , Vol. 02, No. 03, 2011.

[4] Kusum Kumari Bharti, Sanyam Shukla, Sweta Jain, Intrusion Detection Using Clustering, IJCCT Vol-1 Issue 2, 3, 4; 2010 for International Conference on Advances in Computer, Communication Technology & Applications (ACCTA-2010).

[5] Mr. Vivek A. Patole, Mr. V. K. Pachghare, Dr. Parag Kulkarni, Self Organizing Maps to Build Intrusion Detection System, 2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 8

[6] Liberios Vokorokos, Anton Balaz, Martin Chovanec, Intrusion Detection System Using Self Organizing Map, Acta Electrotechnia et Informatica No. 1, Vol 6, ISSN 1335-8243, 2006.

[7] Peter Lichodzijewski, A. Nur Zincir-Heywood, Malcom I. Heywood, Host Based Intrusion Detection Using Self Organizing Maps, In the proceedings of the IEEE International Joint Conference on Neural Networks. IEEE 2002, pages- 1714-1719.

[8] Albert J. Hoglund, Kimmo Hatonen, Antti S. Sorvari, A Computer Host based User Anomaly Detection, System Using The Self Organizing Map, IJCNN 2000, Proceedings of the IEEE International Joint Conference on Neural Network, Volume-5, ISBN-0-7695-0619-4

[9] Self Organizing Map, URL: http://en.wikipedia.org/wiki/Self Organizing_map

[10] Clustering –Introduction, URL: http://home.deib.polimi.it/matteucc/Clustering /tutorial_html.

[11] Understanding Intrusion Detection technique, SANS Institute Info Sec Reading Room