International Journal of Computer Science and Informatics

Volume 2 | Issue 2 Article 10

October 2012

Intrusion Detection System using Genetic Algorithm and Data Mining: An Overview

V.P. Kshirsagar
Dept.of Computer Science & Engg., Govt. College of Engg., An Autonomous Institute Aurangabad (M.S.) IN, vkshirsagar@gmail.com

Sonali M. Tidke
Dept. of Computer Science and Engineering, Govt. Engineering College, Aurangabad (MS), India., sonalitidke11@gmail.com

S.S. Vishnu
Dept.of Computer Science & Engg., Govt. College of Engg., An Autonomous Institute Aurangabad (M.S.) IN, swativishnu@gmail.com

Recommended Citation

Kshirsagar, V.P.; Tidke, Sonali M.; and Vishnu, S.S. (2012) "Intrusion Detection System using Genetic Algorithm and Data Mining: An Overview," *International Journal of Computer Science and Informatics*: Vol. 2 : Iss. 2 , Article 10.
DOI: 10.47893/IJCSI.2012.1076
Available at: https://www.interscience.in/ijcsi/vol2/iss2/10

Intrusion Detection System using Genetic Algorithm and Data Mining: An Overview

Vivek K. Kshirsagar, Sonali M. Tidke & Swati Vishnu

Dept. of Computer Science and Engineering, Govt. Engineering College, Aurangabad (MS), India.
E-mail : vkshirsagar@gmail.com, sonalitidke11@gmail.com, swativishnu@gmail.com

Abstract - Network security is of primary concerned now days for large organizations. Various types of Intrusion Detection Systems (IDS) are available in the market like Host based, Network based or Hybrid depending upon the detection technology used by them. Modern IDS have complex requirements. With data integrity, confidentiality and availability, they must be reliable, easy to manage and with low maintenance cost. Various modifications are being applied to IDS regularly to detect new attacks and handle them. In this paper, we are focusing on genetic algorithm (GA) and data mining based Intrusion Detection System.

Keywords - *Intrusion Detection System, Genetic algorithm, Data mining, Network based IDS, Host based IDS.*

I. INTRODUCTION

Today's network security infrastructure promisingly depends upon Network Intrusion Detection System (NIDS). NIDS provides safety from known intrusion attacks. It is not possible to stop intrusion attacks, so organizations need to be ready to handle them. IDS is a defensive mechanism whose primary purpose is to keep work going on considering all possible attacks on a system.

Intrusion detection is a process used to detect suspicious activity both at network and host level. Two main ID techniques available are anomaly detection and misuse detection. In anomaly based detection system, audit data is used to differentiate abnormal data from normal one. On the other hand, misuse detection system, also called as signature based IDS, uses patterns of well known attacks to match with audit data and identify them as intrusions. Functioning of misuse detection models is in a sense very much similar to that of antivirus applications. Misuse IDS can analyze network or system and compare its activities against signatures of known intrusive computer and network behaviors. For recognizing traffic as attack, IDS must be taught to recognize normal activity. Various ways are available to accomplish this like use of artificial intelligence techniques. Audit data used for testing and creating rules or define patterns can be collected from various sources like network traffic data, system logs from hosts and system calls from various processes. IDS require sensor. Sensor is the system on which an IDS is installed and running. Network sensor monitors network packets like TCP/IP headers, duration of connection, and number of bytes transferred etc. while host sensor monitors system logs, memory usage on host etc.

Figure 1 demonstrates the traditional IDS model. Here sensor machine generates security events, management console monitors those events and controls sensor. The intrusion detector engine records events logged bye the sensor into database and generates alerts based on rules from security events.

In the next sections we describe how IDS can be implemented using data mining and genetic algorithm.

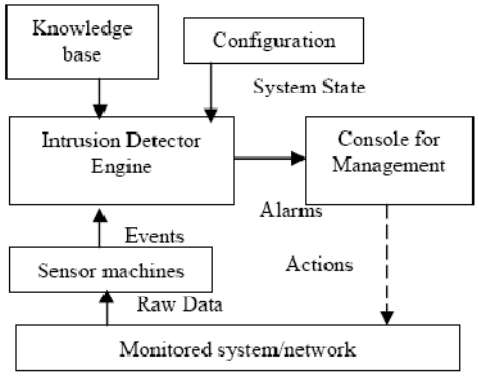


Fig. 1: Traditional IDS model

II. GENETIC ALGORITHM AND IDS

Genetic algorithms [1] employ metaphor from biology and genetics to iteratively evolve a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved and is compose of a fixed number of genes. The number of possible values of each gene is called the cardinality of the gene. Each individual is called as chromosome. The set of chromosomes forms population.

Functioning of genetic algorithm starts with randomly generated population of individuals. Thru various generations these population evolved and individuals' quality gets improved. In every generation, three basic operators of genetic algorithm i.e. selection, crossover and mutation are applied to each individual. Crossover means exchanging the genes between two chromosomes while mutation means random changing of a value of a randomly chosen gene of a chromosome. These individuals are representation of the problem required to be solved. Different positions of each individual can be encoded as bits, characters and numbers [2].

Here, the numbers of best-fit individuals are selected. For this user defined fitness function is used. Fitness function is used to measure quality of each chromosome. Remaining individuals are paired and thru process of crossover, new offspring is produced by partially exchanging their genes. When genetic algorithm is used for problem solving, three factors will have impact on the effectiveness of the algorithm, they are [1]:

a. The selection of fitness function

b. The representation of individuals and

c. The values of the genetic parameters

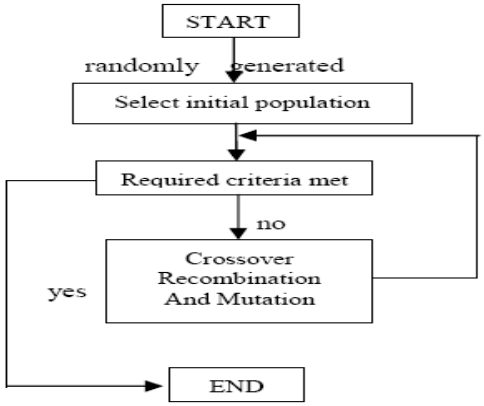


Fig. 2 : Functioning of Genetic Algorithm

Genetic algorithm is used for evolving new rules for IDS. Using these rules normal network traffic or audit data us differentiated from abnormal traffic/data. Rules in the rule set of genetic algorithm are of type if-then. Following is general syntax for rule in genetic algorithm:

if { condition } then { act }

condition refers to the data to be verified and rule in rule set while act is the action to be performed if condition is true. A condition can check for port numbers of network protocols, protocols used, duration of connection, IP address of source and destination etc. while act refers to the action to be performed when condition is true like sending alert message, creating log messages etc.

Benefits of using genetic algorithm for intrusion detection are: [3]

a. Genetic algorithms are intrinsically parallel. Because of multiple offspring, they can explore the solution space in multiple directions at once.

b. Parallelism allows genetic algorithm to implicitly evaluate many schemas at once. This make them well suited to solving problems where space of potential solution is truly huge.

c. Genetic algorithm based systems can be re-trained easily. This improves its possibility to add new rules and evolve intrusion detection system.

A. Data representation in genetic algorithm:

Various network features can be considered for detecting network intrusion like duration of connection, protocols used, source and destination ports, source IP and destination IP, attack- name etc. [4]. If the first six features are connected using logical AND operation to compose the condition part of the rule while feature attack-name is used as act part of rule. Following is the sample example[4] that classifies a network connection as the denial-of-service attack Neptune.

if (duration = "0:0:1" and protocol = "finger" and source_port = 18989 and destination_port = 79 and source_ip = "99.19.99.19" and destination_ip = "192.168.254.10") then (attack_name = "Neptune")

Above rule specifies that if a network packet is originated from IP address 99.19.99.19 and port number 18989 and send to IP address 192.168.254.10 at port number 79 using finger protocol for duration of connection 1 second then most likely it is Neptune attack which eventually make destination host out of service.

Fitness Function:

Every chromosome is selected after applying fitness function to them. If a rule is represented as if A then B [4] then the fitness of the rule is as follows:

support = | A and B | / N

confidence = | A and B | / | A |

fitness = w1 * support + w2 * confidence

Here, N is total number of network connections in audit data, | A | is number of network connections matching the condition A and | A and B | is number of network connections that matches the rule if A then B. The weights w1 and w2 are used to control the balance between two terms.

B. Architecture of genetic algorithm for IDS:

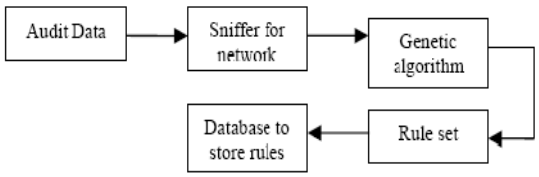


Fig. 3 : Genetic Algorithm for IDS

It requires collecting network data for audit which contains normal and abnormal data. After collecting data, network sniffer will analyze the data and will send it to genetic algorithm. After applying fitness function, rules are added to rule set which are stored in rule base.

III. DATA MINING AND IDS

Data mining is the process of sorting through large amounts of data and picking out relevant information [5]. It helps for extracting hidden useful information from large data warehouses. It helps in predicting future trends and behaviors to help businesses for taking knowledge based decisions. The modern technologies of computers, networks, and sensors have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become useful. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, to data. Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities.

Data mining can contribute in following way to an intrusion detection project [6]:

a. Remove normal activity from alarm data to allow analysts to focus on real attacks.

b. Identify false alarm generators and ”bad” sensor signatures

c. Find anomalous activity that uncovers a real attack

d. Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques: [6]

a. Data summarization with statistics, including finding outliers

b. Presenting a graphical summary of the data

c. Clustering of the data into natural categories

d. Defining normal activity and enabling the discovery of anomalies

e. Predicting the category to which a particular record belongs.

A. Techniques for intrusion detection using data mining

Various techniques can be used for implementing data mining in intrusion detection, each with their own merits. Few of such techniques are presented here:

a. Classification [5] : It is a data mining technique used to map data instances into one of the various predefined categories. It can be used to detect individual attacks but it has high rate of false alarm. Various algorithms like decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are used for classification techniques. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data.

b. Association rule mining : Association describes relationship between various data records. Association rule mining is one of the most popular techniques within data mining. It acts as a sensor which provides source data for meta-learning like techniques which are at higher level of processing. Association rule mining is a slow process and can be replaced by other techniques like classification, clustering etc. An association rule [7] has two parts, an antecedent (if) and a consequent (then). Association rules are created by analyzing data for frequent if/then patterns and support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database and confidence indicates the

number of times the if/then statements have been found to be true. These rules are used for analyzing and predicting the customer behavior.

c. Clustering : In this technique, data points are clustered together based on their similarity factors and is often nearness according to some defined distance. Clustering [8] is an effective way to find hidden patterns in data that humans might miss. It is useful for ID as it can cluster malicious and non-malicious activity separately.

k-means is a clustering algorithm used to cluster observations into different groups of related observations without having prior knowledge about their relationships. Here data is divided in k clusters where k is provided as input.

d. Feature Selection : In this process of machine learning, a set of features from available data is selected and a learning algorithm is trained using selected features for creating classification model. Extraction of features is must as it is not feasible to apply all the available features to learning algorithm. It is also called as feature reduction or variable selection technique.

e. Support Vector Machine (SVM) : It is the technique which maps network connections to the hyper-plane. It attempts to separate data into multiple classes using hyper-plane. SVM algorithm can be modified to operate in the supervised learning domain.

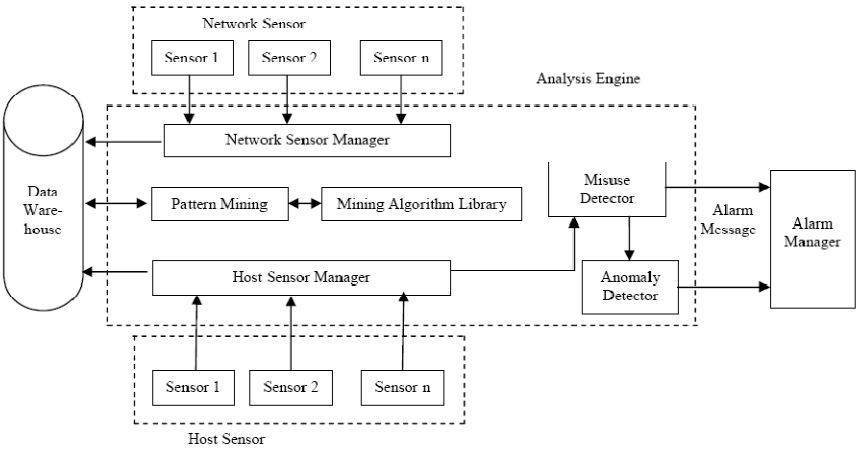
f. Fuzzy logic : Fuzzy logic techniques are being used in computer security since 90's. It allows greater complexity for IDS while it provides some flexibility to the uncertain problem of ID. Most fuzzy IDS require human intervention to determine fuzzy sets and set of fuzzy rules.

g. Meta learning : It is the techniques where new rules are derive from several rule sets which are collected over a period. A meta-rule set [5] relates any two given sets by describing rules that expired, changed, remain unchanged or appeared new.

h. Frequent episodes [5] : It describes relationships in the data stream by recognizing records that occur together. For example, an attack may produce a very typical sequence of records. This technique may produce results for distributed attacks with arbitrary noise inserted within them.

B. Intrusion Detection System model for data mining:

Figure 4 [7] shows various components of Hybrid IDS system in data mining. For network IDS, network sensor and its manager will be required. And to convert hybrid IDS in host based IDS, host sensor and its manager is required, network sensor and its manager will be omitted. Important components in this data mining model are data warehouse, sensors, analysis engine and alarm manager.



Fi g. 4: Data Mining Hybrid IDS

a. Data warehouse : Data warehouse is an overall strategy for building decision support system and knowledge based applications. Analysts can use warehouse for forecasting, competitive analysis and target market research.

b. Sensor : Sensor machines can be network sensor, host sensor or an IDS can have both sensors if it is a hybrid IDS.

c. Analysis engine : Analysis engine consists of network and host sensor manager, pattern mining

and mining algorithm library, misuse detector and anomaly detector. Sensor manager reads data from sensors, analyze them, convert them into database format and store in the data warehouse. Misuse detector and anomaly detector are used to match intrusions with pattern and then alert messages are send to alarm manager unit. Pattern matching and mining algorithm library reads patterns from warehouse and find and match them with intrusions.

d. Alarm manager : Alarm manager has to generate alerts based on how it is implemented.

IV. CONCLUSION

This paper mainly focuses on various IDS models. Various techniques can be used to implement IDS. In the paper, we have mainly concentrated on signature based i.e. misuse detection system. Anomaly based IDS requires to identify new anomalies based on rules stored in IDS while misuse IDS can find only those attacks whose matching rules are already stored in rule set.

REFERENCES

[1] S. Selvakani and R.S. Rajesh, "Genetic Algorithm for Framing Rules for Intrusion Detection" IJCSNS International Journal of Computer Science and Network Security, Vol. 7 No. 11, November 2007.

[2] Wei Li, "Using Genetic Algorithm for Network Intrusion Detection", Department of Computer Science and Engineering, Mississippi, State University, Mississippi State, Ms 39762.

[3] Zorana Bankovic, Jose M. Moya, Alvaro Araujo, Slobodan Bojanic and Octavio Nieto-Taladriz, "A Genetic Algorithm-based Solution for Intrusion Detection", Journal of Information Assurance and Security 4 (2009) 192-199.

[4] Ren Hui Gong, Mohammad Zulkernine, Purang Abolmaesumi, "A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection", Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (CNPD/SAWN '05).

[5] Tamas Abraham, "IDDM: Intrusion Detection using Data Mining Techniques", Information Technology Division, Electronics and Surveillance Research Laboratory, DSTO-GD-0286.

[6] Theodoros Lappas and Konstantinos Pelechrinis, "Data Mining Techniques for (Network) Intrusion Detection Systems", Department of Computer Science and Engineering, UC Riverside, Riverside CA 92521.

[7] Duanyang Zhao, Qingxiang Xu, Zhilin Feng, "Analysis and Design for Intrusion Detection System Based on Data Mining", 2010 Second International Workshop on Education Technology and Computer Science.

[8] S Terry Brugger, "Data Mining Methods for Network Intrusion Detection" University of California, Davis, June 9, 2004.

◈◈◈