International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

A Review on Intrusion Detection System Based Data Mining Techniques

Shivangee Agrawal[1], Gaurav Jain[2]

[1,2] *PG Scholar, Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior, India*

Abstract - *In today's scenario to maintain the security of network system is important. We need a secure and safe network system towards intruders attack. Intrusion detection system is used for identifying the various types of attack in a network. IDS are available in various types network based, host based and hybrid based on the technology detected by them in market. The existing system does not provide that quality of security, so we need a secure and reliable network system. In this paper, we present a review on intrusion detection system (IDS) using data mining and some optimization techniques to efficiently detect various types of intruder attack.*

Key Words: Intrusion Detection System, Data Mining, Particle Swarm Optimization (PSO), Genetic Algorithm (GA).

1. INTRODUCTION

Data mining is the withdrawal of unseen predictive data or information from a big amount of database. It is strong and novel technology has great prospective to companies focus on greatest important information in their information repository. Data mining tools predict future drift and behaviours through permitting businesses to make knowledge-dive decisions [1].

Data mining mechanism can answer trade or profession questions which have been classically taking a colossal period of time consuming to unravel. In the usual data set, knowledge does not alter with time and their nature is static, whereas streaming information generated regularly. Steady data, i.e. Streaming data is unimaginable to retailer; consequently it required to be analyzed in single pass. [2] [3] [4]. Streaming data can be network data which consists of inbound and outbound traffic of the network.

With network technology increase, nowadays more and more people learn various ways of attack through the rich network resources, and carry out extremely destructive attack through simple operation. In recent years, the amount of hackers' attack is increasing 10 times per year. Therefore, it has become urgent topic to confirm the computer systems, network systems as well as the entire information infrastructure security, and it has become the general concern of the computer industry that how to detect and prevent these attacks effectively.

There are numerous approaches to strengthen the network security at moment, for example encryption, VPN, firewall, etc., but each of these is too static to provide an efficient protection.

However, IDS is a dynamic one, which can provide dynamic protection to the security of network in monitoring, attack and counter-attack.

2. INTRUSION DETECTION TECHNOLOGY

Security had become major concern in all fields of network & system infrastructure [5]. The basic challenge is to authorize user identify & the one who is legitimate to system access without abusing their privileges. Insider threats as well as outsider threats are rigorous to the system/network, known as intruders. Intrusion detection methodology can be described as a method that classifies and deals with the malicious use of network and computer resources. It contains the exterior method behavior of intrusion and internal user's non-authorized. It is a methodology designed to confirm about security of computer system that can discover and inform the non-authorized and abnormal occasions, used to detect the violation of network security. Fig. 1 depicts a high level architecture of generic IDS that protects a network.

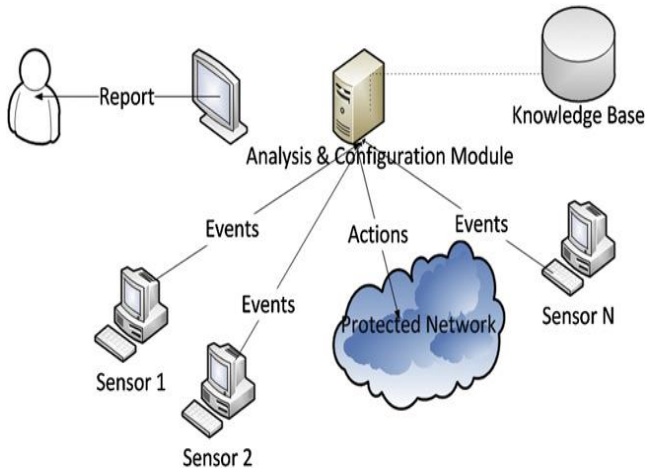

Fig -1: Architecture of a typical IDS.

An IDS is critical technology to detect such intruders who are system harmful. Basic aim of the IDS is to protect the system & network from the intruders. IDS keep track of activities behavior; if they are system malicious then it'll be automatically detected through the IDS [6].

Thus IDS is further categorized into three classes as follows[7]:

1.1 Network Based IDS

It is a platform independent that classifies intrusions through examining traffic network and monitors numerous hosts. NIDS increase access to network traffic through network hub connecting, port mirroring network switch configured, or network tap.

1.2 Host Based IDS

It is an agent contain on a host that classifies intrusions through system calls analyzing, application logs, modifications of file-system (Access control lists, binaries, password files, capability databases, etc.) and other different host state and activities. In a HIDS, sensors commonly consist of a software agent.

1.3 Hybrid IDS

It complements HIDS system through the ability the monitoring network traffic for a particular host; it is various from the NIDS that monitors all network traffic. In computer security, a NIDS is an IDS that attempts to discover unauthorized access to a computer network through analyzing traffic on the network for signs of malicious activity.
In the detecting data target, intrusion detecting method can confidential as host-based, network-based, kernel-based and application-based.

1.4 DRAWBACKS OF IDS

IDS have become a standard factor in the security structures as they permit networks administrators to the detect policy variations. These policy violations range from attackers of external trying to improvement unauthorized access to intruders abusing their access. Present IDS have an amount of significant disadvantages [8]:

1 False positives

A common complaint is the false positives quantity IDS will produce. It is the most challenging task for developing unique signature. Here valid intrusion attempt if a signature also alerts commonly on valid network activity.

2 False negatives

Detecting attack for which there aren't any identified signatures. This leads to the opposite inspiration of false negatives where identification does no longer generate an alert when an intrusion is virtually taking position. Without problems put, if a signature has not been written for a precise take advantage of there is a first-rate hazard that IDS won't observe it.

3 Data overload

Another aspect does not relate to the directly misuse detection but it is particularly important is how much knowledge an analyst can effectively and efficiently analyze.

Data mining can help to the expand intrusion detection with the aid of addressing each and every one among above stated problems. To accomplish these duties, data miners employ one or more of the following tactics:

- knowledge summarization with information, including finding outliers
- Visualization: supplying a graphical abilities abstract
- Clustering of the data into natural categories
- Association rule discovery: describing average endeavor and enabling anomalies discovery
- Classification: predicting style to which a special record belongs.

2. APPLICATION OF DATA MINING IN INTRUSION DETECTION

In classical IDS, security experts firstly categorize attacking actions and system weakness, select statistical approaches because of detecting kinds, then manually enter code and establish the corresponding detecting rules and modes. For complex network system, the limitation of experts' knowledge grows with the change of time and space, so it is not good to increase the effectiveness of detecting the intrusion detecting modes. Safety experts most often predicament in regards to the known attacking aspects and approach weak spot and study on that, which motives the dearth of adaptability of the detecting sample to the unknown intrusion that procedure is set to be dealing with. Meanwhile, lengthy upgrade protection method cycle, the excessive price, these aren't fine for bettering the adaptability of intrusion detecting pattern.

As the experts' rules and statistical approaches often need hardware and software support, it stops the system from reusing and increasing in novel atmosphere, meanwhile it causes the difficulty of embedding new detecting modules. All of these are not good for gaining scalability of intrusion detecting pattern. Therefore, it has become significant issue how to establish an effective, self-adaptable and scalable intrusion detecting pattern in intrusion detecting field. Considering intrusion detection as a data analysis procedure through using data mining predominance in its effective use of knowledge, this is a technique that can automatically create accurate and applicable intrusion patterns from massive audit data, which creates IDS can be useful to any computer environment. This method has become a famous research topic, in inter discipline field of network security

and AI. The analysis association approaches, sequence, classification and clustering in data mining has been proved possible [9].

Intrusions are the activities that violate the security norms of system. An IDS is Mechanism used to identify, screen network or process actions for malicious hobbies and produces reviews to a administration departments. The development of IDS is influenced through following causes: Most current methods have protection was once that render them susceptible to intrusions, and fixing and finding each these deficiencies aren't viable. Prevention methods cannot be ample. It's close to inconceivable to have an undoubtedly relaxed procedure. Even essentially the most secure systems are prone to insider attacks. New intrusions always emerge and novel ways are required to defend towards them.

3. LITERATURE SURVEY

Priyanka Pawar et al.[10] presents the performance of Neural Network for various values of number of clusters, based on experiments. The optimization of output is done using Particle Swarm Optimization (PSO) by selecting initial through PSO. Particle Swarm Optimization is used to optimize the output of our system, by appropriate selecting the input parameters through PSO. An algorithm based on the Particle Swarm Optimization and Neural Network for analyzing program behaviour in intrusion detection is evaluated by experiments. Preliminary experiments with KDD cup'99 Data set show that the PSO optimized Neural Network can effectively detect intrusive attacks and achieves a low false positive rate.

Ketan Sanjay Desale et al. [11] presents the mechanism to improve the efficiency of the IDS using streaming data mining technique. They apply four selected stream data classification algorithms on NSL-KDD datasets and compare their results. Based on the comparative analysis of their results best method is found out for efficiency improvement of IDS.

Seyed Mojtaba Hosseini Bamakan et al. [12] presents a new method based on multiple criteria linear programming and particle swarm optimization to enhance the accuracy of attacks detection. Multiple criteria linear programming is a classification method based on mathematical programming which has been showed a potential ability to solve real-life data mining problems. However, tuning its parameters is an essential steps in training phase. Particle swarm optimization (PSO) is a robust and simple to implement optimization technique has been used in order to improve the performance of MCLP classifier. KDD CUP 99 dataset used to evaluate the performance of proposed method. The result demonstrated the proposed model has comparable performance based on detection rate, false alarm rate and running time compare to two other benchmark classifiers.
Jaina Patel et al. [13] proposed a hybrid model that integrates Anomaly based Intrusion detection technique with Signature based Intrusion detection technique is divided into two stages. In first stage, the signature based IDS SNORT is used to generate alerts for anomaly data. In second stage, data mining techniques "k-means + CART" is used to cascade k-means clustering and CART (Classification and Regression Trees) for classifying normal and abnormal activities. The hybrid IDS model is evaluated using KDD Cup Dataset. The proposed assemblage is introduced to maximize the effectiveness in identifying attacks and achieve high accuracy rate as well as low false alarm rate.

G.V. Nadiammai et al. [14] solved four issues such as Classification of Data, High Level of Human Interaction, Lack of Labeled Data, and Effectiveness of Distributed Denial of Service Attack using the proposed algorithms like EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPERAA Algorithm respectively. Our proposed algorithm has been tested using KDD Cup dataset. All the propose algorithm shows better accuracy and reduced false alarm rate when compared with existing algorithms.

4. IDS WITH OPTIMIZATION TECHNIQUES

4.1 Genetic algorithm

In 1970 John Holland discovers genetic algorithm. GA is an evolutionary algorithm that is based on the survival of fittest. GA finds the optimal solution by searching the solution space. The GA first creates a population of possible solutions then find the optimal solution from the search space by evaluating and using three operators namely crossover, mutation and selection [15].

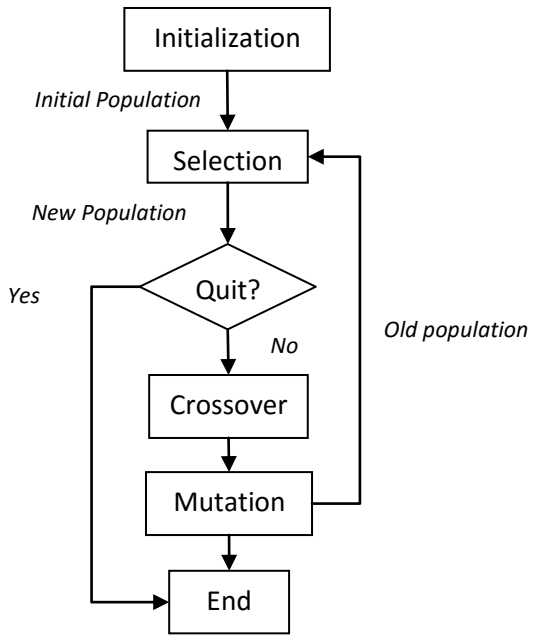

Fig -2: Genetic Algorithm Flow Chart

1. Architecture of 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 [16].

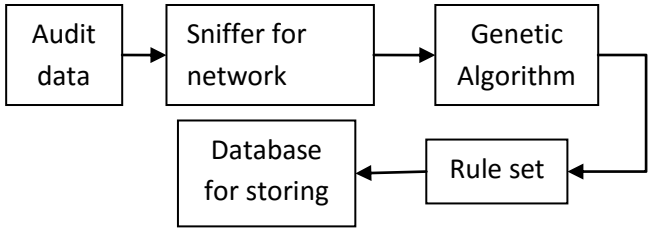

Fig -3: IDS using genetic algorithm

2. Rule representation

New rules for IDS can be generated by using genetic algorithm. These rules can differentiate normal data from abnormal data. The rules in the rule set of genetic algorithm are in the form of IF-THEN. The syntax for the rule is given below:

IF {condition} THEN {act}

When condition is true then act is to be performed. In condition there can be duration, protocol, source port number, destination port number, source IP address, destination IP address and if the condition is true then act can be sending alert message, creating log messages, stopping the connection, etc.

3. Data representation

Intrusion can be detected by considering the various network features like duration, protocol, source and destination port number, source and destination IP address, attack name etc. For example:

If {the connection has following information: duration = "0:0:23" and protocol = "TCP" and source_ port number = "1906" and destination_ port number = "23" and source IP address = "192.168.1.30" and destination IP address = "192.168.0.20"} then {stop the connection}.

Table -1: Rule definition for connection and range of values of each field

Attribute	Range of values	Example values
Source IP address	0.0.0.0~255.255.255.255	192.168.1.30
Destination IP address	0.0.0.0~255.255.255.255	192.168.0.20
Source port number	0~65535	01906
Destination port number	0~65535	00023
Duration	0~99999999	00000023
protocol	0~9	2

Chromosome form of the above example:
(1, 9, 2, 1, 6, 8, 1, 3, 0, 1, 9, 2, 1, 6, 8, 0, 2, 0, 0, 1, 9, 0, 6, 0, 0, 0, 2, 3, 0, 0, 0, 0, 0, 0, 2, 3, 2)

The above rule can be explained as follows: if a network connection has duration time 23 seconds, uses protocol type 2 (TCP), originated from source IP address 192.168.1.30 and port number 1906 for destination IP address 192.168.0.20 and port number 23 then the suspicious behavior is indicated and can be identified as a potential intrusion. This can be identified by matching the rule with the historical data set in which connections are stored marked as anomalous and normal behavior. But a single rule cannot distinguish between anomalous or normal connection. For this a population needs to be evolved to find the optimal rule set.

4. Algorithm

In first step the initialization of the population is performed in which random value is initialized to each gene. Then the evolution of population is performed to several iterations. For every iteration the fitness value of each rule is calculated according to fitness function, the rules which have the highest fitness value are selected and finally the genetic operators, crossover and mutation are applied. Finally the algorithm generates rules for intrusion detection.

1. Initial population
   1.1. define rules
   1.2 Calculate fitness value for each rule.
2. For i = 1 to no. of rules in population
   2.1. do selection for selecting parent 1 and parent 2
   2.2. do crossover for creating child 1 and child 2
   2.3. do mutation
   2.4. Reinsert parent 1, parent 2, child 1 and child 2 to the population

5. Operators

The search capability and convergence of the algorithm is determined by the genetic algorithm. Genetic operators hold the selection, encoding of chromosomes, crossover and mutation on the population and generate the new population [15].

In selection process, chromosomes are selected from the population according to some probability.

For example, Boltzmann selection, roulette wheel selection, rank selection, tournament selection etc. are some methods for selecting chromosomes from the population.

In crossover process, a child is produced from more than one parent solutions. There are different methods like one point crossover, two point crossovers, uniform crossover etc.

Table -2: Example: single point crossover

Chromosome 1	10110|010
Chromosome 2	10101|111
Child 1	10110|111
Child 2	10101|010

In mutation, one or more gene value of chromosome can change. The new solution can change from the previous solution. There are different methods available, for example flipping, interchange mutation, reversing mutation, bit string mutation etc.

Table -3: Example: bit string mutation

Chromosome 1	10110101
Child 1	10111101

6. Fitness Function

Performance of genetic algorithm is dependent on the calculation of fitness value. For calculating the fitness value of the rules, following fitness function can be used.

$$\text{Fitness value} = (a/A)-(b/B) \quad \ldots 1$$

Where, a gives the total number of correctly detected attacks, A is the total number of attacks in the datasets, b is the total number of false positives (normal connections falsely identified as attacks), and B is the total number of normal connections in the dataset. [17].

4.2 Particle Swarm Optimization (PSO)

In 1995 Eberhart and Kennedy introduced particle swarm optimization. In particle swarm optimization particle refers to the each individual of the population. When the initialization of particle is done, each particle updates its position and velocity according to their local best position (pbest) and global best position (gbest) of all particles [18,19].

Particle swarm optimization algorithm steps:

1. Initialization of position and velocity of particles with randomly chosen value.
2. Find fitness value of each particle according to fitness function.
3. If fitness value of particle i is better than the pbest then update pbest = fitness value.
4. If pbest is updated and it is better than current gbest then update gbest = pbest.
5. Update velocity and position of particle according to equation (a) and (b).
6. If the best fitness value or stopping criteria is reached then stop the process, otherwise repeat the process from step 2.

For updating particle's velocity

$$Vi[t+1] = w.Vi[t] + c1.rand1\ (pi,best[t] - pi,current[t]) + c2.rand2\ (pg,best[t] - pi,current[t]) \quad \ldots.(2)$$

For updating particle's position

$$pi[t+1] = pi[t] + Vi[t+1] \quad \ldots.(3)$$

1. Flow Chart

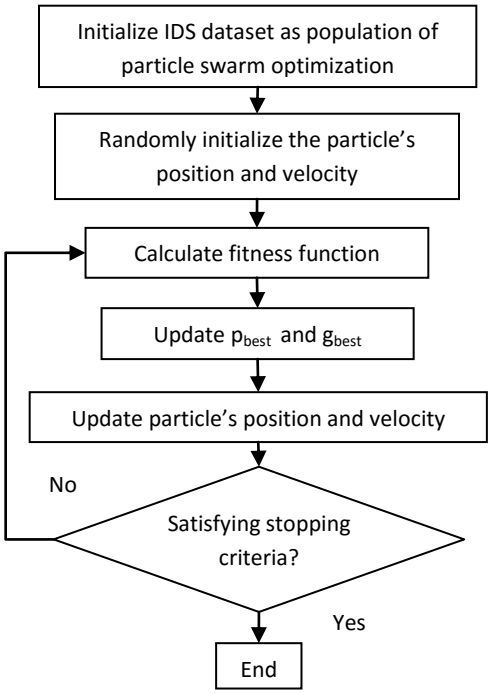

Fig -4: Flow chart of IDS using PSO

5. CONCLUSION

This paper shows the study about intrusion detection system with its application and drawback. We focus on genetic based intrusion detection and other swarm intelligence based technique so that performance of IDS can improve. The IDS with genetic algorithm and particle swarm optimization is also explained by flow chart. How to represent chromosome in GA is also explained in brief.

REFERENCES

[1] G. Trupti Phutane and Apashabi Pathan, "A Survey of Intrusion Detection System Using Different Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 11, November 2014, pp: 6801-6807

[2] Anthony Raj. A, "A Study on Data Mining Based Intrusion Detection System", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 1 (March 2014), pp: 21-25

[3] Harshna and Navneet Kaur, "Survey paper on Data Mining techniques of Intrusion Detection", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013, pp: 799-802

[4] Changxin Song and Ke Ma, "Design of Intrusion Detection System Based on Data Mining Algorithm," Proceedings of 2009 International Conference on Signal Processing Systems, IEEE 2009, pp. 307-373.

[5] Manikandan R, Oviya P and Hemalatha C, "A New Data Mining Based Network Intrusion Detection Model," Journal of Computer Application, Volume 5, Issue EICA2012-1, pp. 1-10 February 10, 2012.

[6] Daejoon Joo, Taeho Hong and Ingoo Han, "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors," Expert System with Applications 25, 2003, pp.69-75.

[7] Wenke Lee and Salvatore J.Stolfo, "Data Mining Approaches for Intrusion Detection," Proceedings of the 7th USENIX Security Symposium San Antonio, Texas, January 26-29, 1998.

[8] Sahilpreet Singh and Meenakshi Bansal,"A Survey on Intrusion Detection System in Data Mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume No. 2, Issue No. 6, June 2013, pp: 2190- 2194

[9] Sunita Jahirabadkar and Parag Kulkarni (2013) "Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Vol 63– No.20, pp. 29-35.

[10] Priyanka Pawar and Damodar Tiwari, "Intrusion Detection System based on Particle Swarm Optimized Neural Network", International Journal of Digital Application & Contemporary Research, Volume 4, Issue 11, June 2016.

[11] Ketan Sanjay Desale, Chandrakant Namdev Kumathekar and Arjun Pramod Chavan, "Efficient Intrusion Detection System using Stream Data Mining Classification Technique", International Conference on Computing Communication Control and Automation, 2015 IEEE.

[12] Seyed Mojtaba Hosseini Bamakan, Behnam Amiric, Mahboubeh Mirzabagherib and Yong Shia, "A New Intrusion Detection Approach using PSO based Multiple Criteria Linear Programming", Information Technology and Quantitative Management (ITQM 2015), © 2015 The Authors. Published by Elsevier B.V.

[13] Jaina Patel and Mr. Krunal Panchal, "Effective Intrusion Detection System using Data Mining Technique", Journal of Emerging Technologies and Innovative Research (JETIR), June 2015, Volume 2, Issue 6.

[14] G.V. Nadiammai and M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal (2014) 15, 37–50.

[15] Li, Wei. (2004). Using genetic algorithm for network intrusion detection.

[16] Vivek K. Kshirsagar, Sonali M. Tidke and Swati Vishnu, "Intrusion Detection System using Genetic Algorithm and Data Mining: An Overview", International Journal of Computer Science and Informatics ISSN (PRINT): 2231 – 5292, Vol-1, Iss-4, 2012.

[17] V. Moraveji Hashemi, Z. Muda and W. Yassin, "Improving Intrusion Detection Using Genetic Algorithm", Information Technology Journal 12(11): 2167-2173,2013.

[18] Shivangee Agrawal and Vikas Sejwar, "Frequent Pattern Model for Crime Recognition", International Journal of Computational Intelligence Research, Volume 13, Number 6 (2017), pp. 1405-1417.

[19] Shivangee Agrawal and Nivedita Bairagi, "A Novel Approach for Association Rule Mining using Modified Shuffled Frog-Leaping Algorithm", International Journals of Advanced Research in Computer Science and Software Engineering, Volume 7, Number 8 (2017).

[20] Shivangee Agrawal and Vikas sejwar, "Crime Identification using FP-Growth and Multi Objective Particle Swarm Optimization", in ICEI 2017 unpublished.