Intrusion Detection Techniques for Cloud Computing

K.Swathi[1] M. Ram Gopal[2] V.V. Subrahmanyam[3]
1, 2.Prasad. V. Potluri Siddhartha Institute of Technology, Vijayawada, India
3. Indira Gandhi National Open University, New Delhi, India.

Abstract : Attacks on systems and data are a reality in the world we live in. Detecting and responding to those attacks has become the norm and is considered due diligence when it comes to security. As a matter of fact, most of the standards and regulations applied in the technology space today have explicit instructions regarding the need for monitoring and alerting or intrusion detection. Cloud resources, data and applications are vulnerable for attacks too. So, Intrusion Detection Systems (IDS) are employed in the cloud to detect malicious behaviour in the networks and in the hosts.
In this paper a detailed study on the Network Intrusion Detection Systems (NIDS), Intrusion Detection Systems for Cloud Systems and various techniques available for ID for Cloud systems was done. Finally, it also provides a comprehensive overview of various data sets for Cloud security such as CIDD, KDD CUP'99 data set for researchers to implement and analyze their techniques.

IndexTerms - ***Cloud Security, NIDS, CIDD, KDD CUP'99.***

I. INTRODUCTION

As the growth is recording in large scale in the network based applications as well as the usage of networks in organizations security is a major issue for networks. Even in the social networks or a cloud, network security is always plays a key role. Several defacing mechanisms are exists in this broad area and several researchers are working from past few decades. Firewalls and Intrusion detection systems play major roles in maintaining the network securely. Network Intrusion Detection is an attack defense mechanism in which intruders are identified based on various characteristics of their request in the network.

Due to the availability of vast set of historical data regarding various types of normal as well as attack requests of the users of networks, machine learning algorithms came to existence in the field of intrusion detection system. Several machine learning techniques like pattern recognition, anomaly detection, clustering techniques, support vector machines that can be applied on the above data to find a specific request is legitimate or not.

In the remaining of this paper a detailed description of Intrusion Detection mechanism and some of the issues are presented in section II.. Section III elaborates some techniques available in intrusion detection systems for Cloud systems. Section IV describes various data sets available for the researchers in the field of intrusion detection to implement and test their algorithms. Finally conclusion is presented in section V.

II INTRUSION DETECTION SYSTEM

One of the types of classification of Intrusion Detection System (IDS) is based on the type of data that is used for detection mechanism. Host based intrusion detection System (HIDS) and network based intrusion detection System (NIDS) are two major types in this category of intrusion detection System.

HIDS rely on the information available from various sources like host systems, that includes the contents of operating systems, system and application files. NIDS analyzes the packets that travel across the network by capturing from network communications.

Another type of classification for intrusion detection systems are based on the mechanism applied to detect the intruders. Misuse/Signature-Based Detection and Anomaly/Statistical-Based Detection are two types fall in this classification. The misuse detection mechanism identifies the legitimate users by comparing the available request data with well-known patterns of attacks.

The limitation of this mechanism is that it can identify only the signatures attacks i.e., known vulnerabilities. Anomaly/Statistical Detection: In anomaly based detection system, an observation is made to identify an unusual activity of a network request. Some pattern recognition techniques are applied on the system event streams to find patterns of activity that appear to be abnormal. The major drawbacks of these kinds of systems are that they are very expensive to implement and the false-alarm rate of these kinds of algorithms is high due to lack of sufficient data.

III. INTRUSION DETECTION TECHNIQUES IN CLOUD

Cloud computing provides various services like application and storage to its clients on remote servers. This makes the client of the cloud to free from maintenance. Clients do not have to worry about its maintenance and software or hardware up-gradations. A hypervisor server in cloud data center will host the client services on physical machines and for the clients it visualizes the resources. Flooding attack, Economical Denial of Sustainability attack (EDoS), User to Root attack, Port Scanning Attack, Backdoor Channel attack, attacks on Virtual Machines and Hypervisors etc., are some of the attacks on Cloud systems.

Deploying Cloud Based Intrusion Detection System (CIDS) in hypervisor or host machine would allow the administrator to monitor the hypervisor and virtual machines on that hypervisor. But with the rapid flow of high

volume of data as in cloud model, there would be issues of performance like overloading of VM hosting IDS and dropping of data packets. Also if host is compromised by an offending attack the HIDS employed on that host would be neutralized. In such a scenario, a network based IDS would be more suitable for deployment in cloud like infrastructure. Various techniques as shown below can be used to detect attacks in cloud systems.

A. Machine Learning Algorithms for CIDS

In this section a brief explanation of several machine learning algorithms those can be applied on CIDS are presented.
Figure 1 shows a basic workflow of any CIDS.

- A request packet Network is considered as a new request and the data related to the request packet is captured.
- The captured packet is transformed same as the data samples that are available in the sample patterns which is treated as a knowledge base.
- Now the transformed data is supplied as an input for CIDS which applies any machine learning algorithm.
- The algorithm's goal is to find whether the new arrived packet is the legitimate request or not.
- To fulfill this goal the CIDS applies pattern recognition techniques.
- If the CIDS mapped the request to a normal (legitimate) pattern then the system allows the request into the cloud environment treating it as a legitimate one.
- Otherwise some defense mechanisms were considered and the packet is discarded.

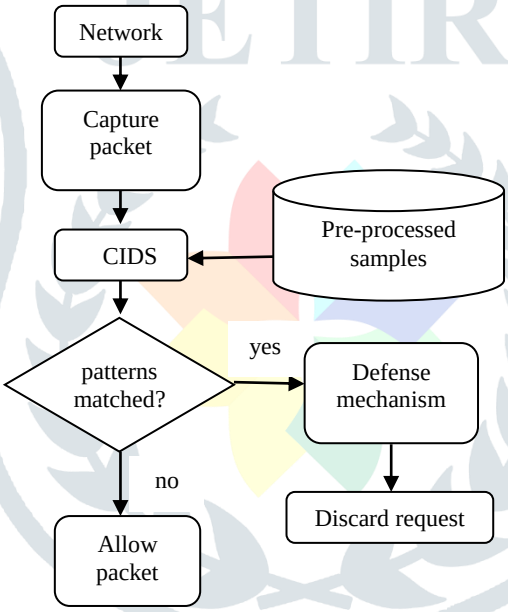

***Figure 1**: Workflow of CID*

B. Classification algorithms

Classification algorithms are the supervised learning algorithms that are based on the class label on the data set. In this learning technique the available samples are categorized as training set and test set. Trained samples are used to build the machine and the test is applied on that machine to find the accuracy of the machine. If the required accuracy is achieved the trained machine is now readyto find unknown (new) requests. Decision tree based methods, Rule based methods, Memory based reasoning, neural networks, Naïve Bayes and Bayesian Belief networks and Support vector machines are examples of classification algorithms. With the help of confusion matrix Accuracy, Recall, Precision and other measures are calculated for the classification algorithms.

C. Clustering algorithms

It is an unsupervised algorithm. When lack of the prior knowledge of the class labels these types of algorithms is applied. In clustering unlabeled samples are grouped so that the inter cluster similarity is maximized and intra cluster similarity minimized. These groups are treated as clusters. Unknown (new) request is marked to a specific cluster based on its nearestness to that cluster. The nearestness can be measured using distances, grids. Root mean square error is the one of the measures of the clustering accuracy. Algorithms try to minimize the mean square error rate in order to increase the performance of clustering algorithms.

D. Neural Networks

Neural networks are algorithmic techniques used to first learn the relationship between the two sets of information, and then "generalize" to obtain new input-output pairs in a reasonable way. Neural networks could theoretically be used in knowledge-based intrusion-detection systems to identify the attacks and seek them in the audit stream. However, as there is currently no reliable way to understand what triggered the association; the neural network cannot explain the reasoning that led to the identification of the attack.

The main reason behind the use of Neural Networks for intrusion detection is its ability to generalize data (from incomplete data) and to be able to classify data as being normal or intrusive [14]. Types of Artificial Neural network which can be used in IDSs can be classified into following three categories [14]: Multi-Layer Feed-Forward (MLFF) neural nets, Multi-Layer Perceptron (MLP) and Back Propagation (BP).

E. Signature Based Intrusion Detection

A Signature based IDS uses a database of rules (signatures) of different attacks known previously. A signature is a pre-determined attack pattern. These signatures are used to compare the incoming network pattern, if the incoming network pattern matches the signature, an intrusion is detected. This type of detection methods has an advantage, that by knowing the network behavior signatures are easy to develop and understand. For example [10], you might use a signature that looks for particular strings within an exploit payload to detect attacks that are attempting to exploit particular buffer-overflow vulnerability. Signature based IDS have very high accuracy in detecting known attacks and minimum number of false positives. The new signatures can be added into the database without modifying existing ones. The main drawback of Signature based IDSs is that these types of IDSs are not able to detect unknown attacks; even a slight variation in the pattern can fool it.

F. Genetic Algorithm (GA)

A Genetic Algorithm (GA) mimics biological evolution to solve the problems. It is based on Darwinian's principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness [1]. GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators [6]. The process usually begins with randomly generated population of chromosomes, which represent all possible solution of a problem that are considered candidate solutions. From each chromosome different positions are encoded as bits, characters or numbers. These positions could be referred to as genes. An evaluation function is used to calculate the goodness of each chromosome according to the desired solution; this function is known as "Fitness Function". During the process of evaluation "Crossover" is used to simulate natural reproduction and "Mutation" is used to mutation of species [6]. For survival and combination the selection of chromosomes is biased towards the fittest chromosomes.

G. Anomaly based Intrusion Detection

Anomaly based IDS uses behaviour based approach. It identifies the event that seems to be malicious as compared to the normal system behavior. It checks the deviation between the normal behavior and current user's behavior. It collects the information of legitimate user's action or behaviour over a period of time. This information is used to train the system. Then a statistical test is performed to check whether this behaviour belongs to legitimate user's behaviour.Anomaly based IDS can detect unknown or zero-day attacks [11] even though system is not updated [12]. For example [13], suppose that a computer becomes infected with a new type of malware. The malware could perform such type of behaviour like sending large numbers of e-mails and consumption of the computer's processing resources that would be significantly different from the normal system user's behaviour.

H. Rough Sets

Rough Set Theory Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others. Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The in-discernible relation generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an elementary set, and forms a basic granule (atom) of knowledge about the universe.

I. Support Vector Machine Based

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data recognize patterns, used for classification and regression analysis of both linear and nonlinear data. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. SVM is useful in detection of intrusions even with the availability of less sample data. SVM has good generalization ability even with the high dimensional data.SVM requires lesser number of training samples as compared to Neural Network based classifiers. SVM can only be used for binary data.

In the next section, we provide the data sets available for the researchers to test their designed algorithms.

IV. DATA SETS FOR CIDS

Many benchmark data sets are available for researchers to develop, implement and analyze their models for the CIDS. This section elaborates some of these data sets in brief.

A. KDD CUP'99 Data Set

KDD Cup '99 is a bench mark data set is available for the researchers of Intrusion Detection System (IDS) that contain several features and patterns of the network requests[3][4]. Number of researchers are using this data set and applying their algorithms such as Clustering, Classification, regression and other machine learning techniques to improve the accuracy of IDS as well as decrease the computational time.

KDD CUP'99 data set is a raw TCP dump dataon a local area network over a period of nine weeks. A training dataset and test dataset are also available under KDD CUP'99[1]. Total nine weeks of data is divided into seven weeks of data for the training data set containing five million records that covers 22 different types of attacks and two weeks of testing data set that contains about 39 attack types. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test datasets not available in the training datasets.

This data set provides 41 features, one class label and about 9 lakhs of sample data set that includes normal as well as four different classes of attacks. Denial of Service (DOS) attacks, Probing, Remote to Local (R2L) and User to Root are the available classes of attacks in KDD cup data set.

KDD CUP'99 provides about 22 different types of attacks that fall into these classes of attacks based on their characteristics and behavior.

Limitations of the KDD CUP'99 Data Set

The KDD CUP'99 data set [1] contains a large collection of redundant records which may misleads the learning algorithms.

It lacks of equal distribution of the sample data for various classes of attacks and normal samples.

Especially U2R and R2L attack samples are very less when compared with the normal, DOS and probe attack samples that may causes the misclassification of these type attacks.

Several researchers are still working in this area to increase the detection rates for these two classes of attacks.

Darpa'98 uses TCP dump traffic controller that drops the packets because of a quick overload. However, there was no observation to identify these dropped packets.

B. NSL-KDD Data Set

The NSL-KDD data set is a modified version of KDD CUP'99 data set. Many research papers have used NSL-KDD data set due to its advantage over KDD CUP'99 data set[2].

The KDD CUP'99 data set contains about 78% and 75% of duplicated records in train and test data sets respectively. This redundancy may leads the detection algorithms mislead and produce biased results. The redundant records that are present in KDD CUP'99 are eliminated in NSL-KDD data set. Removing of the redundant records leads to reduce the size of the data set.

The availability of different sizes and different formats of NSL-KDD data set provides greater flexibility for the researchers to apply their algorithms.

C. CIDD Data Set

CIDD consists of both network and host audit data. CIDD data correlated according to the user IP address and audit time [7]. The CIDD data is a labeled data which can be used in machine learning algorithms for signature based CIDSs. The audit data of CIDD is divided into 2 categories according to the environment in which the data is generated [7]. Unix Solaris audits and their corresponding TCP dump data is present in the first category. The second category includes Windows NT audits and their corresponding TCP dump data. Each category in CIDD consists of both training and testing data sets.

V CONCLUSION

In this paper a detailed study of the IDs, intrusion detection for clouds, some of the data sets that are available for Cloud Intrusion detection like CIDD, KDD CUP99 and NSL-KDD data sets were discussed. An overview of various Cloud ID techniques was discussed.

REFERENCES

[1]KDD Cup 1999, Available on: http://kdd.ics.uci.edu/databases/kddc up 99/kddcup99.html, October 2007.
[2] MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, CISDA, 2009.
[3] Roschke, Sebastian, Feng Cheng, and ChristophMeinel. "Intrusion detection in the cloud." 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE, 2009.
[4] Vieira, Kleber, et al. "Intrusion detection for grid and cloud computing." IT Professional Magazine 12.4, 2010.

[5] Subashini, Subashini, and VeerarunaKavitha. "A survey on security issues in service delivery models of cloud computing." Journal of network and computer applications 34.1, 2011.
[6] Lo, Chi-Chun, Chun-Chieh Huang, and Joy Ku. "A cooperative intrusion detection system framework for cloud computing networks. Parallel processing workshops (ICPPW), 2010 39th international conference on. IEEE, 2010.

[7] Kholidy, Hisham A., and FabrizioBaiardi. Cidd: A cloud intrusion detection dataset for cloud computing and masquerade attacks, Information Technology: New Generations (ITNG), 2012 Ninth International Conference on. IEEE, 2012.

[8] Modi, Chirag, et al. "A novel framework for intrusion detection in cloud."Proceedings of the Fifth International Conference on Security of Information and Networks.ACM, 2012.

[9] Mohammad SazzadulHoque, Md. AbdulMukit and Md.AbuNaserBikas, An implementation of intrusion detection system using genetic algorithm, International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.
[10] James C. Foster, IDS: Signature versus anomaly detection. http://searchsecurity.techtarget.com/tip/IDS-Signature-versus-anomaly-detection, 2005.

[11] Dotan Cohen, What is a Zero-Day Exploit? http://what-is-what.com/what_is/zero_day_exploit.html, 2007.

[12] Mudzingwa, D.; Agrawal, R, A study of methodologies used in intrusion detection and prevention systems (IDPS), Proceedings of IEEE Southeastcon, pp. 1-6, 2012.

[13] Karen Scarfone and Peter Mell, Guide to Intrusion Detection and Prevention Systems (IDPS), Computer Security Division, Information Technology Laboratory NIST Gaithersburg, 2007.

[14] Ibrahim L M, Anomaly network intrusion detection system based on distributed time-delay neural network, Journal of Engineering Science and Technology, Vo. 5, Issue: 4, Start page: 457, 2010.

[15] Wei Li, Using Genetic Algorithm for Network Intrusion Detection, In Proceedings of the United States Department of Energy Cyber Security Group Training Conference, pp. 24-27, 2004.

[16] Goyal Anup and Chetan Kumar, GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System, not published, Electrical Engineering and Computer Science, North Western University, Evanston, IL, 2007.