ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume No. 2, Issue No. 6, June 2013

A Survey on Intrusion Detection System in Data Mining

[1]Sahilpreet Singh,[2]Meenakshi Bansal

[1,2]Departmentof Computer Engineering, Punjabi University
Yadavindra College of Engineering, Talwandi Sabo
Punjab, India

ABSTRACT

This paper presents a survey of techniques of intrusion detection system using supervised and unsupervised learning. The techniques are categorized based upon different approaches like Statistics, Data mining, Neural Network Based and Self Organizing Maps Based approaches. The detection type is borrowed from intrusion detection as either misuse detection or anomaly detection. It provides the reader with the major advancement in the malware research using these approaches the features and categories in the surveyed work based upon the above stated categories. This served as the major contribution of this paper.

***Keywords*: - Intrusion Detection System, Neural Network, Data Mining**

1. INTRODUCTION

Computer networks and systems have become indispensable tools for modern business much of this information is, to some degree, confidential and its protection is required. Not surprisingly then, intrusion detection systems (IDS) have been developed to help uncover attempts by unauthorized persons and/or devices to gain access to computer networks and the information stored therein. An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station. Some systems may attempt to stop an intrusion attempt but this is neither required nor expected of a monitoring system. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. The development of IDS is motivated by the following factors because Most existing systems have security was that render them susceptible to intrusions, and finding and fixing all these deficiencies are not feasible. Prevention techniques cannot be sufficient. It is almost impossible to have an absolutely secure system. Even the most secure systems are vulnerable to insider attacks. New intrusions continually emerge and new techniques are needed to defend against them. Since there are always new intrusions that cannot be prevented, IDS is introduced to detect possible violations of a security policy by monitoring system activities and response. IDSs are aptly called the second line of defence, since IDS comes into the picture after an intrusion has occurred. If we detect the attack once it comes into the network, a response can be initiated to prevent or minimize the damage to the system.

It also helps prevention techniques improve by providing information about intrusion techniques. Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack. We are also interested in link and sequence analysis. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions, instance-based examples, and probability models.

DATA MINING BASED INTRUSION DETECTION SYSTEM ARCHITECTURE

The overall system architecture is designed to support a data mining-based IDS with the properties described. The architecture is consists of sensors, detectors, a data warehouse, and a model generation component. This architecture is capable of supporting not only data gathering, sharing, and analysis, but also data archiving and model generation and distribution. The system is designed to be independent of the sensor data format and model representation. A piece of sensor data can contain an arbitrary number of features. Each

feature can be continuous or discrete, numerical or symbolic.

1.1 Sensors

Sensors observe raw data on a monitored system and compute features for use in model evaluation. Sensors insulate the rest of the IDS from the specific low level properties of the target system being monitored. This is done by having the entire sensors implement a Basic Auditing Module (BAM) framework. In a BAM, features are computed from the raw data and encoded in XML.

1.2 Detectors

Detectors take processed data from sensors and use a detection model to evaluate the data and determine if it is an attack. The detectors also send back the result to the data warehouse for further analysis and report. There can be several (or multiple layers of) detectors monitoring the same system. There can also be a "back-end" detector, which employs very sophisticated models for correlation or trend analysis, and several "front-end" detectors that perform quick and simple intrusion detection.

1.3 Data Warehouse

The data warehouse serves as a centralized storage for data and models. One advantage of a centralized repository for the data is that different components can manipulate the same piece of data asynchronously with the existence of a database, such as off-line training and manually labeling. The data warehouse also facilitates the integration of data from multiple sensors. By correlating data/results from different IDSs or data collected over a longer period of time, the detection of complicated and large scale attacks becomes possible.

1.4 Model Generator

The main purpose of the model generator is to facilitate the rapid development and distribution of new (or updated) intrusion detection models. In this architecture, an attack detected first as an anomaly may have its exemplary data processed by the model generator, which in turn, using the archived normal and intrusion data sets from the data warehouse, automatically generates a model that can detect the new intrusion and distributes it to the detectors. Especially useful are unsupervised anomaly detection algorithms because they can operate on unlabeled data which can be directly collected by the sensors.

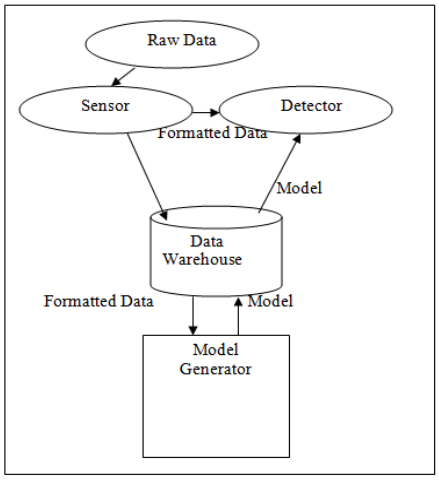

Fig 1.The Architecture of Data Mining based IDS

2. DATA MINING BASED APPROACHES

Data mining is used in intrusion detection to construct rules describing normal network behaviors. The rules include association rules that describe frequency associations between any two fields of the network record database and also frequent episodes that describe the frequency with which a field takes a certain value after two other fields have particular values in a definite time interval. Deviations from these rules indicate an attack on the network.

2.1 Supervised Learning-Based Approaches:

Recently, methods from machine learning and pattern recognition have been utilized to detect intrusions. Supervised learning and unsupervised learning are both used. For supervised learning for intrusion detection, there are mainly supervised neural network (NN)-based approaches & support vector machine (SVM)-based approaches

2.2 Unsupervised Learning-Based Approaches:

Supervised learning methods for intrusion detection can only detect known intrusions. Unsupervised learning methods can detect the intrusions that have not been previously learned. An example of unsupervised learning for intrusion detection includes *K*-means-based approaches and self-organizing feature map (SOM).Current approaches for intrusion detection have the following two problems.

a) Current approaches often suffer from relatively high false alarm rates, whereas they have high detection rates. As most network behaviors are normal, resources are wasted on checking a large number of alarms that turn out to be false.
b)Their computational complexities are oppressively high. This limits the practical applications of these approaches.

3. RELATED WORK

Anazida Zainal et al. (2008) in paper has discussed the Efficiency is one of the major issues in intrusion detection. Inefficiency is often attributed to high overhead and this is caused by several reasons. The purpose of the paper is to address the issue of continuous detection by introducing traffic monitoring mechanism. In traffic monitoring, a new recognition paradigm is proposed in which it minimizes unnecessary recognition. Therefore, the purpose of traffic monitoring is two-folds; to reduce amount of data to be recognized and to avoid unnecessary recognition. For this Adaptive Neural Fuzzy Inference System and Linear Genetic Programming to form ensemble classifiers that shows a small improvement using the ensemble approach for DoS and R2L classes (attacks).

G. Zhai et al. (2010) in paper has discussed that ID3 algorithm was a classic classification of data mining. It always selected the attribute with many values. The attribute with many values wasn't the correct one, it would created fault alarm and omission alarm. To this fault, an improved decision tree algorithm was proposed. The decision tree was created after the data collected classified correctly. With the help of using Decision tree algorithm it shows the maximum attacks and also increases the alert level after modified the decision tree.

Jorge Blasco et al. (2010) in paper **has** studied that one of the central areas in network intrusion detection is how to build effective systems that are able to distinguish normal from intrusive traffic. To avoid the blind use of GP, it provides the search by means of a fitness function based on recent advances on IDS evaluation. For the experimental work use of a well-known dataset (i.e. KDD- 99) that has become a standard to compare research although its drawbacks. Results clearly show that an intelligent use of GP provides better accuracy and also compare the Hit rate and False Rate to detect the number of attacks.

Ahmed Youssef et al. (2011) in paper has studied that Intrusion detection has become a critical component of network administration due to the vast number of attacks persistently threaten our computers. Traditional intrusion detection systems are limited and do not provide a complete solution For the problem. However, in many cases, they fail to detect malicious behaviors (false negative) or They fire alarms when nothing wrong in the network (false positive). For this combination of Data Mining Techniques and Network behavior analysis were applied and overcome the limitations of traditional Intrusion Detection System.

Mohd. Junedul Haque et al. (2012) in paper has said that the Intrusion Detection system is an active and driving secure technology to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a network. The main part of Intrusion Detection Systems (IDSs) is to produce huge volumes of alarms. The interesting alarms are always mixed with unwanted, non-interesting and duplicate alarms. For this Data mining algorithm, K means clustering, Distributed IDS are applied to improve the detection rate and decrease the false alarm rate.

S. Devaraju et al. (2013) in paper has discussed about the security purpose in information system. To deal with the problems of networks different classifiers are used to detect the different kinds of attacks. In this, the performance of intrusion detection with various neural network classifiers is compared. In this proposed research there are five types of classifiers used. They are Feed Forward Neural Network (FFNN), Elman Neural Network (ENN), Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) and Radial Basis Neural Network (RBNN). Finally it is clear that Probabilistic Neural Network has better accuracy than rest of other neural networks.

S.A.Joshi et al. (2013) in paper has presented that with the tremendous growth in information technology, network security is one of the challenging issue and so as Intrusion Detection system (IDS). The traditional IDS are unable to manage various newly arising attacks. To overcome this type of problem Data Mining techniques, Feature Selection, Multiboosting were applied. With data mining, it is easy to identify valid, useful and understandable pattern in large volume of data. Features are selected using binary classifiers for more accuracy in each type of attack. Multiboosting is used to reduce both the variance and bias. Thus the efficiency and accuracy of Intrusion Detection system are increased and security of network so is also enhanced.

4. COMPARATIVE STUDY

Author(s)	Year	Paper Name	Technique	Results
S.A.Joshi, et al.	2013	Network Intrusion Detection System (NIDS) based on Data Mining	Data Mining, Feature Selection, Multiboosting	Find high detection rates for U2R and R2L and also to detect attacks.
S. Devaraju, et al.	2013	Detection of Accuracy for IDS in Neural Network	Different types of Neural Networks and KDD cup	Probabilistic Neural network has better accuracy than others Neural network.
Mohd. Junedul Haque et al.	2011	An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques	Data mining algorithm, K means clustering, Distributed IDS	False alarm rate has been decreased also clustering helps in to identify the attacked data.
Ahmed Youssef, et al.	2011	Network Intrusion Detection using Data Mining and Network behavior analysis	Data Mining Techniques and Network behavior analysis	Combination of both DM and NBA overcome the limitation of traditional IDS
Jorge Blasco, et al.	2010	Improving Network Intrusion Detection by Means of Domain-Aware Genetic Programming	Use of Genetic Programming	Explore the Hit rate and False Rate on data set to detect no. of attacks
G. Zhai et al.	2010	Research and Improvement on ID3 Algorithm in Intrusion Detection System	Decision tree Algorithm	Shows maximum attacks and also increases the alert level after modified the decision tree
Anazida Zainal, et al.	2008	Data Reduction and Ensemble Classifiers in Intrusion Detection	Adaptive Neural Fuzzy Inference System and Linear Genetic Programming	LGP has better detection accuracy than ANFIS

5. CONCLUSION

It is shown in the paper that there is several intrusion detections tools with competing features which are develop for detection of attacks like known attacks and unknown attacks and also supervised and Un-supervised approaches are used to detect the attacks. Unsupervised learning methods can detect the intrusions that have not been learned by supervised approaches.

REFRENCES

[1]. Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin "Data Reduction and Ensemble Classifiers in Intrusion Detection" in *2008 IEEE.*

[2]. Guangqun Zhai, Chunyan Liu "Research and Improvement on ID3 Algorithm in Intrusion Detection System" in *2010 IEEE*

[3]. Jorge Blasco, Agustin Orfila, Arturo Ribagorda "Improving Network Intrusion Detection by Means of Domain-Aware Genetic Programming" *DOI 10.1109/ARES.2010.53 in IEEE 2010.*

[4]. Ahmed Youssef and Ahmed Emam "Network Intrusion Detection using Data Mining and Network

Behavior Analysis" *International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.*

[5]. Mohd. Junedul Haque, Khalid.W. Magld, Nisar Hundewale "An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques" in *2012 IEEE.*

[6] .N.S.Chandolikar, V.D.Nandavadekar "Comparative analysis of two algorithm for Intrusion attack classification using dataset" *in International Journal of Computer Science and Engineering (IJCSE) in 2012*
.

[7]. Devendra kailashiya, Dr. R.C. Jain "Improve Intrusion Detection Using Decision Tree with Sampling" in *IJCTA | MAY-JUNE 2012*

[8]. S.A.Joshi, Varsha S.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining" *International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013*

[9]. S. Devaraju, S .Ramakrishnan "Detection of Accuracy for Intrusion Detection System using Neural Network Classifier" *International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)*

[10] Yacine Bouzida, Frederic Cuppens "Neural networks vs. decision trees for intrusion detection" *in 2011.*

Sahilpreet Singh received his B.Tech degree in Information Technology from Swami Vivekanand Institute of Engineering & Technology (Ramnagar, Banur) under Punjab Technical University in 2011 and pursuing M Tech. (Regular) degree in computer engineering from Yadavindra College of Engineering Punjabi University Guru Kashi Campus Talwandi Sabo (Bathinda), batch 2011-2013. His research interests include improvement of Intrusion Detection System in Data Mining.