International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

A Survey and Comparative Study of Different Data Mining Techniques for Implementation of Intrusion Detection System

Tanmayee S. Sawant[Å*] and Suhasini A. Itkar[Å]

[Å]Department of Computer Engineering, Pune University, PES's Modern College of Engineering, Pune, India

Accepted 25 April 2014, Available online 01 June 2014, **Vol.4, No.3 (June 2014)**

Abstract

With the increased use of internet, computerized applications and online transactions, it is most important to handle and prevent the different types of attacks and information security from intruders. Intrusion Detection System (IDS) is the most reliable system that can handle the intrusions of the computer environment and alert the network administrator so that they can take corrective actions to prevent that intrusion. Current intrusion detection systems may not able to detect unknown attacks as they are emerged swiftly every day. Up till now Intrusion detection system implemented using many techniques such as data mining technique, neural network technique, using combination of different classifiers, hybrid approach as well as layered approach. In this paper we showed the comparative study of different techniques of implementation of IDS

Keywords: *Data mining, Intrusion detection system.*

1. Introduction

Intrusions are the violations or impending threads of violations of computer security policies. Attack is any attempt to destroy, expose, alter, disable, steal or gain the unauthorized access to or make the unauthorized use of assets. Attack can be active or passive. An active attack attempts to alter system resources or affect their operations, hence comprises the integrity or availability. A passive attack attempts to learn or make use of information from the system but does not affect system resources, hence comprises confidentiality. An attempt of attack can takes place from inside or outside the organization. Insider attacker is one who has authorized access to system resources but use them in illegitimate way. Outside attacker is the illegal user of the system. A close-in attack involves someone attempting to get physically close to network components, data, and systems in order to learn more about a network. In phishing attack the hacker creates a fake web site that looks exactly like an original website, when the user attempts to log on with their account information, the hacker records the username and password and then tries that information on the real website. In a hijack attack, a hacker takes over a session between you and another person and disconnects the other person from the communication and you still consider that you are talking to the original party and may send private information to the hacker by an accident. In a spoofing attack, the hacker modifies the source address of the packets he or she is sending so that they appear to be coming from someone else. This may be an attempt to bypass firewall rules. A buffer overflow attack is when the attacker sends more data to an application than the estimated capacity. In Exploit type of attack, the attacker knows of a security problem within an operating system or a piece of software and leverages that knowledge by exploiting the vulnerability. In Password attack, an attacker tries to break the passwords stored in a network account database or in a password-protected file. Intrusion Detection System (IDS) is the most authoritative system that can handle the intrusions of the computer environment by alerting the analyst so that they can take corrective actions to prevent that intrusion**.** Major functions of Intrusion detection system involves,

1. Used to examine the network traffic.
2. Identifying possible events by monitoring both user and system.
3. Logging information about user and system.
4. Analyzing system configuration and vulnerabilities.
5. Assessing file and system integrity.
6. Recognizing abnormal activities and patterns of typical types of attacks.
7. Reporting them to network security administrator.

Additional to this, many organizations use Intrusion detection system for other purposes such as identifying problems with security policies, documenting the existing threats, deterring the individuals from violating the security policies.

2. IDS Terminology

Following are the terminologies used in Intrusion Detection System,

1. **Burglar Alarm:** A signal suggesting that a system has been or is being attacked.

*Corresponding author: **Tanmayee S. Sawant**

2. **Detection Rate:** The detection rate is defined as the number of intrusion samples detected by the system (True Positive) divided by the total number of intrusion samples present in the test set.
3. **False Alarm Rate:** defined as the number of 'normal' patterns classified as attacks (False Positive) divided by the total number of 'normal' patterns.
4. **True Positive:** A legitimate attack which triggers IDS to produce an alarm.
5. **False Positive:** An event signaling IDS to produce an alarm when no attack has taken place.
6. **False Negative:** When no alarm is raised when an attack has taken place.
7. **True Negative:** An event when no attack has taken place and no detection is made.
8. **Noise:** Data or obstruction that can trigger a false positive or indefinite a true positive.
9. **Site policy:** Guidelines within an organization that control the rules and configurations of IDS.
10. **Site policy awareness:** An IDS's ability with dynamism to change its rules and configurations in response to changing environmental activity
11. **Confidence value:** A value an organization places on an IDS based on past performance and analysis to help determine its ability to effectively identify an attack.
12. **Alarm filtering:** The process of categorizing attack alerts produced from an IDS in order to distinguish false positives from actual attacks.
13. **Attacker** or **Intruder:** An entity who tries to find a way to gain unauthorized access to information, inflict harm or engage in other malicious activities.
14. **Masquerader:** A user who does not have the authority to a system, but tries to access the information as an authorized user. They are generally outside users.
15. **Misfeasor:** They are commonly internal users and can be of two types:
    a. An authorized user with limited permissions.
    b. A user with full permissions and who misuse their powers

2.1 Approaches of Intrusion Detection system

An active Intrusion Detection Systems (IDS) is also known as Intrusion Detection and Prevention System (IDPS). Intrusion Detection and Prevention System (IDPS) is configured to automatically block suspected attacks without any participation required by an operator. Intrusion Detection and Prevention System (IDPS) has the advantage of providing real-time corrective action in response to an attack. A passive IDS is a system that's configured to only supervise and analyze network traffic activity and alert an operator to potential vulnerabilities and attacks. A passive IDS is not capable of performing any protective or corrective functions on its own.

Network Intrusion Detection Systems (NIDS) usually consists of a network appliance with a Network Interface Card (NIC) operating in profligate mode and a separate management interface. The IDS is placed along a network segment or boundary and monitors all traffic on that segment.

A Host Intrusion Detection Systems (HIDS) and software application installed on workstations which are to be monitored. The agents monitor the operating system and write data to log files and/or activate alarms. Host Intrusion detection systems (HIDS) can only monitor the individual workstations on which the agents are installed and it cannot monitor the entire network. Host based IDS systems are used to monitor any intrusion attempts on critical servers. The drawbacks of Host Intrusion Detection Systems (HIDS) are-

1. Difficult to examine the intrusion attempts on multiple computers.
2. Host Intrusion Detection Systems (HIDS) can be very difficult to maintain in large networks with different operating systems and configurations
3. Host Intrusion Detection Systems (HIDS) can be disabled by attackers after the system is compromised.

A knowledge-based (Signature-based) Intrusion Detection Systems (IDS) references a database of previous attack signatures and known system vulnerabilities. The meaning of word signature, when we talk about Intrusion Detection Systems (IDS) is recorded substantiation of an intrusion or attack e.g., nature of data packets, failed attempt to run an application, failed logins, file and folder access etc. These footprints are called signatures and can be used to identify and prevent the same attacks in the future. Based on these signatures Knowledge-based (Signature-based) IDS identify intrusion attempts.

The disadvantages of Signature-based Intrusion Detection Systems (IDS) are signature database is not continually updated and maintained so signature-based Intrusion Detection Systems (IDS) may fail to identify a new attack.

A Behavior-based (Anomaly-based) Intrusion Detection Systems (IDS) references a baseline or learned pattern of normal system activity to identify active intrusion attempts. Deviations from this baseline or pattern cause an alarm to be triggered. Higher false alarms are often related with Behavior-based Intrusion Detection Systems (IDS).

3. Introduction to Data Mining

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating enormous and rising amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- Nonoperational data, such as industry sales, forecast data, and macro economic data meta data - data about the data itself, such as logical database design or data dictionary definitions

Information is the patterns, associations, or relationships among all this data which can provide information.

Knowledge is the information which can be converted into knowledge about historical patterns and future trends. The Knowledge Discovery in Database (KDD) process is generally defined with the stages:

1. Selection
2. Pre-processing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation

Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. It is a suitable way of extracting patterns, which represents mining completely stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability.

Data mining consists of five major elements

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

3.1 Advantages of Data Mining Techniques

1. Problems with large databases may contain valuable implicit regularities that can be discovered automatically.
2. Difficult-to-program applications, which are too difficult for traditional manual programming.
3. Software applications that modify to the individual users preferences, such as modified advertising.

4. Different Techniques to implement IDS

Intrusion detection system can be implemented using different techniques like data mining classification technique, neural network techniques, artificial intelligence techniques etc. Till now Intrusion detection system has been developed using data mining classification algorithms such as decision trees , Bayes classifiers, k- nearest neighbor classifier, case based reasoning, fuzzy logic technique, neural networks, clustering algorithms, genetic algorithm etc, Some researcher has implemented it using combination of different classifiers, hybrid approach as well as layered approach. Several types of algorithms are particularly significant for implementation of intrusion detection system. Several types of algorithm are used in implementation of IDS such as,

Classification: maps a data item into one of several pre-defined categories. These algorithms normally output "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 will determine future audit data as belonging to the normal class or the abnormal class.

Link analysis: determines relations between fields in the database. Finding out the correlations in audit data will provide insight for selecting the right set of system features for intrusion detection.

Sequence analysis: models sequential patterns. These algorithms can help us recognize what (time-based) sequence of audit events are frequently encountered together. These recurrent event patterns are significant elements of the behavior profile of a user or program.

Decision trees are a technique from data mining that categorize new pieces of information into a number of predefined categories. Decision trees use a pre-classified dataset to learn to categorize data based on existing trends and patterns. After the tree is created, the logic from the decision tree can be included into a number of different intrusion detection technologies including firewalls and IDS signatures. Advantages of Decision Tree:

1. Easy to understand and interpret.
2. Implicitly performs the feature selection.
3. Requires less effort from user for data preparation.
4. A non linear relationship between the parameters does not affect the tree performance.

5. Experimental Methodology

5.1Dataset used

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99.This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. The data set used by many course of research is the DARPA KDD99 benchmark data set, also known as "DARPA Intrusion Detection Evaluation data set" that not only includes a large quantity of network traffic but also collects a wide variety of attacks. Attack fall into following four major categories:

1. **Denial of service (DoS) attacks:** Attackers disrupt a host or network service to make legitimate users can not access to a machine, E.g. Apache2, Mail bomb, Back, Smurf, Land, SYN Flood, Ping of death, Process table, Teardrop.

2. **Remote to Local (R2L) attacks:** Unauthorized attackers gain local access from a remote machine and then exploit the machine's vulnerabilities, E.g. Ipsweep, Mscan, Saint, Satan and Nmap.

3. **User to Root (U2R) attacks:** Local users get access to local machine without authorization and then exploit the machine's vulnerabilities, E.g. Eject, Xterm, Loadmodule, Ps, Perl and Fdformat.

4. **Probes:** It is a category of attacks where an attacker examines a network to discover well-known vulnerabilities. E.g. Guest Dictionary, Phf, Ftp_write, Imap, Named, Sendmail and Xlock.

Following table shows comparative study of various data mining techniques along with its advantages and disadvantages.

Table 1 Comparative study of various Data mining Techniques

Classifier	Method used	Parameters considered	Advantages	Disadvantages
Support Vector Machine	A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks.	Its efficiency lies in the selection of kernel and soft margin parameters. For kernels, different pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is picked. Trying exponentially increasing sequences of C is a practical method to recognize good parameters.	1.Highly Accurate 2.Able to model complex nonlinear decision boundaries 3.Less prone to over fitting than other methods	1.High algorithmic complexity and extensive memory requirements in large-scale tasks. 2. The choice of the kernel is difficult 3. The speed both in training and testing is slow.
Artificial Neural Network	It is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.	It uses the cost function *C* is a key concept in learning, as it is a measure of how far away a particular solution is from an optimal solution to the problem to be solved.	1. Requires less formal statistical training. 2. Able to absolutely detect composite nonlinear relationships between dependent and independent variables. 3. High tolerance to noisy data. 4. Availability of multiple training algorithms.	1."Black box" nature. 2. Greater computational burden. 3. Proneness to over fitting. 4. Requires long training time.
Bayesian Method	Based on the rule, using the joint probabilities of sample interpretation and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.	In Bayes, all model parameters (i.e., class priors and feature probability distributions) can be approximated with relative frequencies from the training set.	1. Naive Bayesian classifier simplifies the computations. 2. Reveal high accuracy and speed when applied to large databases.	1 The assumptions made in class conditional independence. 2. Lack of available probability data.
Decision Tree	Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute; one branch corresponds to the positive instances of the predicate and the other to the negative instances.	Decision Tree Induction uses parameters like a set of contender attributes and an attribute selection method.	1. Construction does not require any domain knowledge. 2. Can handle high dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. Output attribute must be categorical. 2. Limited to one output attribute. 3. Decision tree algorithms are unstable. 4. Trees created from numeric datasets can be complex.

Conclusions

This paper has presented a survey and comparative study of the various data mining techniques that have been proposed towards the improvement of implementation of Intrusion Detection Systems. We have shown the ways in which data mining has been known to help the process of Intrusion Detection and the ways in which the various techniques have been applied. Finally we proposed a comparative study of data mining approaches that we feel can contribute extensively in the attempt to create better and more effective Intrusion Detection Systems.

References

Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu, and Stephen Maybank (2013) Online Adaboost-Based Parameterized Methods for Dynamic Distributed Network Intrusion Detection, IEEE Transactions on Cybernetics.

Luigi Coppolino, Salvatore D'Antonio, Alessia Garofalo, Luigi Romano (2013) Applying Data Mining Techniques to Intrusion detection in Wireless Sensor networks, Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing.

T. Subbulakshmi, Ms. A. Farah Afroze (2013) Multiple Learning based Classifiers using Layered Approach and Feature Selection for Attack Detection, IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN).

Vikas Sharma, Aditi Nema (2013) Innovative Genetic approaches For Intrusion Detection by Using Decision Tree, International Conference on Communication Systems and Network Technologies.

Manish Kumar, Dr. M. Hanumanthappa, Dr. T. V. Suresh Kumar (2012) Intrusion Detection System Using Decision Tree Algorithm, proceeding for IEEE.

Shina Sheen, R Rajesh, Member IEEE, (2012) Network Intrusion Detection using Feature Selection and Decision tree classifier, IEEE International Conference on Tools with Artificial Intelligence.

V. K. Pachghare, Parag Kulkarni (2011) Pattern Based Network Security using Decision Trees and Support Vector Machine, System engineering and Electronics, Vol.27, No.7, July.

Mrutyunjaya Panda, Manas Ranjan Patra (2008) A Comparative Study of Data Mining Algorithms for Network Intrusion Detection, First International Conference on Emerging Trends in Engineering and Technology.

W. Lee and S. J. Stolfo (.2000) A framework for constructing features and models for intrusion detection systems, ACM Trans. Information System security, Vol 3, no. 4.

Wenke Lee S alvatore J. Stolfo Kui W. Mok (2000), A Data Mining Framework for Building Intrusion Detection Models", International virus bulletin.

W. M. Hu, W. Hu, and S. Maybank (2008), Adaboost-based algorithm for network Intrusion detection, IEEE Trans.Syst., Man, Cybern, part B: Cybern, vol. 38, No. 2.

Matthew G. Schultz and Eleazar Eskin, Erez Zadok (2008) Data Mining Methods for Detection of New Malicious Executables, Proceedings of the 2008 International Virus Bulletin Conference.