ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

CLASSIFICATION OF DATA MINING-BASED INTRUSION DETECTION SYSTEMS

S.PRADEEPKUMAR

M.Phil Research Scholar, Department of Computer Science,

RVS College of Arts & Science, Sulur.

ABSTRACT

In this paper, we present an overview of real time data mining-based IDS Intrusion Detection Systems (IDSs).We focus on issues related to deploying a data mining-based IDS in a real time environment. IDS technology is one of the important tools used now-a-days, to counter such threats. Various IDS techniques has been proposed, which identifies and alarms for such threats or attacks .Intrusion Detection Systems are designed to detect system attacks and it classifies system activities into normal and abnormal form. Data Mining based intrusion detection system model generalizes and detects both known attacks and normal behavior in order to detect unknown attacks and fails to generalize and detect new attack without known signatures. To improve efficiency, the computational costs of features are analyzed and a multiple-model cost- based approach is used to produce detection models with low cost and high accuracy. To deal with these new problems of networks, data mining based IDS are opening new research avenues. The paper provides a study on the various data mining based intrusion detection techniques. This architecture facilitates the sharing and storage of audit data and the distribution of new or updated models. This architecture also improves the efficiency and scalability of the IDS.

I. INTRODUCTION

Internet is widely spread in each corner of the world; computers all over are exposed to diverse intrusions from the World Wide Web. To protect the computers from these unauthorized attacks, effective intrusion detection systems (IDS) need to be employed. Traditional instance based learning methods for Intrusion Detection can only detect known intrusions since these methods classify instances based on what they have learned. Security of network systems is becoming increasingly important as more and more sensitive information is being stored and manipulated online. Intrusion Detection Systems (IDSs) have thus become a critical technology to help protect these systems. Most IDSs are based on hand-crafted signatures that are developed by manual encoding of expert knowledge. Most IDSs are based on hand-crafted signatures that are developed by manual encoding of expert knowledge. These systems match activity on the system being monitored to known signatures of attacks. The major problem with this approach is that these IDSs fail to generalize to detect new attacks or attacks without known signatures. Recently, there has been an increased interest in data mining- based approaches to building detection models for IDSs. These models generalize from both known attacks and r- mal behavior in order to detect unknown attacks. There is an increasing interest in data mining based approaches to building detection models for IDSs. These models generalize from both known attacks and normal behaviour in order to detect unknown attacks. IDS can also be generated in a quicker and more automated method than manually encoded models that require difficult analysis of audit data by domain experts.IDS better detection

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

performance and generalization ability of data mining based IDSs are some difficulties in the implementation of the system. We discuss several problems inherent in developing and deploying a real-time data mining-based IDS and present an overview of our research, which addresses these problems. These problems are independent of the actual learning algorithms or models used by IDS and must be overcome in order to implement data mining methods in a deployable system. An effective data mining-based IDS must address each of these three groups of issues. Although there are tradeoffs between these groups, each can generally be handled separately. We present the key design elements and group them into which general issues they address.

2. LITERATURE REVIEW

Intrusion detection system plays an important role in detecting malicious activities in computer systems. The following discusses the various terms related to intrusion detection. Intrusion is a type of malicious activity that tries to deny the security aspects of a computer system. It is defined as any set of actions that attempts to compromise the integrity, confidentiality or availability of any resource.

i) Data integrity: It ensures that the data being transmitted by the sender is not altered during its transmission until it reaches the intended receiver. It maintains and assures the accuracy and consistency of the data from its transmission to reception.

ii) Data confidentiality: It ensures that the data being transmitted through the network is accessible to only those receivers who are authorized to receive the respective data. It assures that the data has not been read by unauthorized users.

iii) Data availability: The network or a system resource ensures that the required data is accessible and usable by the authorized system users upon demand or whenever they need it.

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect malicious activities taking place through the network. ID is an area growing in significance as more and more sensitive data are stored and processed in networked systems. Intrusion Detection system is a combination of hardware and software that detects intrusions in the network. IDS monitor all the events taking place in the network by gathering and analyzing information from various areas within the network. It identifies possible security breaches, which include attacks from within and outside the organization and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system.

3. ACCURACY

Crucial to the design and implementation of effective data mining-based IDS is defining specifically how detection performance, or accuracy, of these systems is measured. Because of the difference in nature between a data mining-based system and a typical IDS, the evaluation metrics must take into account factors which are not important for traditional IDSs.

At the most basic level, accuracy measures how well an IDS detects attacks. There are several key components of an accuracy measurement. One important component is detection rate, which is the percentage of attacks that a system detects. Another component is the false positive rate, which is the percentage of normal data that the system falsely ermines to be intrusive. These quantities are typically measured by testing the system on a set of data (normal and intrusions) that are not seen during the training of the system in order to simulate an actual deployment.

There is an inherent tradeoff between detection rate and false positive rate. One way to represent this tradeoff is by plotting the detection rate versus false positive rate on a curve under

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

different parameter values creating a ROC curve1. A method to compare accuracy between two IDSs is to examine their ROC curves.

In practice, only the small portion of a ROC curve corresponding to acceptably low false positives is of interest, as in a deployable system, only a low false positive rate can be tolerated. Hand-crafted methods typically have a fixed detection threshold. They perform at a constant detection rate across different false positive rates. In a ROC curve, we can assume that their curve is a straight line at each detection level. Data mining-based systems have the advantage of potentially being able to detect new attacks that hand-crafted methods tend to miss. Data mining-based IDSs are only useful if their detection rate is higher than a hand-crafted method's detection rate with an acceptably low false positive rate. Given this framework, our goal is to develop a data mining-based IDS that is capable of outperforming hand-crafted signature-based systems at the tolerated false positive rate.

We have developed and applied a number of algorithm- independent techniques to improve the performance of data mining-based IDSs. In this section, we focus on a few particular techniques that have been proven to be empirically successful. We first present a generic framework for extracting features from audit data which help discriminate attacks from normal data. These features can then be used by any detection model building algorithm. We then de- scribe a method for generating artificial anomalies in order to decrease the false positive rate of anomaly detection algorithms. Our research has shown that by generating artificial anomalies, we can improve the accuracy of these ID models. Finally, we present a method for combining anomaly and misuse (or signature) detection models. Typically misuse and anomaly detection models are trained and used in complete isolation from each other. Our research has shown that by combining the two types of models, we can improve the overall detection rate of the system without compromising the benefits of either detection method.

4 Efficiency

In typical applications of data mining to intrusion detection, detection models are produced off-line because the learning algorithms must process tremendous amounts of archived audit data. These models can naturally be used for off-line intrusion detection (i.e., analyzing audit data off- line after intrusions have run their course). Effective intrusion detection should happen in real-time, as intrusions take place, to minimize security compromises. In this section, we discuss our approaches to make data mining-based ID models work efficiently for real-time intrusion detection.

In contrast to off-line IDSs, a key objective of real-time IDS is to detect intrusions as early as possible. Therefore, the efficiency of the detection model is a very important consideration. Because our data mining-based models are computed using off-line data, they implicitly assume that when an event is being inspected (i.e., classified using an ID model), all activities related to the event have completed so that all features have meaningful values available for model checking. As a consequence, if we use these models in real time without any modification, then an event is not inspected until complete information about that event has arrived and been summarized, and all temporal and statistical features (i.e., the various temporal statistics of the events in the past seconds, see Section 2.1) are computed. This scheme can fail miserably under real-time constraints. When the volume of an event stream is high, the amount of time taken to process the event records within the past seconds and calculate statistical features is also very high. Many subsequent events may have terminated (and thus completed with attack actions) when the "current" event is finally inspected by the model. That is, the detection of intrusions is severely delayed. Unfortunately, DoS attacks, which typically generate a large amount of traffic in a very short period time, are often used by intruders to first overload an IDS, and use the detection delay as a window of opportunity to

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

quickly perform their malicious intent. For example, they can even seize control of the host on which the IDS lives, thus eliminating the effectiveness of intrusion detection altogether.

It is necessary to examine the time delay associated with computing each feature in order to speed up model evaluation. The time delay of a feature includes not only the time.

5. USABILITY

A data mining-based IDS are significantly more complex than a traditional system. The main cause for this is that data mining systems require large sets of data from which to prepare. The expectation to reduce the complexity of data mining systems has led to many active research areas [13, 14]. First, management of both training and historical data sets is a difficult task, especially if the system handles many different kinds of data. Second, once new data has been analyzed, models need to be updated. Third, many data mining-based IDSs are difficult to deploy because they need a large set of clean labeled training data. Typically the attacks within the data must either be manually labeled for training signature detection models, or removed for training anomaly detection models.

6. System Architecture

The overall system architecture is designed to support a data mining-based IDS with the properties described throughout this paper. As shown in Figure 2, the architecture 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. In this framework, a model can be anything from a neural network, to a set of rules, to a probabilistic model. To deal with this heterogeneity, an XML encoding is used so each component can easily exchange data and/or models.

Our design was influenced by the work in standardizing the message formats and protocols for IDS communication and collaboration: the Common Intrusion Detection Frame- work (CIDF, funded by DARPA) [29] and the more recent Intrusion Detection Message Exchange Format (IDMEF, by the Intrusion Detection Working Group of IETF, the Inter- net Engineering Task Force). Using CIDF or IDMEF, IDSs can securely exchange attack information, encoded in the standard formats, to collaboratively detect distributed intrusions. In our architecture, data and model exchanged between the components are encoded in our standard message format, which can be trivially mapped to either CIDF or ID- MEF formats. The key advantage of our architecture is its high performance and scalability. That is, all components can reside in the same local network, in which case, the work load is distributed among the components; or the components can be in different networks, in which case, they can also participate in the collaboration with other IDSs in the Internet.

In the following sections we describe the components depicted in more detail. A complete description of the system architecture is given in .

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

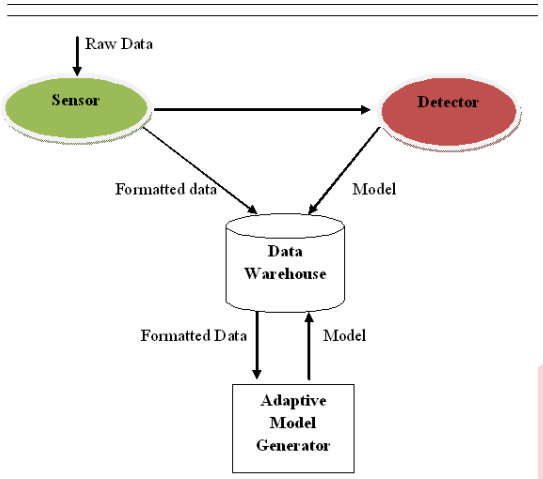

Fig: The Architecture of Data Mining –Based IDS

I. 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 all of the sensors implement a Basic Auditing Module (BAM) framework. In a BAM, features are computed from the raw data and encoded in XML.

II. 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. For example, workloads can be distributed to different detectors to analyze events in parallel. 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. The front-end detectors keep up with high-speed and high-volume traffic, and must pass data to the back-end detector to perform more thorough and time consuming analysis.

III. 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 same type of components, such as multiple sensors, can manipulate data concurrently. Relational database features support "stored procedure calls" which enable easy implementation of complicated calculations, such as efficient data sampling carried out automatically on the server.

7. Where to do Intrusion Detection?

According to the monitored system, the source of input information can be on a host or network. Thus IDS is further classified into three categories as follows:

i) Network-based intrusion detection system (NIDS)

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

It is an independent platform that identifies intrusions by examining network traffic and monitors multiple hosts. Network intrusion detection systems gain access to network traffic by connecting to a network hub, network switch configured for port mirroring, or network tap.

ii) Host-based intrusion detection system (HIDS)

It consists of an agent on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications (binaries, password files, capability databases, Access control lists, etc.) and other host activities and state. In a HIDS, sensors usually consist of a software agent.

iii) Hybrid Intrusion detection system (Hybrid IDS)

It complements HIDS system by the ability of monitoring the network traffic for a specific host; it is different from the NIDS that monitors all network traffic . In computer security, a Network Intrusion Detection System (NIDS) is an intrusion detection system that attempts to discover unauthorized access to a computer network by analyzing traffic on the network for signs of malicious activity.

8. Data Mining and Real Time IDSs

Even though offline processing has a number of significant advantages, data mining techniques can also be used to enhance IDSs in real time. Lee was one of the first to address important and challenging issues of accuracy, efficiency, and usability of real-time IDSs. They implemented feature extraction and construction algorithms for labeled audit data. Ex. entropy, conditional entropy, relative entropy, information gain, and information cost to capture intrinsic characteristics of normal data and use such measures to guide the process of building and evaluating anomaly detection models. A serious limitation of their approaches (as well as with most existing IDSs) is that they only do intrusion detection at the network or system level. However, with the rapid growth of e-Commerce and e-Government applications, there is an urgent need to do intrusion detection at the application-level.

9. RESULT AND DISCUSSION

The result of intrusion detection system is to compare the ROC curve of two systems accuracy is low. Intrusion detection system compares the result in before detection and after detection with accuracy, efficiency, usability of low, medium and high. The result of algorithms will then be scrutinized and techniques for reducing false alarm rate and increasing the accuracy will be tried and tested. Table.1 shows before detection with intrusion detection system. Table.2 shows after intrusion detection system with accuracy is high compare with efficiency and usability.

Table 1: Before Detection

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

	System1			System2		
	Low	Medium	High	Low	Medium	High
Accuracy	No	Yes	No	No	No	Yes
Efficiency	Yes	No	No	Yes	No	No
Usability	No	Yes	No	No	Yes	No

Table 2: After Detection

	System1			System2		
	Low	Medium	High	Low	Medium	High
Accuracy	No	No	Yes	No	No	Yes
Efficiency	Yes	No	No	Yes	No	No
Usability	No	Yes	No	No	Yes	No

10. CONCLUSION AND FUTURE WORK

In this paper, outlined the breadth of research efforts to address important and challenging issues of accuracy, efficiency, and usability of real-time IDSs. To implement the feature extraction and construction algorithms for labeled audit data. The application of Data Mining in Intrusion Detection System is emerging trend in the recent years. The Data Mining techniques can extract characteristics of sample data, thus reduces the difficulties involved in the collection of training data. Thereby achieving the active defense for Intrusion Detection System. So far, there have been very few or no tries at attempting to develop a real-time IDS. All prior systems focus on Offline traffic only. To examined various algorithms like Decision Tree, k-NN, BPNN and Ripper Rule and will try to implement them in practice. Hence, using best algorithms can implement real time online network intrusion detection system. The research efforts on IDSs for e-Commerce and e-Government applications in the near future. This research will be to take traffic that comes directly from the network .We are developing algorithms for data mining over the output of multiple sensors. This is strongly motivated by the fact that single sensors do not typically observe entire attack scenarios. By combining the information from multiple sensors we hope to improve detection accuracy.

REFERENCES

1. E. Eskin. Anomaly detection over noisy data using learned probability distributions 2009.

2. A. Ghosh and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. 2009.
3. W. Lee, S. J. Stolfo, and K. Mok. Data mining in work flow environments: Experiences in

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

intrusion detection. 2010