Genetic Algorithm Methodology for Intrusion Detection System

Bhavana G.Rathwa[1]
Prof.Purnima Singh[2]

[1] Dept.of Computer Engineering, Parul Institute of Engineering and Technology, Waghodia Gujarat, INDIA

[2] Dept.of Computer Science & Engineering, Parul Institute of Engineering and Technology, Waghodia Gujarat, INDIA

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:-Data mining, Intrusion Detection System (IDS), Genetic Algorithm.

1. INTRODUCTION

In recent years, Intrusion Detection System (IDS) has become one of the hottest research areas in Computer Security. It is an important detection technology and is used as a countermeasure to preserve data integrity and system availability during an intrusion.

When an intruder attempts to break into an information system or performs an action not legally allowed, we refer to this activity as an intrusion. Intruders can be divided into two groups, external and internal. Intrusion techniques may include exploiting software bugs and system misconfigurations, password cracking, sniffing unsecured traffic, or exploiting the design flaw of specific protocols. An Intrusion Detection System is a system for detecting intrusions and reporting them accurately to the proper authority. Intrusion Detection Systems are usually specific to the operating system that they operate in and are an important tool in the overall implementation an organization's information security policy, which reflects an organization's statement by defining the rules and practices to provide security, handle intrusions, and recover from damage caused by security breaches.

There are two generally accepted categories of intrusion detection techniques: misuse detection and anomaly detection. Misuse detection refers to techniques that characterize known methods to penetrate a system. These penetrations are characterized as a 'pattern' or a 'signature' that the IDS look for. The pattern/signature might be a static string or a set sequence of actions. System responses are based on identified penetrations. Anomaly detection refers to techniques that define and characterize normal or acceptable behaviours of the system (e.g., CPU usage, job execution time, system calls). Behaviours that deviate from the expected normal behaviour are considered intrusions.

IDSs can also be divided into two groups depending on where they look for intrusive behaviour: Network-based IDS (NIDS) and Host-based IDS. The former refers to systems that identify intrusions by monitoring traffic through network devices (e.g. Network Interface Card, NIC). A host-based IDS monitors file and process activities related to a software environment associated with a specific host. Some host-based IDSs also listen to network traffic to identify attacks against a host.

2. INTRODUCTION TO GENETIC ALGORITHM

Genetic Algorithms are a searching algorithm designed to mimic the way nature reproduces itself

and betters itself in doing so. Genetic algorithms have several individuals in its population, each one of which could be a potential solution to a problem. Each one of those individuals would be following a path to a possible solution, which means that it is even possible for the search to find more than one solution. Different researchers have produced many Genetic Algorithms, and all of them are very different from each other.

They all, however, display the characteristics of the genetic algorithm, which follow these basic steps:
1. Step one it to randomly create a population of individuals.
2. Step two is to evaluate the population to see which individuals will contribute to the next Generation.
3. Step three is to alter the new generation of individuals once they have been paired off.
4. The final step is to discard the old population and perform step two on the new population.
Once step three, above, has been completed, the algorithm jumps back to step two. The loop will only stop when one of the individuals has been evaluated and is said to be either very close to the solution, or it has found the solution.

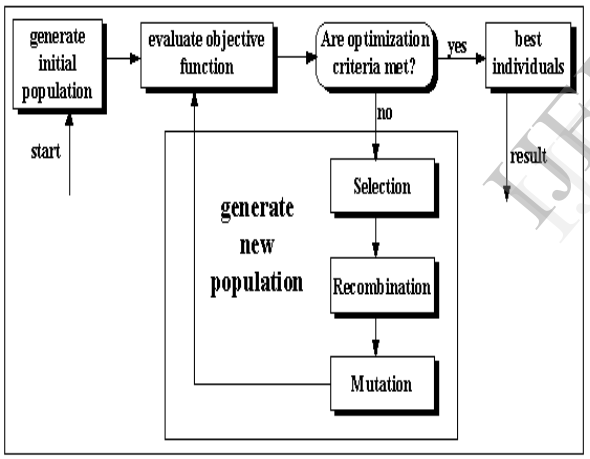

Fig1: Structure of GA

Once the new population has been created, it is time to alter the individuals so that they are different from the other generation. There are two possible operations, which can be performed here, crossover and mutation.
1) Crossover
This occurs when two individuals, which have been paired off, exchange chromosomes. A random number is generated. The number must be between 1 and the 1-(the maximum number of bits in the bit string). The bit string is considered to be the individual and the bits are the chromosomes. Once the random number has been generated, then all digits after that position in the bit string are exchanged.

2) Mutation
Another change an individual may go through is known as mutation. Unlike crossover, it only involves one individual. Depending one what algorithm it is (there are many genetic algorithms), mutation is not very likely to occur. Usually, the possibility of a mutation occurring to one of the chromosomes is set to 1 in thousands.
There are many different genetic algorithms, some of which do not incorporate mutation, and when they do, they handle mutation in a different way to the rest! The basic example may have the computer generate a random number, which decides whether or not a bit is to undergo mutation.
Once a bit is selected for mutation, some algorithms simply which do not incorporate mutation, and when they do, they handle mutation in a different way to the rest! The basic example may have the computer generate a random number, which decides whether or not a bit is to undergo mutation.
Once a bit is selected for mutation, some algorithms simply

3. GENETIC ALGORITHM APPLIED TO INTRUSION DETECTION

Applying genetic algorithm to intrusion detection seems to be a promising area. We discuss the motivation and implementation details in this section.

3.1 OVERVIEW

Genetic algorithms can be used to evolve simple rules for network traffic.

These rules are used to differentiate normal network connections from anomalous connections. These anomalous connections refer to events with probability of intrusions. The rules stored in the rule base are usually in the following form:

If {condition} then {act}

For the problems we presented above, the condition usually refers to a match between current network connection and the rules in IDS, such as source and destination IP addresses and port numbers (used in TCP/IP network protocols), duration of the connection, protocol used, etc., indicating the probability of an intrusion. The act field usually refers to an action defined by the security policies within an organization, such as reporting an alert to the system administrator, stopping the connection, logging a message into system audit files, or all of the above. For Example, a rule can be defined as:

If {the connection has following information: source IP address 124.12.5.18; destination IP address: 130.18.206.55; destination port number: 21; connection time: 10.1 seconds}
Then {stop the connection}

This rule can be explained as follows: if there exists a network connection request with the source IP address 124.12.5.18, destination IP address 130.18.206.55, destination port number 21, and connection time 10.1 seconds, then stop this connection establishment. This is because the IP address 124.12.5.18 is recognized by the IDS as one of the blacklisted IP addresses; therefore, any service request initiated from it is rejected.

The final goal of applying GA is to generate rules that match only the anomalous connections. These rules are tested on historical connections and are used to filter new connections to find suspicious network traffic.

3.2 BENEFITS OF USING GENETIC ALGORITHM FOR INTRUSION DETECTION ARE:

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

4. DATA MINING AND IDS

Data mining is the process of sorting through large amounts of data and picking out relevant information. It helps for extracting hidden useful information from large data warehouses. It helps in predicting future trends and behaviours 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: [5]

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: [5]
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.

4.1 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:

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 neighbour 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. [4]

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 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 behaviour.

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 no 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.

CONCLUSIONS

In this paper, we discussed a methodology of applying classification technique for IDS using data mining, in use genetic algorithm into network intrusion detection techniques. A brief overview of Intrusion Detection System (IDS), genetic algorithm, and related detection techniques are discussed. The system architecture is also introduced. Factors affecting the GA are addressed in detail. This implementation of genetic algorithm is unique as it considers both temporal and spatial information of network connection during the encoding of the problem; therefore, it should be more helpful for identification of network anomalous behaviours.

REFERENCES

[1] Song Naiping, Zhou Genuine," A study on Intrusion Detection Based on Data Mining"2010 International Conference of Information Science and Management Engineering.

[2] Wang Pu and Wang Jun-Qing," Intrusion Detection System with the Data Mining Technologies, "proceedings of the pp.978-1-61284- 486-2/111-2011 IEEE.

[3] YedukondaluGangolu, Ravi Regulagadda, M.Mamatha Devi, Ravindra Changala,"Intrusion Detection System Using Genetic Algorithm" International Journal of Emerging trends in Engineering and Development, Issue 2, Vol.4, May 2012.

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

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