http://dx.doi.org/10.18090/samriddhi.v7i2.8632

Study of Data Mining Based Approaches For Network Intrusion Detection System

Neeraj Shukla[*1] and Nagendra Yadav[2]

ABSTRACT

In the current era, there is ample knowledge in using Internet in social networks (such as instant messaging, video conferencing, etc.), the field of healthcare, various areas related to electronic commerce, banking, and services several other fields. As computer systems based on the network plays an ever more important in modern society once they have become the target of our enemies and criminals. Therefore, we must find the best way to protect our systems. The security of a computer system is compromised during an intrusion occurs. Intrusion can be defined as "a set of actions that aim to compromise the integrity, confidentiality or availability of a resource," for example, illegally obtain superuser privileges to attack and make out of the system (ie, Denial of Service), etc. The purpose of this document is to provide a study of some works that use data mining techniques to detect intrusions and answer some technical questions. An advance effective idea is discussed in the chapter that will detect intruders in a data storage perspective and integrate data mining and online analytical processing (OLAP) for the purpose of intrusion detection.

Keywords :*Intrusion, IDS, Data Mining.*

1. INTRODUCTION

An intrusion can be defined come "a set of actions that attempt to compromise the integrity, confidentiality or availability of a resource," for example, illegally obtaining root privileges, attacking and making non system out of service (ie, Denial-of-service), etc. These applications used over the Internet need a satisfactory or say high level of security and privacy, and eat. On the other hand, the problem is that our equipment fills low and vulnerable attacks by many threats. The aim of this paper is a survey providing some works that employ data mining techniques for detecting intruders and faced by a few glitches.

An intrusion can be defined as a set of actions that endanger the safety requirements (for example, integrity, confidentiality, availability) of a computer resource / rouge (for example, user accounts, file systems and cores in the system) [1, 2]. Intruders have promoted ourselves if invented innovative tools that support various types of attacks rouge. Therefore, effective intrusion detection methods (ID) have become a necessity to insist to protect our computers against intruders. In general, two types of foin him intrusion detection systems (IDS); Systems misuse detection and anomaly detection systems [1, 2, 3]. Expert Systems (eg IDES computer clock, nidx, P-BEST, ISOA) They use non set of rules to describe attacks, signature analysis (for example, Haystack, Net Ranger, Real Secure, Musig) where attacks fils characteristics captured in the audit trail, state transition analysis (eg, STAT, USTAT and netstat) using diagrams transition states, colored Petri Nets

1.* Neeraj Shukla, Assistant Professor-I, Computer Science & Engineering Department, School of Management Sciences, Lucknow, India, e-mail: neer1990@gmail.com

2. Nagendra Yadav, Electronics and Communication Engg. Department, ATC, Lucknow, India, e-mail : nagendras8@gmail.com

(eg NIT) or reasoning based on cases (eg AutoGuard) [1]. Anomaly detection [4, 5] in contrast to misuse detection, you can identify new attacks. Build behavioral models of normaux rouges (called profiles) and uses these profiles to detect new patterns that deviate significantly from them. These patterns may represent real suspect's intrusions or simply might be new behaviors that need to be added to the profiles. Current anomaly detection systems used Côme multivariate statistical methods and analysis temporelle pair identify anomalies; Examples of these systems fils IDES Nides and emerald. Other anomaly detection systems soi built on the basis of the Expert computer systems Côme clock, wisdom and sense [6], [7].

Table-1 : A comparsion between the two types of intrusion detection

	Misuse Detection	**Anomaly Detection**
Characteristics	Use patterns of well-known attacked (signatures) to identify intrusions, any match with signatures is reported as a possible attack	Use deviation from normal usage patterns to identify intrusions, any significant deviations from the expected behaviour are reported as possible attacks.
Drawback	- False negatives - Unable to detect new attacks - Need signatures update - Known attacks has to be hand-coded - Overwhelming security analysts	- False positives. - Selecting the right set of system features to be measured is ad hoc and based on experience - Has to study sequential interrelation between transactions - Overwhelming security analysts

We conclude that traditional IDS face many limitations. This has led to an increased interest in improving current IDS. Applying Data Mining (DM) techniques such as classification, clustering, association rules, etc, on network traffic data is a promising solution that helps improves IDS. In this paper, we discuss DM approaches for network intrusion detection and suggest that a combination of both approaches has the potential to detect intrusions in networks more effectively.

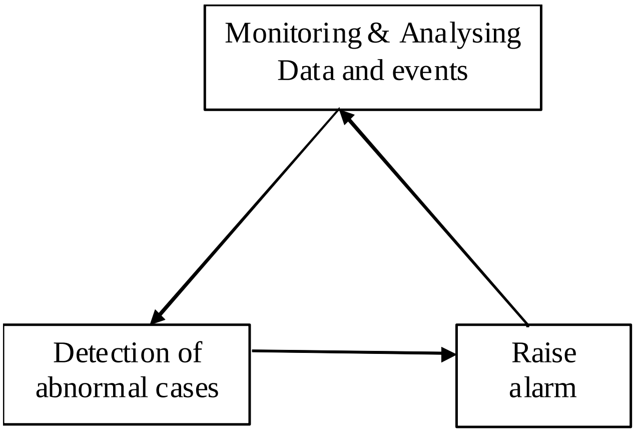
Monitoring & Analysing Data and events
Detection of abnormal cases
Raise alarm

Fig.1: Traditional IDS Framework

Network-based intrusion detection can be broken down into two categories: packet-based anomaly detection and flow-based anomaly detection. Flow-based anomaly detection tends to rely on existing network elements, such as routers and switches, to make a flow of information available for analysis. On the other hand, packet-based anomaly detection doesn't rely on other network components; it observes network traffic for the detection of anomalies. Flow-based anomaly detection is based on the concept of a network flow and flow records. A flow record is summarized indicator that a certain network flow took place and that two hosts have communicated with each other previously at some point in time.

2. REQUIREMENT OF NIDS

An effective contemporary production-quality IDS needs an array of diverse components and features, including

- Centralized view of the data
- Data transformation capabilities
- Analytic and data mining methods
- Flexible detector deployment, including scheduling that enables periodic model relation and distribution
- Real-time detection and alert infrastructure
- Reporting capabilities
- Distributed processing
- High system availability
- Scalability with system load

3. DATA MINING AND NIDS

Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms [8]. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [9]. We are also interested in link and sequence analysis [10]. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern [10]. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets), instance-based examples, and probability models [8]. Data mining algorithms can be used for misuse detection and anomaly detection. In misuse detection, training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect anomalies & known intrusions. Research in this area has included the application of classification algorithms, association rule mining, and cost-sensitive modeling. Anomaly detection builds models of normal behaviour and automatically detects significant deviations from it. Supervised or unsupervised learning can be used. In a supervised approach, the model is developed based on training data that are known to be "normal." In an unsupervised approach, no information is given about the training data. Anomaly detection research has included the application of classification algorithms, statistical approaches, clustering, and outlier analysis. The techniques used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity. Classification algorithm about Data Mining can be used to construct classifier, after the invasion of a large number of data sets being trained. Classifier can be used for intrusion detection. Clustering analysis algorithm can also be used to construct the network model of normal behaviour, or intrusion behaviour model. Association analysis algorithm can be used to describe the invasion of behaviour patterns of association rules, through these rules intrusion detection can come. The goal of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. Examples of security violations include the abuse of privileges or the use of attacks to exploit software or protocol vulnerabilities. Traditionally, intrusion detection techniques are classified into two broad categories: misuse detection and anomaly detection [11].

Misuse detection works by searching for the traces or patterns of well- known attacks. Clearly, only known attacks that leave characteristic traces can be detected that way. Anomaly detection, on the other hand, uses a model of normal user or system behaviour and ages significant deviations from this model as potentially malicious. This model of normal user or system behaviour is commonly known as the user or system profile. Strength of anomaly detection is its ability to detect previously unknown attacks. Additionally, intrusion detection systems (IDSs) are categorized according to the kind of input information they analyze. This leads to the distinction between host-based and network-based IDSs. Host-based IDSs analyze host-bound audit sources such as operating system audit trails, system logs, or application logs.

4. SURVEY OF APPLIED TECHNIQUES ON DATA MINING APPROACH

In this section we present a survey of data mining techniques that have been applied to IDSs by various research groups.

4.1 Feature Selection

It is also known as subset selection or variable selection, is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. Feature selection is necessary either because it is computationally infeasible to use all available features, or because of problems of estimation when limited data samples (but a large number of Features) are present." Feature selection from the available data is vital to the effectiveness of the methods employed. Researchers apply various analysis procedures to the accumulated data, in order to select the set of features that they think maximizes the effectiveness of their data mining techniques. Two basic premises of intrusion detection are that system activities are observable, e.g., via auditing, and there is distinct evidence that can distinguish normal and intrusive activities. We call the evidence extracted from raw audit data features, and use these features for building and evaluating intrusion detection models. Feature extraction (or construction) is the processes of determining what evidence that can be taken from raw audit data is most useful for analysis. Feature extraction is thus a critical step in building IDS that is, having a set of features whose values in normal audit records differ signiûcantly from the values in intrusion records is essential for having good detection performance.

4.2 Machine Learning

Machine learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. In contrast to statistical techniques, machine learning techniques are well suited to learning patterns with no a priori knowledge of what those patterns may be. Clustering and Classification are probably the two most popular machine learning problems. Techniques that address both of these problems have been applied to IDSs.

5. DATA MINING TECHNIQUES - KDD

Data mining techniques and tools are the subject of the growing field of knowledge discovery in databases (KDD). According to Fayyad, Knowledge Discovery in Databases can be defined as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". Data mining is a particular step in which specific algorithms are applied to extract patterns from data.

KDD Process- The KDD process involves a number of steps and is very much interactive, iterative and user-driven process and are explained as follows:

1. Getting to know the application domain which means trying to understand the data and the discovery task.
2. Data preparation that includes creating a target dataset, removing noise from data and identifying a subset of the variables.
3. Data mining that includes deciding the model such as summarization, classification or clustering is to be derived from the dataset and then applying an appropriate algorithm to generate classification rules.
4. Interpretation which means trying to understand the discovered patters, returning to the previous steps to restart the process using different settings, removing redundant or trivial patterns and presenting the useful patterns to users.
5. Using the discovered knowledge – includes incorporating the knowledge into a production system or simply reporting them to interested third parties.

Data mining is the most critical step in the KDD process and a lot of research has been done to develop general, accurate and fast data mining algorithms.

6. DATA MINING APPLICATIONS IN IDS

The various data mining techniques are classified based on their functions, representation, preference criterion and algorithms [37, 38]. The main function of the model that we are interested in is classification, as malicious or malicious or as a particular type of attack [12][13][14]. We are also interested in link and sequence analysis [15][16]. Additionally, data mining system provide the means to easily perform data summarization and visualization, aiding the security analyst and identifying areas of concern. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions including neural networks, instance based examples and probability models [14].

7. CONCLUSION

A serious limitation of these approaches as well as with most existing IDSs is that they only do intrusion detection at the network or at system level. However, there is an urgent need to do intrusion and fraud detection at the application-level. This is because many attacks may focus on applications that have no effect on underlying system activities or the network used more for various domain specific researches like medical, educational etc.

REFERENCES

[1] Jiawei Han and. Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kufmann, 2nd edition 2006, 3rd edition 2011.

[2] S.J. Stolfo, W. Lee. P. Chan, W. Fan and E. Eskin, "Data Mining – based Intrusion Detector: An overview of the Columbia IDS Project" ACM SIGMOD Records vol. 30, Issue 4, 2001.

[3] S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy". Technical Report 99-15, Chalmers Univ., March 2000. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.6603.

[4] Tanase, Matthew, " One of These Things is not Like the Others: The State of Anomaly Detection", 2010, http://www.symantec.com/connect/articles/one-these-things-not-others-state-anomaly-detection

[5] C. Kruegel and G. Vigna. "Anomaly detection of web-based attacks", in ACM CCS'03.

[6] Esphion: Packet vs. flow-based anomaly detection. Technical White Paper, July 2005. http:/trendmap.net/support/wp ESP_WP_4_ PACKET_V_ FLOWS. pdf.

[7] Enterprise Strategy Group, "Network Behaviour Analysis Systems: The New Foundation of Defense-in-Depth", Technical White Paper, November 2005. http://www.enterprisestrategygroup.com/

[8] Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful Knowledge from volumes of data," Communications of the ACM 39 (11),November 1996, 2734.

[9] Ghosh, A. K., A. Schwartzbard, and M. Schatz," Learning program behavior profiles for intrusion detection", In Proc. 1st USENIX, 9- 12 April, 1999.

[10] Eric Bloedorn et al, "Data Mining for Network Intrusion Detection: How to Get Started," Technical paper, 2001.

[11] Mounji, A. (1997). Languages and Tools for Rule-Based Distributed Intrusion Detection. PhD thesis, Faculties Universitaires Notre-Dame dela Paix Namur (Belgium).

[12] S. Mukkamala et al. "Intrusion detection using neural networks and support vector machines", in IEEE IJCNN May 2002.

[13] Kumar, S.,"Classification and Detection of Computer Intrusion", PhD. thesis, 1995, Purdue Univ., West Lafayette, IN.

[14] Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth , "The KDD process for extracting useful knowledge from volumes of data," Communications of the ACM 39 (11). November 1996, 2734.

[15] Ghosh, A. K., A. Schwartzband, and M. Schatz, "Learning program behaviour profiles for intrusion detection", In Proc. 1st USENIX, 9-12 April, 1999.

[16] Lee W. and S. J. Stolfo, "Data mining approaches for intrusion detection", In Proc. Of the 7th.