International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Intrusion Detection Using Data Mining Approach

Kamble Jayshree R.[1], Rangdale S.P.[2]

[1]Siddhant College of Engineering, Sudumbare, Pune-412 109, India

[2]Assistant Professor, Information Technology, Siddhant College of Engineering, Sudumbare, Pune, India

Abstract: ***Intrusion Detection is one of intelligent application in which data mining technique can be used. This paper presents data mining approach in Intrusion detection. Intrusion detection is nothing but the detection of action that attempts to compromise the system security factors such as integrity, availability of a resource, confidentiality. Intrusion detection does not involve the Intrusion prevention functionality. This paper represents the approach for expert system doing data mining embedded with Intrusion Detection System. Complete information about intrusion mechanism is required to generate the appropriate decision. This paper discusses the frame work for the data mining based IDS and architecture of Data Mining based IDS.***

Keywords: Intrusion Detection System, Anomaly Detection, Misuse Detection DOS, Probing, U2R, R2L

1. Introduction

1.1 Intrusion Detection System and Its Types

Intrusions are the activities that violate the security policy of system or it is attempts to compromise the system security factors. Intrusion Detection is the process used to identify intrusions in system. Based on the sources of the audit information used by each IDS, the IDSs may be classified into Host-base IDSs, Distributed IDSs and Network-based IDSs

1.1.1 Host-based IDSs
In this IDS input audit data is given by host audit trails and it detects attacks against a single host.

1.1.2 Distributed IDSs
In this IDS audit data is gather from multiple host and possibly the network that connects the hosts and detect attacks involving multiple hosts.

1.1.3 Network-Based IDSs
It uses network traffic as the audit data source, relieving the burden on the hosts that usually provide normal computing services and detect attacks from network.

1.2 Drawbacks of Traditional IDS

Traditional intrusion detection system IDS tools are based on signatures of known attacks and have well known limitations which are as follows

- Traditional IDS are normally detects known service level network attacks.
- Traditional IDS has to be manually revised for each new type of discovered intrusion
- Unable to detect emerging cyber threats
- Not suitable for detecting policy violations and insider abuse
- Do not provide understanding of network traffic
- Generate too many false alarms
- Not suited for detecting multi-step attacks
- Data overload: The amount of data the analyst needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

Data mining can help improve intrusion detection by addressing each and every one of the above mentioned problems.

1.3 Contribution of Data Mining in Intrusion Detection

Data Mining [2] is the process of extracting the knowledge by automatically searching large volumes of data for patterns using association rules.

1.3.1Data Mining Techniques used in IDS

- Data summarization
- Visualization: Which presents a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs

One or combination of above mentioned data mining techniques can be employed to find anomalous activity that uncovers a real attack, remove normal activity from alarm data to allow analysts to focus on real attacks and to identify long, ongoing patterns (different IP address, same activity)

1.4 Intrusion Detection Techniques

1.4.1 Misuse detection
In this technique the intrusion is detected in terms of characteristics of known attacks or system vulnerabilities.

Characteristics

a) It is based on known attack actions.
b) Features are extracted from known intrusions.
c) It is integrated with the Human knowledge.
d) The rules are pre-defined

Disadvantage

Cannot detect novel or unknown attacks

1.4.2 Anomaly detection

Detect any action that significantly deviates from the normal behavior.

Characteristics
It is based on the normal behavior of a subject. Sometime assume the training audit data does not include intrusion data. Any action that significantly deviates from the normal behavior is considered intrusion.

2. Literature Survey

In this section we present a Literature Survey

2.1 Machine Learning

Machine Learning [6] 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 Clustering and Classification are probably the two most popular machine learning problems.

2.2 Feature Selection

Feature selection is common process in machine learning. It is also known as subset selection because features are the subset of data values. 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.

2.3 Statistical Techniques

Statistical techniques are opposite to the Machine Learning. It is also known as "top-down" learning , When we employ the mathematics to aid our search then these statistical techniques can be applied. Three basic classes of statistical techniques are linear, nonlinear (such as a regression-curve), and decision trees .

2.4 Ensemble Approaches

Combined use of multiple data mining methods or techniques is known as Ensemble Approach. Data mining correlate this techniques obtained result of each technique is combined. Advantage of this approach is if one technique fails another can detect the attack.

3. Architecture of Data Mining based IDS

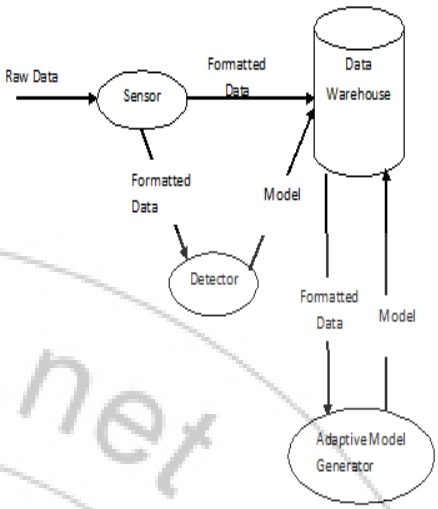

Figure 1 Architecture of Data Mining Based IDS

The above Figure 1 shows overall system architecture of data mining based IDS. As shown in Figure 1, the architecture consists of sensors, detectors, a data warehouse, and a Adaptive model generation component. This architecture supports the gathering, sharing and analyzing of data. It generates models and distributes them. The system is independent of the sensor data format and model representation.

In above architecture, data and model exchanged between the components are encoded in standard message format, High performance and scalability are the advantages of this architecture.
The complete description of architecture is as follows:

Sensors
Sensors receive the raw data and compute features for use in model evaluation. All the sensors implement a Basic Auditing Module (BAM) framework. In BAM, features are computed from the raw data and encoded in XML.

Detectors
Detectors take processed data from sensors and use detection model to evaluate the data and determine if it is an attack. There can be multiple layers of detectors in the same system. The detectors can send back the result to the data warehouse for further analysis and report.

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. The data warehouse 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.

Adaptive Model Generator (AMG)

The model generator [3] facilitates 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 (historical) 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 (or any other IDSs that may use these models). Especially useful are unsupervised anomaly detection algorithms because they can operate on unlabeled data which can be directly collected by the sensors.

4. Framework for Data Mining Based IDS

Framework Data Mining Based IDS [4], When audit mechanisms are enabled to record system events, distinct evidence of legitimate activities and intrusions will be manifested in the audit data. Because of the large volume of audit data, both in the amount of audit records and in the number of system features (fields of the audit records), efficient and intelligent data analysis tools are required to discover the behavior of system activities.

In data mining generally wide verity of algorithms are available which are drawn from statistics, pattern recognition, machine learning and databases. Types of algorithms are particularly useful for mining audit data:

4.1 Classification

Classification maps a data item into one of several predefined categories. Output of these algorithms are Classifiers e.g. in the form of decision trees or rules. Data Mining based IDS gathers the normal and abnormal audit and applies the classification algorithm. Further this algorithm can rectify the unseen audit data which belongs to normal class or the abnormal class.

4.2 Link Analysis

Link Analysis [2] determines relations between fields in the database records. Correlations of system features in audit data can serve as the basis for constructing normal usage profile.

4.3 Sequence Analysis

Sequence analysis [2] models sequential patterns. These algorithms can discover what time-based sequence of audit events are frequently occurring together. These frequent event patterns provide guidelines for incorporating temporal statistical measures into intrusion detection models. For example, patterns from audit data containing network-based denial-of-service (DOS) attacks suggest that several per-host and per-service measures should be included.

5. Experiment

The experiments were performed on the network traffic offline data of Defense Advanced Research Projects Agency (DARPA) 2000 Intrusion Detection Evaluation program. The data consists of network connection records generated by TCP dump. A network connection record is a set of information, such as duration, protocol type, number of transmitted bytes etc, which represents a sequence of data flow to and from a well defined source and target. Each record in this data was marked as normal or attack, with exact specification about the attack type. All the attacks fall into four main categories.

- DOS: denial-of-service, e.g. syn flood;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local super user (root) privileges, e.g., various buffer overflow attacks;
- probing: surveillance and other probing, e.g., port scanning

There are two types of data sets one is training dataset while other is test dataset.

Table 1: Attack types in the training data set

S.No.	*Attack Name*	*Category*
1	Back	DoS
2	buffer_overflow	U2R
3	ftp_write	R2L
4	guess_passwd	R2L
5	Imap	R2L
6	Ipsweep	Probe
7	Land	DoS
8	Loadmodule	U2R
9	Multihop	R2L
10	Neptune	DoS
11	Nmap	Probe
13	Perl	U2R
14	Phf	R2L
15	Pod	DoS

5.1 Data Preprocessing

The processing performed on the raw data to prepare it for any other processing procedure is known as data preprocessing. Data preprocessing is an important step in data mining. It transforms the data into a format that is acceptable to the actual data processing algorithm e.g. a neural network

5.2 Computing Environment

The computing environment used for experiments was composed of a cluster of 32 PCs running LINUX. Each of these machines was equipped with a 900 MHz AMD Athlon processor with 1 GB memory and 100BT networking. Two different machine learning algorithms were parallelized using different message passing tools for mining the KDD data. The programming tool used to implement parallel back propagation was CRLib

5.3 Experimental Result

As we used two data set training and test data set .Following table 2 shows the detection rate of old and new attack on

both data set. New attacks found in only test data set but not in training data set.

Table 2: Detection rate of Intrusion

Category	*Old Attack*	*New attack*
DOS	77.5	21.2
Probing	98	95.4
U2R	23	79.9
R2L	57.8	4.6
Total	256.3	201.1

The figure below shows the graphical representation of detection rate of old and new attacks

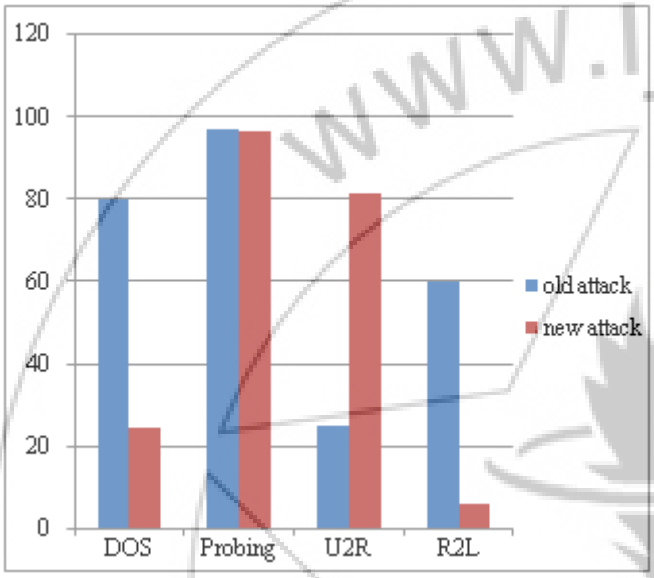

Figure 2: Detection rate of Intrusion

References

[1] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Nig Tan, "Data Mining for Network Intrusion Detection", Computer Science Department, University of Minnesota, MN 55455, USA
[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process of extracting useful knowledge", Volumes of data Communications of the ACM, 39(11):27–34, November 1996.
[3] Andrew Honig, Andrew Howard, Eleazar Eskin, Sal Stalfo, " Adaptive Model Generation :An Architecture for deployment of data mining based IDS", Department of Computer Science, Columbia University, New York.
[4] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok., "A Data Mining Framework for Building Intrusion Detection Models", Computer Science Department, Columbia University.
[5] Rob Schapire, "Machine Learning Algorithms for Classification", Princeton University
[6] Fabrizio Sebastiani, "Machine learning in automated text categorization" ,ACM Computing Surveys(CSUR),Vol. 34, Issue 1,March 2002.
[7] Muazzam Siddiqui, "High Performance Data Mining Techniques for Intrusion Detection", B.E.NED. University of Engineering and Technology, 2000.

Author Profile

Ms. Kamble Jayshree R. received B.E. (Computer) degree from Bharti Vidyapeeth College of Engg. For Women, Pune University in 2008, Now pursuing M.E.(Information Technology) degree from Siddhant College of Engg., Pune University.