International Research Journal in Advanced Engineering and Technology (IRJAET)

ISSN (Print) : 2454-4744 | ISSN (Online) : 2454-4752 (www.irjaet.com)

Vol. 1, Issue 1, pp.74-82, June, 2015

A new approach of an Efficient Network Intrusion Detection System Based on Genetic Algorithm

S.Vijayarangam
Research Scholar, St.Peter's University, Avadi, Chennai
Dr.A.Rajesh
Supervisor, C.Abdul Hakeem College of Engineering and Technology

ARTICLE INFO

Article History:

Received 7th, Sep, 2015
Received in revised form 10th, Sep, 2015
Accepted 24th, Sep, 2015
Published online 26th, Sep, 2015

Keywords:

Algorithm, Intrusion Detection, KDD99, Training Accuracy.

ABSTRACT

The need for an efficient Intrusion Detection System arises to preserve data integrity and system reliability for computer network and security. Though the internet has been designed to withstand various forms of failure, the intrusion detection tools and attacks are becoming increasingly sophisticated, exposing the internet to new threats. Under this decisive factor we propose a novel Genetic Algorithm for intrusion detection to effectively detect various types of computer network intrusions. Here we propose a novel feature extraction and classification algorithm. First the binary host event data is first processed into ASCII information; this is reviewed in terms of service and duration and summarized into host session records. The uniqueness to reduce the computational over head is maintained by applying our data mining algorithm after the preprocessing stage. Then the classification rule is applied to learn the detection model. KDD99 data set has been used for benchmark test verification. To evaluate our system, standard metric like training accuracy, testing accuracy, detection rate and false positive rate will be estimated for fair comparison. And it is found that, the proposed system delivers better accuracy under 21 out of 41 features testing case.

INTRODUCTION

Computer networks are usually protected by a number of access restriction polices (anti-virus software, firewall, message encryption, secure network protocols, password protection, etc.) Since it has been proven that a potential attacker can always find a way to infiltrate into a network, there is a need for additional support that would detect this type of security breaches. These systems are known as intrusion detection systems (IDS) and are placed inside the protected network, looking for potential threats in network traffic and/or audit data recorded by hosts.

IDS have three common issues: speed, accuracy and adaptability. The speed issue arises from the extensive amount of data that needs to be monitored in order to observe the entire situation. An existing approach to solving this problem is to split the network stream into few more manageable streams and analyze each in real-time using separate IDSs. The event stream must be split in a way that covers all relevant attack scenarios, but this assumes that all the attack scenarios must be known a prior.

We are deploying a different approach. Instead of defining different attack scenarios, we exact the features of network traffic that are likely to take part in an attack. This provides higher flexibility since a feature can be relevant for more than one attack or is prone to be abused by an unknown attack. Moreover, we need only one IDS to perform the detection. Finally, in this way the total amount of data to be processed by the IDS is highly reduced. Consequently, the amount of time spent for offline training of the system and afterwards the time spent for attacks detection are also reduced.

Integration of learning algorithms provides a potential solution for the adaptation and accuracy issues. Many different solution based on machine learning techniques have been deployed for intrusion detection in both commercial systems and the state-of-art. These techniques introduce certain amount of intelligence in the detection process, and the capable of processing large amount of ‘intelligence’ in the detection process, and are capable of processing large amount of data much faster than a human. However, these systems often introduce significant computational overhead. Furthermore, many of them do not deal property with so called ’rare’ classes i.e. the classes that have significantly smaller number of elements then the rest of the classes. This problem occurs mostly because of the tendency for generalization that most of these techniques exhibit. Intrusions can be considered rare classes since it is reasonable to assume that the amount of intrusive traffic is considerably smaller than the amount of normal traffic. Thus, we need a machine learning technique that is capable of dealing with this issue.

In this work we are presenting a genetic algorithm (GA) approach for classifying network connections. GAs are robust, inherently parallel, adaptable and suitable for dealing with the classification of rare classes. Moreover, due to its inherent parallelism, it offers a possibility to implement the system using reconfigurable devices without the need of deploying a microprocessor. In this way, the implementation cost would be much lower than the cost of implementing traditional IDS providing at the same time higher level of adaptability, as these devises can be dynamically reconfigured.

This work represents a continuation of our previous one where we investigated the possibilities of applying GA to intrusion detection while deploying small subset of features. The experiments have confirmed the robustness of GA and inspired us to further continue experimenting on the subset.

Here we further investigate a combination of two GA based intrusion detection systems. The first system in the line is an anomaly-based IDS implemented as a simple linear classifier. This system exhibits high both detection and false-positive rate. For that reason, we have added a simple system based on if-then rules that filter the detection of the linear classifier and in that way significantly reduces false-positive rate. We actually create a strong-classifier built upon weak-classifiers, but without the need to follow the process of boosting algorithm as both of the created systems can be trained separately.

For evolving our GA-based system KDD99 cup training and testing dataset was used. KDD99 cup dataset was found to have quite drawbacks, but it is still prevailing dataset used for training and testing of IDSs due to its good structure and availability. Because of these shortcomings, the presented results do not illustrate the behavior of the system in a real-world environment, but they do reflect its possibilities. In 1987 Dorothy E. Denning proposed intrusion detection as is an approach to counter the computer and networking attacks and misuses [1]. Intrusion detection is implemented by an intrusion detection system and today there are many commercial intrusion detection systems available. To evaluate the capability of intrusion detection system, two key indicators called Detection Rate (DR) and False Positive Rate are needed [2].

In the beginning days, the researchers focused on rule based and statistical intrusion detection systems. But, the results become unsatisfactory, with large datasets. After that, artificial intelligence based techniques have been introduced, and shows certain improvement in detecting the intrusions [3]. Most of the existing techniques make great efforts to achieve a single solution [4]. There was no single technique to detect all kinds of attacks to a certain level of detection accuracy and false alarm rate [5], also they are not capable of modeling correct hypothesis space of the problem [6]. Some of the techniques are unstable; some could not process the larger size like high dimensional data [7]. This paper proposes a novel Genetic Algorithm which utilizes some existing feature extraction techniques to reduce the amount of data in the process.

LITERATURE REVIEW

In the recent past there has been a growing recognition of deploying intelligent techniques for the creation of efficient and reliable intrusion detection systems. A complete survey of these techniques is hard to be presented at this point, since there are more than hundred IDSs based on machine learning techniques. Some of the best-performed techniques used in the state-of-the-art apply GA, combination of neural networks, genetic programming (GP) ensemble, support vector machines, fuzzy logic, clustering techniques, hidden Markov models, junction tree algorithm, Naïve Bayes Classifier, ant colonies, etc.

All of the techniques mentioned above have two steps: training and testing. The systems have to be constantly retrained using new data since new attacks are emerging every day. The advantage of all GA or GP-based techniques lies in their easy retraining. It’s enough to use the best population evolved in the previous iteration as initial population and repeat the process, but this time including new data.

Previous studies in the intrusion detection field have come across many techniques to generate effective ensembles. Ensemble techniques or classifiers have been applied to overcome the limitations of a single classifier [8]. Roli [9] proposed a multi classifier based system of neural networks. The various neural networks were used different features of KDD cup 99 datasets. This paper concluded that a multi strategy combination technique like belief function outperforms other representative techniques. In [10] J. Gomez et al. proposed a linear representation scheme for evolving fuzzy rules using the concept of complete binary tree structure. Genetic Algorithm is used to generate genetic operators for producing useful and minimal structural modification to the fuzzy expression tree represented by chromosomes. The biggest limitation of the proposed approach was that the training was time consuming. Middlemiss et al [11] have used GA for weighted feature extraction with specific application to intrusion detection. A k-nearest neighbor classifier was used for the fitness function of GA as well as to evaluate the performance of the new weighted feature set. These weighted features are used to scale the input variables provided to the classifier system.

Xiao et al. [12] detect anomalous network behavior based on information theory using Genetic Algorithm. Some network features can be identified with network attacks based on mutual information between network features and type of intrusions and then using these features a linear structure rule and also a Genetic Algorithm is derived. The approach of using mutual information and resulting linear rule seems very effective because of the reduced complexity and higher detection rate. The main problem is, it considered only the discrete features. This paper measured the fitness of a chromosome using the standard deviation equation with respect to distance. But the detection rate was poor and they failed to reduce the false positive rate in Intrusion Detection System.

Finally, if we want to consider the possibility of hardware implementation using reconfigurable hardware, not all the systems are appropriate due to their sequential nature or high computational complexity. Due to the parallelism of our algorithm a hardware implementation using reconfigurable devices is possible. This can lead to lower implementation cost with higher level of adaptability compared to the existing solutions and reduced amount of time for system training and testing.

In short, the main advantage of our solution lies in the fact that it includes important characteristics (high accuracy and performance, dealing with rare classes, inherent adaptability and feasibility of hardware implementation) in one solution. We are not familiar with any existing solution that would cover all the characteristics mentioned above.

GENETIC ALGORITHMS FOR INTRUSION DETECTION

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics. The most important idea that stands beyond the initial creation of GAs is the aim of developing a system as robust and as adaptable to the environment as the natural systems. GA operates on a population of potential solutions applying the principle of the survival of the fittest to produce better and better approximations to the solution of the problem that GA is trying to solve. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness value in the problem domain and breeding them together using the operators borrowed from the genetic process performed in the nature, i.e. crossover and mutation. This process leads to the evolution of the populations of individuals that are better adapted to their environment than the individuals that they were created from, just as it happens in natural adaptation.

Intrusion Detection Systems – Types and Issues

According to the detection mechanism they use, there are two general categories of IDSs: misuse detection and anomaly based. Misuse detection systems detect intruders with known patterns. As only the attacks that already exist in the attack database can be detected, this model needs continuous updating. Their advantage is very low false positive rate. Anomaly detection systems identify deviations from an established normal behavior and alert to potential unknown or novel attacks without having any prior knowledge of them. They exhibit higher rate of false alarms, but they have the ability of detecting unknown attacks.

Another classification of IDSs is established by the resource they monitor. According to this classification, IDSs are divided into two categories: host based and network based. Host based intrusion detection systems monitor host resources for intrusion traces whereas network based intrusion detection systems try to find intrusion signs in the network data. The current trend in intrusion detection is to combine both host based and network based information to hybrid systems and therefore not rely on only one methodology.

As already stated in the introduction, IDSs have three common problems: speed, accuracy and adaptability. The speed problem arises from the extensive amount of data that intrusion detection systems need to monitor in order to observe the entire situation. In order to cope with it, the most important piece of information should be extracted so as to provide efficient detection of attacks. The adaptation and accuracy issues of the intrusion detection can be solved by in cooperating learning algorithms. In the case of intrusion detection, to learn means to discover patterns of normal behavior or pattern of attacks. This formulation of intrusion detection problem combines the advantages of signature-based and anomaly-based IDS. Thanks to the generalization capability of learning algorithms, it is also possible to detect new attacks that exploit the same vulnerabilities of known attacks.

Genetic Algorithm Overview

GA evolves a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved. Each individual is called chromosome, and is composed of a certain number of genes that in general case does not have to be fixed. The quality of each rule is measured by a fitness function which is the quantitative representation of each rule’s adaptation to the environment, i.e. the problem to be solved. The procedure starts from an initial population of randomly generated individuals. Then the population is evolved for a number of generations while the qualities of the individuals are being gradually improved in the sense of increasing the fitness value as the measure of quality. During each generation, three basic genetic operators are sequentially applied to each individual with certain probabilities, i.e. Selection, crossover and mutation. Crossover consists of exchanging of the genes between two chromosomes performed in a certain way, while mutation consists of random changing of a value of a randomly chosen gene of a chromosome. Both crossover and mutation are performed with a certain possibility, called crossover/mutation rate.

Understanding Kddcup99 Data

KDD99 is built based on the data captured in DARPA’98 IDS evaluation program. DARPA’98 is about 4 gigabytes of compressed raw (binary) tcp dump data of seven weeks of network traffic that can be processed into about 5 million connection records and each about 100 bytes. And the remaining two weeks of test data have around 2 million connection records. The KDD training dataset was approximately consists of 4,900,000 single connection vectors each of which contains 41 features (34 features are numeric and 7 features are symbolic) and is labeled as either normal or an attack, with exactly one specific attack type [16].

KDD has evolved from interaction and cooperation among such different fields like pattern recognition, database, statistics, AI, and knowledge acquisition for intelligent systems. Discovering a high level knowledge from lower levels of relatively raw data, or to discover a higher level of abstraction and interpretation than those previously known, is the main idea in KDD. KDD applies machine learning and pattern recognition techniques to extract patterns implicit in a database.

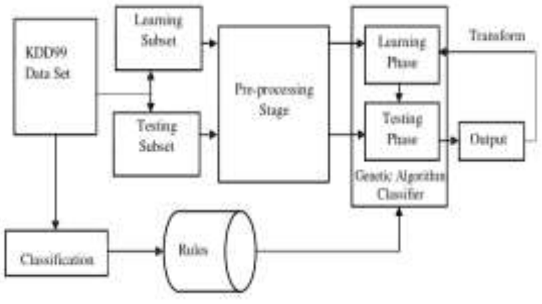


Figure 1. Proposed Model

PROPOSED WORK:

The problem of intrusion detection is not just identifying the attacks, but also to know the type of the connection. A genetic algorithm is a very good way to find an efficient solution to the problem. This process usually begins with the chromosomes which are randomly generated and represent all the possible solutions of the problem. Different positions are encoded as bits, numbers or characters from each chromosome. The goodness of each chromosome is calculated according to the desired solution using evaluation function, which is known as fitness function. Selection, crossover and mutation are the genetic operators that are sequentially applied to each individual with some certain probabilities [19]. Along with these IDS using GA approach a strong classification algorithm is proposed. This classification algorithm distinguishes important alerts from redundant one. It is a data mining technique used to map data instances in one of the various predefined categories. It can be used to detect individual attacks but it has a high rate of false alarm. It employs frequent item set mining for detecting patterns that describe frequently occurring redundant alerts. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data. The main aim of this algorithm is to reduce the false positive rate. The false positive rate is basically known as the percentage of a true thing which is wrongly recognized as an attack. This can be calculated as,

$$\text{False positive rate} = \frac{\#False\ positive}{\#True\ Negative + \#False\ positive} * 100$$

If the false positive rate is increased the alert system has to be called for the positive connections too. This algorithm will first get the event data of the host in each and every session and convert that information into session records, which describe strong associations between alert attribute values for each host ID. Such session records consist of host-bound audit sources such as operating system audit trails, system logs, or application logs, IP addresses, source, duration, timestamp and flags. This will help to understand the behavior of the host so that classification algorithm can alert the data mining algorithm to save the particular data, if the source behavior crosses a certain limit of the classification rule. A classification algorithm classifies the intrusion into some matrices that are, false positive, false negative, and true positive and true negative. Classification algorithm will form rules under these matrices, which is called as classification rules. This classification rule will apply to the data mining algorithm. So the whole mining system will be under a smart alert.

Rational Difference= (DVal– Dval_{max})/(Dval_{max}–Dval_{min}) (1)

Normalized Output (D_{norm}) = 2 * Rationaldifference (2)

Where,

DVal_{min}, DVal_{max} are the maximum and minimum value of the original inputs $\text{DVal}_{\text{norm}}$ is the normalized output.

The normalized output will fall in the range [-1, +1]. The training and learning dataset will have $(f, R_t, I_t)$and (f, R_l, I_l) respectively. Where f :features , R_{t} : Training Records , I_{t} : Training Intrusion type , R_{l} : Learning Records, I_{l} : Learning Intrusion type. Classification rules are prepared from the available data using the GA in an offline environment. In the real time environment, these rules are used to classify the inward network connections.

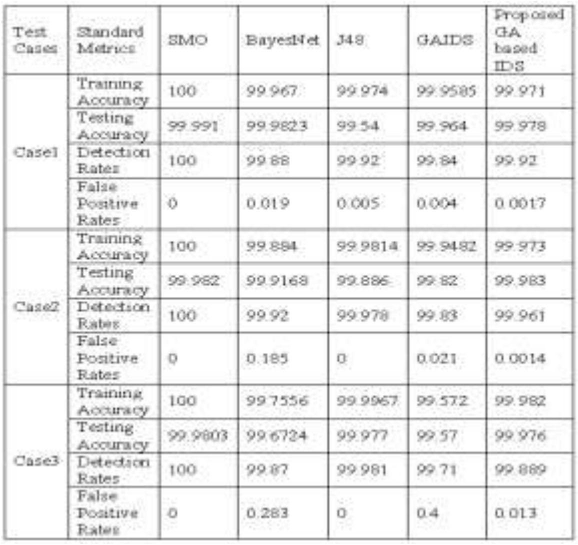
$$Fitness = \frac{Correctly\ detected\ attacks}{Total\ attacks\ in\ the\ training\ data\ set} - \frac{False\ attacks}{Total\ normal\ connections\ in\ the\ training\ dataset}$$

The range of the fitness values will be [-1,+1]. For a high fitness value (+1), high detection rate and low rate of false-positive is needed. Inversely, low detection rate and high rate of false-positive result in low fitness value (-1).The 41 features of the KDD cup data set are loaded in each network connection which will be either normal or attack.

IMPLEMENTATION AND EVALUATION:

To evaluate the effectiveness of our proposed system, KDD Cup data set is utilized for the implementation process. The preliminaries are 498,042 records among which 97,280 are normal connection records in training set. The test set contains 321,039 records among which 60,593 are normal connection records. Our implementation is done in three different experimental scenarios. In the first scenario 17 out of 41 features, second scenario will take 21 out of 41 features and 30 out of 41 in the last case. The experimental results are tabulated in Table1. The proposed GA based IDS, is compared with the other conventional techniques such as SMO, BayesNet, J48 and GAIDS.

Table 1.Evaluation Results



Test Cases	Standard Metrics	SMO	BayesNet	J48	GAIDS	Proposed GA based IDS
Case1	Training Accuracy	100	99.967	99.974	99.9585	99.971
	Testing Accuracy	99.991	99.9823	99.54	99.964	99.978
	Detection Rates	100	99.88	99.92	99.84	99.92
	False Positive Rates	0	0.019	0.005	0.004	0.0017
Case2	Training Accuracy	100	99.884	99.9814	99.9482	99.973
	Testing Accuracy	99.982	99.9168	99.886	99.82	99.983
	Detection Rates	100	99.92	99.978	99.83	99.961
	False Positive Rates	0	0.185	0	0.021	0.0014
Case3	Training Accuracy	100	99.7556	99.9967	99.572	99.982
	Testing Accuracy	99.9803	99.6724	99.977	99.57	99.976
	Detection Rates	100	99.87	99.981	99.71	99.889
	False Positive Rates	0	0.283	0	0.4	0.013

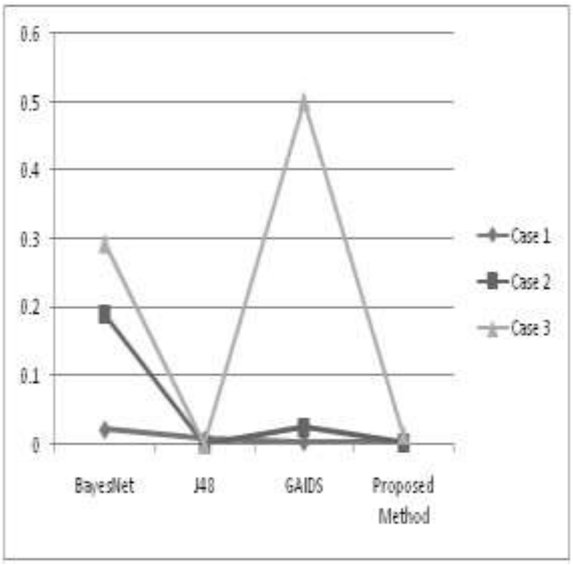


Fig 2. Comparison of False Positive Rate

J48 and our method performed really well in all the cases of our experiments and those methods gave a very small range of false positive rate, which is graphically represented in Fig 2.

CONCLUSION:

In this paper, we have proposed a novel technique of deploying genetic algorithm to detect the network intrusion. We have investigated new techniques for intrusion detection and evaluated their performance in terms of training accuracy, testing accuracy, detection rates and false positive rates. We have used the feature extraction and classification algorithm in the KDD99 Cup data set to evaluate the proposed GA based intrusion detection system. Three different testing cases i.e 17 out of 41 features, 21 out of 41 features and 30 out of 41 were simulated and our proposed system has higher training accuracy of 99.982% in the third case than the conventional GA based intrusion detection system. Our proposed system is capable of achieving lower false positive rate of 0.0017%, 0.014% and 0.13% in the three testing cases. From the empirical results, it is proved that the proposed GA based intrusion detection system has the best training and testing accuracy under 21 out of 41 features. However in future, our research work will be focused on developing more prominent intrusion detection system to achieve 100% accuracy.

REFERENCES:

[1] D. E. Denning, “An intrusion detection model,” IEEE Trans. Software Eng., vol. 13-2, p. 222, Feb. 1987.

[2] A. Patcha and J. M. Park, “An overview of anomaly detection techniques: existing solutions and latest technological trends,” Computer Networks, vol. 51, pp. 3448–3470, 2007.

[3] S. X. Wu and W. Banzhaf, “The use of computational intelligence in intrusion detection systems: a review,” Applied Soft Computing Journal, vol. 10, no. 1, pp. 1–35, 2010.

[4] V. Engen, Machine learning for network based intrusion detection: an investigation into discrepancies in findings with the KDD cup’99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data [Ph.D. thesis], Bournemouth University, 2010.

[5] M. Re and G. Valentini, “Integration of heterogeneous data sources for gene function prediction using decision templates and ensembles of learning machines,” Neurocomputing, vol. 73, no. 7-9, pp. 1533–1537, 2010.

[6] T. Dietterich and G. Bakiri, “Error-correcting output codes: a general method for improving multiclass inductive learning programs,” in Proceedings of the of Santa fe

Institute Studies in the Sciences of Complexity, pp. 395–395, Citeseer, 1994.

[7] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: a survey,” ACM Computing Surveys, vol. 41, no. 3, article 15, 2009.

[8] T. Dietterich, “Ensemble methods in machine learning,” in Proceedings of Workshop on Multiple Classifier Systems, pp. 1–15, 2000.

[9] G. Giacinto and F. Roli, “Approach to the automatic design of multiple classifier systems,” Pattern Recognition Letters, vol. 22, no. 1, pp. 25–33, 2001.

[10] Gomez.J, Dasgupta.D, Nasaroui.D and Gonzalez.F, “Complete expression Trees for evolving Fuzzy classifiers system with Genetic Algorithms and Applications to Network Intrusion Detection”, June 2002, PP.469-474.

[11] Melanie Middlemiss, Grant Dick, “Weighted Feature Extraction Using a Genetic Algorithm for Intrusion Detection”, 2003 Congress on Evolutionary Computation (cec-03) 2003, pp.1669-1675.

[12] T. Xia, G. Qu, S. Hariri, M. Yousif, “An Efficient Network Intrusion Detection Method Basedon Information Theory and Genetic Algorithm”, Proceedings of the 24th IEEE International Performance Computing and Communications Conference (IPCCC 05), Phoenix, AZ, USA.2005.

[13] R. Hu and R. I. Damper, “A 'No Panacea Theorem' for classifier combination,”Pattern Recognition, vol. 41, no. 8, pp. 2665–2673, 2008.

[14] Anup Goyal, Chetan Kumar, “GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System”, 2008.

[15] K. Ilgun, R. Kemmerer, P. A. Porras, “State Transition Analysis: A Rule-Based Intrusion Detection Approach”, IEEE Transaction on Software Engineering, 21(3):pp. 181-199. 1995