ISSN: 2321-7782 (Online)
Volume 1, Issue 5, October 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

An Intrusion Detection System using Data Mining Techniques

Dr. P. Devaraaju
Assistant Professor
Dept. of Computer Science and Technology
S K University
Anantapuramu - AP

Abstract: ***Network security is becoming an increasingly important issue, since the rapid development of the Internet. Network Intrusion Detection System (IDS), as the main security defending technique, is widely used against such malicious attacks. Data mining and machine learning technology has been extensively applied in network intrusion detection and prevention systems by discovering user behavior patterns from the network traffic data. The aim of a network-based intrusion detection system (IDS) is to identify patterns of known intrusions (misuse detection) or to differentiate anomalous network activity from normal network traffic (anomaly detection).There are generally two main methods in intrusion detection: misuse detection and anomaly detection. A major difficulty of any anomaly-based intrusion detection system is that patterns of normal behavior change over time and the system must be retrained. Data mining is one of the technologies applied to intrusion detection to invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behavior patterns. This article reviews the current state of art data mining techniques, compares various data mining techniques used to implement an intrusion detection system.***

I. INTRODUCTION

With the ever-increasing growth of computer networks and emergence of electronic commerce in recent years, computer security has become a priority. The security of a computer system is compromised when an intrusion takes place. Intrusion detection is the process of identifying and responding to malicious activity targeted at computing and networking sources [1]. Intrusion detection is the process of observing the events occurring in a computer system or network and analyzing them for instances which violates related security policies or practices.

An intrusion into a computer system is any activity that violates system integrity, confidentiality, or data accessibility. In order to meet this challenge, Intrusion Detection System is being designed to protect the availability, confidentiality and integrity of critical networked information systems. [9,10]. This connectivity between information systems is spreading worldwide and becoming more and more universal, thus making it increasingly vulnerable to breaches, cyber attacks, and pervasive failures. Because these infrastructures are highly interdependent, attacks on one infrastructure can have serious cascading effects on others resulting in potentially catastrophic damage and disrupt Intrusion detection is an important component of that infrastructure protection mechanism. A major problem for such an IDS is that it can give false alarms in cases where there are modifications in the normal system behavior. The IDS must be capable of adapting to these changes and the user profile must be updated at regular intervals.

Using data mining technology, some useful knowledge can be discovered from network data, and invasive behavior and normal behavior rule base can be established. Then we can divide abnormal acts from much real-time data.

II. INSTRUCTION DETECTION

Intrusion Detection System (IDS) is a system that is responsible for detecting anomalous, inappropriate, or other data that may be considered unauthorized occurring on a network or host. It needs to be accurate, adaptive and extensible. There are generally two main methods in intrusion detection; misuse detection and anomaly detection. Misuse detection is based on knowledge of system vulnerabilities and known attack patterns, while anomaly detection assumes that an intrusion will always reflect some deviation from normal patterns. Misuse Detection is a method that provides the ability to identify intrusions based on a known pattern (signatures) for malicious activity.

Attack fall into following four main classes:

- **Denial of service (DoS) attacks:** Attackers disrupt a host or network service to make legitimate users can not access to a machine, e.g. ping-of-death and SYN flood;
- **Remote to Local (R2L) attacks:** Unauthorized attackers gain local access from a remote machine and then exploit the machines vulnerabilities, e.g. guessing password;
- **User to Root (U2R) attacks:** Local users get access to local machine without authorization and then exploit the machines vulnerabilities, e.g. various "buffer overflow" attacks; and
- **Probes:** It is a category of attacks where an attacker examines a network to discover well-known vulnerabilities. These network investigations are reasonably valuable for an attacker who is staging an attack in future.

There are two useful method of classification for intrusion detection systems is according to data source. Each has a distinct approach for monitoring, securing data and systems. There are two following general categories under this classification:

- **Host-based IDSs (HIDS)** – examine data held on individual computers that serve as hosts. The network architecture of host-based is agent-based, which means that a software agent resides on each of the hosts that will be governed by the system [4].
- **Network-based IDSs (NIDS)** – examine data exchanged between computers. Most efficient host-based intrusion detection systems are capable of monitoring and collecting system audit in real time as well as on a scheduled basis, thus distributing both CPU utilization and network overhead and providing for a flexible means of security administration [4].

III. DATA MINING

With the recent rapid development in KDD, a better understanding of the techniques and process frameworks that can support systematic data analysis on the vast amount of audit data that can be made available. Data mining is a relatively new approach for intrusion detection. Data mining is defined as [3] "the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data". There exist many different types of data mining algorithms to include classification, link analysis, clustering, association, rule abduction, deviation analysis, and sequence analysis. Using these algorithms data mining extracts knowledge from the large data sets by analyzing them and presents it in the intrusion detection model. Data mining generally refers to the process of (automatically) extracting models from large stores of data [8]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database.

IV. DATA MINING TECHNIQUES FOR IDS

The central theme of our approach is to apply data mining techniques for intrusion detection in email system Internet-based. Several types of algorithms [8] are particularly relevant to our research:

Classification: Maps a data item into one of several pre-defined categories. These algorithms normally out-put "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that can label or predict new unseen audit data as belonging to the normal class or the abnormal class [7]. The Classification algorithm is inductively learned to construct a model from the preclassified data set. Each data item is defined by values of the attributes. Classification may be viewed as mapping from a set of attributes to a particular class.

K-Nearest Neighbour:

K-Nearest Neighbour (k-NN) is instance based learning for classifying objects based on closest training examples in the feature space. It is a type of lazy learning where the function is only approximated locally and all computation s deferred until classification. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor. The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function. Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances. Generally it is used for intrusion detection in combination with statistical schemes (anomaly detection). **Decision Tree:**

Decision tree is a predictive modeling technique most often used for classification in data mining. The Decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. The main approach is to select the attributes, which best divides the data items into their classes. According to the values of these attributes the data items are partitioned. This process is recursively applied to each partitioned subset of the data items. The process terminates when all the data items in current subset belongs to the same class. A node of a decision tree specifies an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to a possible value of the

Neural Network (NN):

Neural networks have been used both in anomaly intrusion detection as well as in misuse intrusion detection. For anomaly intrusion detection, neural networks were modeled to learn the typical characteristics of system users and identify statistically significant variations from the user's established behavior. In misuse intrusion detection the neural network would receive data from the network stream and analyze the information for instances of misuse. A NN for misuse detection is implemented [5] in two ways. The first approach incorporates the neural network component into an existing or modified expert system. This method uses the neural network to filter the incoming data for suspicious events and forward them to the expert system. This improves the effectiveness of the detection system. The second approach uses the neural network as a standalone misuse detection system. In this method, the neural network would receive data from the network stream and analyze it for misuse intrusion.

Support Vector Machine:

Support Vector Machines [7] have been proposed as a novel technique for intrusion detection.An SVM maps input (real-valued) feature vectors into a higher-dimensiona feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [6]. Structural risk minimization seeks to find a hypothesis h for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. SVM uses a linear separating hyper plane to create a classifier but all the problems cannot be separated linearly in the original input space. SVM uses a feature called kernel to solve

this problem. The Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; including polynomial, radial basis functions, two layer sigmoid neural nets etc.

Association rule mining:

Association describes relationship between various data records. Association rule mining is one of the most popular techniques within data mining. It acts as a sensor which provides source data for meta-learning like techniques which are at higher level of processing. Association rule mining is a slow process and can be replaced by other techniques like classification, clustering etc. An association rule [2] has two parts, an antecedent (if) and a consequent (then). Association rules are created by analyzing data for frequent if/then patterns and support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database and confidence indicates the number of times the if/then statements have been found to be true. These rules are used for analyzing and predicting the customer behavior.

Clustering:

In this technique, data points are clustered together based on their similarity factors and is often nearness according to some defined distance. Clustering is an effective way to find hidden patterns in data that humans might miss. It is useful for ID as it can cluster malicious and non malicious activity separately. k-means is a clustering algorithm used to cluster observations into different groups of related observations without having prior knowledge about their relationships. Here data is divided in k clusters where k is provided as input.

V. CONCLUSION

Intrusion detection systems are one of the key areas of application of data mining techniques. This paper mainly focuses on various IDS models. Various techniques can be used to implement Intrusion detection system. It uses the known patterns to detect the unauthorized behavior attacks. While the main objective is to design a more systematic and automated approach to intrusion detection, elimination of human intervention is not total. Establishing good security policy and key support structure need to be in place. Continuous maintenance and monitoring with sound system administration practices are amongst the heart of best practices in network security.

With the increasing incidents of cyber attacks, building an effective intrusion detection models with good accuracy and real-time performance are essential. Data mining is relatively new approach for intrusion detection. More data mining techniques should be investigated and their efficiency should be evaluated as intrusion detection models.

References

1. Amoroso EG (1999) Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion.Net Books, NJ
2. Duanyang Zhao, Qingxiang Xu, Zhilin Feng, “Analysis and Design for Intrusion Detection System Based on Data Mining”, 2010 Second International Workshop on Education Technology and Computer Science.
3. R. Grossman, S. Kasif, R. Moore, D. Rocke, and J. Ullman. Data Mining Research: Opportunities and Challenges, A report of three NSF workshops on Mining Large, Massive, and Distributed Data, January 1998.
4. J. Cannady. Artificial Neural Networks for Misuse Detection. National Information Systems Security Conference, 1998.
5. G.V.Nadiammai, S.Krishaveni, M.Hemalatha – “A comprehensive Analysis and study in intrusion detection system using data mining Techniques”. IJCA, Volume 35 –No.8, December 2011.
6. J. Cannady. Artificial Neural Networks for Misuse Detection. National Information Systems Security Conference, 1998.
7. S. Mukkamala, G. Janoski, A. Sung. Intrusion Detection Using Neural Networks and Support Vector Machines. Proceedings of IEEE International Joint Conference n Neural Networks, pp.1702-1707, 2002
8. Valdimir V. N. The Nature of Statistical Learning Theory, Springer, 1995.
9. W. Lee and S. Stolfo. Data Mining Approaches for Intrusion Detection. In proceedings of the 7th USENIX Security Symposium, 1998.
10. W. Lee, S.J. Stolfo, K.W. Mok, Algorithms for Mining System Audit Data, in Proc. KDD, 1999.
11. Ya-Li Ding, Lei Li, Hong-Qi Luo, “A NOVEL SIGNATUR SEARCHING FOR INTRUSION DETECTION SYSTEM USINGDATA MINING”, Machine Learning and Cybernetics, 2009 International Conference on Volume 1, pp.122 - 126, 2009.

12. Zhan Jiuhua, "Intrusion Detection System Based on Data Mining", First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pp.402-405, 2008.

