International Journal of Mathematical Archive-2(10), 2011, Page: 1870-1874
Available online through www.ijma.info ISSN 2229 – 5046

The Role of Data Mining Techniques in Network Intrusion Detection System

[1]Gajanan D. Kurundkar*, [2]Quadri M. N. and [3]Nitin A. Naik

[1]Dept. of Computer Science, Shri Guru Budhiswami Mahavidyalya, Purna Dist. Parbhani (M.S.), India
E-mail: gaju_k_2001@yahoo.com

[2]Dept. of Computer Science, Yeshwant Mahavidyalya, Nanded-(M.S.) India
E-mail: mnq_1977@yahoo.com

[3]Dept.of Computer Science, Yeshwant Mahavidyalya Nanded-(M.S.) India
E-mail: nanresearch@rediffmail.com

(Received on: 28-09-11; Accepted on: 12-10-11)

ABSTRACT

Now a days it's becoming an important task to maintain security for computer system and the data contain in it. Security becomes an important task for everyone, In Information Security; intrusion detection is the process of detecting illegal actions or misuse of user for confidentiality, the process of detecting such kind of activity of unauthorized user. This paper try to focused on data mining techniques that are being used for detecting intruder by using such purposes. In conclusion we near by a new idea on how data mining can support for IDS detection.

Keywords: *Data mining, Security, Intruder, Confidentiality, unauthorized*

I. INTRODUCTION

An intrusion detection system (IDS) monitors network traffic and monitors for doubtful activity and alerts the system or network administrator. In several cases the IDS may also react to abnormal or malicious traffic by taking action such as blocking the user or source IP address from accessing the network.

There are number of ways for identifying intruder main task of IDS is to detect suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS) intrusion detection systems. There are IDS that detect based on looking for specific signatures of known threats- similar to the way antivirus software typically detects and protects against malware and there are IDS that detect based on comparing traffic patterns against a baseline and looking for anomalies. There are IDS that simply monitor and alert and there are IDS that perform an action or actions in response to a detected threat. Different types of IDS types have been discussed in shortly.

1. NIDS: Network Intrusion Detection Systems are placed at a strategic point within the network to monitor traffic to and from all devices on the network. Preferably it is try to scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network. The network-based approached relies on the tcpdump data as input, which gives per packet information. [1]

2. HIDS: Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator of suspicious activity is detected

3. SIGNATURE BASED: A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

4. ANOMALY BASED: An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline.

**Corresponding author: [1]Gajanan D. Kurundkar*, *E-mail: gaju_k_2001@yahoo.com*

What is Data mining: The drawing out of hidden predictive information from large databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining (DM), also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules [3]. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Future analyses offered by data mining move beyond the analyses of past events provided by showing tools typical of decision support systems.

Nearby all companies already collect and process massive quantities of data. Data mining techniques can be implemented speedily on existing software and hardware platforms to improve the value of existing information resources, and can be integrated with new products and systems as they are brought online. When implemented on high performance client/server or parallel processing computers, data mining tools can study huge databases for getting results.

This paper gives an introduction to the basic technologies of data mining. Examples of gainful applications to display its significance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

Architecture for Data Mining

The in general system architecture is designed to support a data mining-based IDS with the properties described all through this paper. As shown in Fig. 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 Framework (CIDF, funded by DARPA) [4] and the more recent Intrusion Detection Message Exchange Format (IDMEF, by the Intrusion Detection Working Group of IETF, the Internet Engineering Task Force).

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 IDMEF 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 Figure 2 in more detail. A complete description of the system architecture is given in [5].

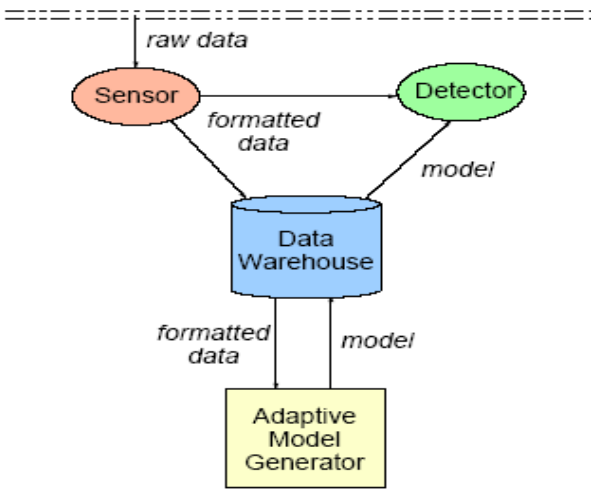

Fig. The Architecture of Data Mining Based IDS.

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 all of the sensors implement a Basic Auditing Module (BAM) framework.

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. 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.

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 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. Arbitrary amount of sensor data can also be retrieved by a single SQL query. Distribution of detection models can be configured to push or pull. 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.

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 (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.[2] It receives audit data for anomalous events (encoded as a GIDO, the Generalized Intrusion Detection Objects) from a detector, computes patterns from the data, compares them with historical normal patterns to identify the "unique" intrusion patterns, and constructs features accordingly. Machine learning algorithms are then applied to compute the detection model, which is encoded as a GIDO and sent to all the detectors. Much of the design and implementation efforts had been on extending the Common Intrusion Specification Language (CISL) to represent intrusion detection models (see [20] for details). Our preliminary experiments show that the model generator is able to produce and distribute new effective models upon receiving audit data.

Data mining for intrusion detection: Intrusion Detection Systems (IDS) have become a standard section in security infrastructures as they allow network administrators to detect rule violations. These policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS have drawbacks:

• Current IDS are usually tuned to detect known service level network attacks. This leaves them vulnerable to original and novel malicious attacks.

• Data overload: Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he 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.

• False positives: A false positive occurs when normal attack is incorrectly classified as malicious and treated for that reason.

• False negatives: This is the case where an IDS does not generate an alert when an intrusion is actually taking place. Data mining can help improve intrusion detection by addressing each and every one of the above mentioned problems. Remove normal activity from alarm data to allow analysts to focus on real attacks

• Identify false alarm generators and bad signatures

• Find anomalous activity that uncovers a real attack

• Identify long, ongoing patterns (different IP address, same activity). To accomplish these tasks, data miners employ one or more of the following techniques:

• Data summarization with statistics, including finding outliers

• Visualization: presenting a graphical summary of the data

• Clustering of the data into natural categories

DATA MINING AND IDS: Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms [6]. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [7]. We are also interested in link and sequence analysis [8][9][10]. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern [11].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.

A. Off Line Processing: The use of data mining techniques in IDSs, usually implies analysis of the collected data in an offline environment. There are important advantages in performing intrusion detection in an offline environment, in addition to the real-time detection tasks typically employed. In off-line analysis, it is assumed that all connections have already finished and, thus, we can compute all the features and check the detection rules one by one. The estimation and detection process is generally very demanding and, therefore, the problem cannot be addressed in an online environment because of the various the realtime constraints [12]. Many real-time IDSs will start to drop packets when flooded with data faster than they can process it. An offline environment provides the ability to transfer logs from remote sites to a central site for analysis during off-peak times.

B. 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 et al. [13] were 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. They developed several anomaly detection algorithms. In the paper, the authors explore the use of information-theoretic measures, i.e., 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.

C. Multi sensor Correlation: The use of multiple sensors to collect data by various sources has been presented by numerous researchers as a way to increase the performance of an IDS. Lee et al.[13], state that using multiple sensors for ID should increase the accuracy of IDSs. Lee et al. note that, “an IDS should consist of multiple cooperative lightweight subsystems that each monitor a separate part (such as an access point) of the entire environment.” • Dickerson and Dickerson [14] also explore a possible implementation of such a mechanism. Their architecture consists of three layers: – A set of Data Collectors (packet collectors) – A set of Data Processors – A Threat analyzer that utilizes fuzzy logic and basically performs a risk assessment of the collected data. • Honig et al. [15] propose a model similar to the one by Dickerson and Dickerson [14] and also has components.

CONCLUSION:

This paper has presented a survey of the data mining techniques that have been planned towards the improvement of IDSs. We have made known the ways in which data mining has been known to support the process of Intrusion Detection and the traditions in which the various techniques have been useful and evaluated by researchers. in conclusion, in the we proposed a data mining approach that we consider can supply significantly in the Attempt to create better and more effective Intrusion Detection Systems.

REFERENCES:

[1] Intrusion Detection Using Datamining Techniques by Anshu Veda(04329022) KReSIT,IIT Bombay, Prajakta Kalekar(04329008) ,KReSIT,IIT Bombay Anirudha Bodhankar(04329003) KReSIT,IIT Bombay.

[2] Real Time Data Mining-based Intrusion Detection Wenke Lee_ , Salvatore J. Stolfo_ , Philip K. Chan , Eleazar Eskin_ , Wei Fan_ , Matthew Miller_ , Shlomo Hershkop_ , and Junxin Zhang.

[3] Data Mining Techniques for (Network) Intrusion Detection Systems by Theodoros Lappas and Konstantinos Pelechrinis Department of Computer Science and Engineering.

[4] S. Stainford-Chen. Common intrusion detection framework. http://seclab.cs.ucdavis.edu/cidf.

[5] E. Eskin, M. Miller, Z.-D. Zhong, G. Yi, W.-A. Lee, and S. Stolfo. Adaptive model generation for intrusion detection. In *Proceedings of the ACMCCS Workshop on Intrusion Detection and Prevention*, Athens, Greece, 2000.

[6] 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.

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

[8] Lee, W. and S. J. Stolfo, "Data mining approaches for intrusion detection", In Proc. of the 7th USENIX Security Symp., San Antonio, TX. USENIX, 1998.

[9] W. Lee, S. J. Stolfo et al, "A data mining and CIDF based approach for detecting novel and distributed intrusions", Proc. of Third International Workshop on Recent Advancesin Intrusion Detection (RAID 2000), Toulouse, France.

[10] Lee, W., S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," In Proc. of the 1999 IEEE Symp. On Security and Privacy, Oakland, CA, pp. 120132. IEEE Computer Society Press, 9-12 May 1999

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

[12] Singh, S. and S. Kandula, "Argus - a distributed network-intrusion detection system," Undergraduate Thesis, Indian Institute of Technology, May 2001.

[13] Lee, W. and D. Xiang, "Information-theoretic measures for anomaly detection", In Proc. of the 2001 IEEE Symp. on Security and Privacy, Oakland, CA, pp. 130143. IEEE Computer Society Press, May 2001

[14] Dickerson, J. E. and J. A. Dickerson, "Fuzzy network profiling for intrusion detection", In Proc. of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta, pp. 301306. North American Fuzzy Information Processing Society (NAFIPS), July 2000.

[15] Honig, A., A. Howard, E. Eskin, and S. J. Stolfo,"Adaptive model generation: An architecture for the deployment of data mining based intrusion detection systems", In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications. Boston: Kluwer Academic Publishers May 2002.

****

