A Genetic Algorithm Approach to Network Intrusion Detection Systems

Srinivasa K G[1], Jagdeesh M N[1], Jayasimha R[1], Kiran Kumar N[1], Venugopal K R[2], L M Patnaik [2]
[1] Data mining Laboratory, MSRIT, Bangalore, India.
[2]

***Abstract*—We present an anomaly based network intrusion detection system applying a genetic algorithm approach. The method we choose exhibits a good detection rate with low false positives. The training time required is less compared to other NIDSes. The IDS designed is payload based and uses an adaptive genetic algorithm for both learning and detection. We have benchmarked our work with PAYL and POSEIDON using the 1999 DARPA dataset.**

I. Introduction

An intrusion detection system is used to detect many types of malicious network traffic and computer usage. This includes network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorized logins and access to sensitive files.

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

A. Intrusion Detection Systems

IDS come in a variety of flavors and approach the goal of detecting suspicious traffic in different ways. 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.

NIDS: Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally you would scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network.

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 The other types of intrusion detection systems are signature based and anomaly based. Signature based systems work with an intrusion database populated offline by knowing of the characteristics of the attack. Thus the IDS has to compare the input and classify it into normal and abnormal categories.

Anomaly based systems have only the normal behaviors in their profiles and any deviation above a threshold is signaled as an anomaly. Unlike signature based systems which give low detection rates and low false positive rates anomaly based system suffer from high false-positive rates; however they have a good detection rate.

Signature based systems cannot detect new attacks until they are known and added to the database. This results in lower detection rates. Signature based systems are preferable to detect attacks on the operating systems. However, anomaly based systems have the ability to detect zero-day worms. And hence are preferable for network related attacks.

Most of the systems used till now are predominantly signature based however a considerable amount of research is going on for reducing the false positive rates and increasing the detection rates in anomaly based systems.

An anomaly based system can classify the input based on either the header information or the payload. In this paper we describe a payload based system. Other payload based systems include PAYL [ ] and POSEIDON [ ].

B. Genetic Alogrithms

Genetic algorithms are a branch of evolutionary algorithms used in search and optimization techniques. The motivation behind the usage of genetic algorithms is to use evolutionary techniques in performing a more human like search.

The GA usually contains three primordial steps selection, crossover and mutation. These operations mimic the biological process to select the fittest chromosome. As the process continues over a period of time, our data centers (the chromosomes) will have only the fit individuals representing the solution to the problem set. Ascribing to the no free lunch theorem, there is no specific algorithm suited for all searches and problem sets. The genetic algorithms offer a unique set of advantages for search and optimization and choosing to opt for a GA solution should be motivated based on the problem set at hand.

1) Contribution: In this work we propose a genetic algorithm based approach to network intrusion detection system. The use of Genetic Algorithm is that it has a better classification than any other neural network architecture, takes less time for training and comes out with a better detection rate.

GANIDS is payload based in the sense that it uses only the destination address and the service port numbers for building profiles and all the header information is ignored. GANIDS

uses a single tier architecture where a GA is used for both classification and detection.

We have benchmarked our system w.r.t POSEIDON [ ] using the 1999 DARPA dataset [ ]. On this dataset our system shows a reasonable detection rate with low false positives and a faster running time than POSEIDON.

Genetic Algorithms belong to the evolutionary algorithms and is very efficient in machine learning and it has emerged to be a very effective tool in data mining applications. In an IDS the incoming packet needs to be classified into normal and abnormal categories. A GA functions best in this job since it can classify with a higher accuracy than any other methods such as Neural Networks etc.

By using a GA we obtain a better classification of the input data and hence will result in higher detection rates and lesser false positives which are major concerns for an Anomaly Based IDS. In addition to that genetic algorithms are relatively faster than neural networks and requires less time for training and hence the performance of the system increases considerably.

II. Related Work

Network intrusion detection systems have undergone rapid changes and are using new evolved techniques to get better results. So it is quite obvious that there are a lot of other methods to develop an IDS that are suitable to your needs.

In this part we go on to explain the techniques quite similar to the ideas presented here and then go on to describe the various other techniques in use.

Wei li, in his work describes an IDS that exploits both temporal and spatial information of network connections, in encoding the network information rules in to an IDS. He manages to highlight the various parameters required to implement the GA and also discusses the architecture and the various functions involved. His *evaluation* function calculates an *outcome* value for a chromosome as a sum of the product of the matched value and weights. He considers 57 genes in each chromosome, i.e., parameters of actual network connections. For every mismatch that occurs a penalty is calculated, which determines how easy it is to classify an intrusion. The fitness of the chromosome is computed using this penalty as *1-penalty*. He goes on to talk about the crossover and mutation techniques, including some niching techniques such as crowding and sharing. Using a data set such as the MIT Lincoln Library data set can be used to generate the rule set for the IDS. A very sound architecture is presented in this paper.

In the same vein, Zhang lian-hua et al. present an IDS using *rough set classification*, using fast hybrid genetic algorithms to form the reducts. Using rough set classification they manage to achieve an improvement over the SVM based systems. The main concept used is to reduce the number of attributes to consider for the rule generation. The rule generation is achieved here through the reducts. During testing on the KDD 99 data set they manage to achieve significant improvement in misclassification rate, though for the U2R&R2L attacks the SVM based system performs better.

Stein et al. have developed a IDS based on genetic algorithms and decision trees. The feature selection algorithm is based on a wrapper model. They use GA for searching and a decision tree for evaluation. The fitness of an individual is determined by the classification error rate. The next generation individuals are chosen according to the rank selection method. They have used a two point crossover mechanism and a bit level mutation strategy. Replacement is done by keeping two elite parents and replacing all else with the offspring. In their experiments they managed to achieve an improvement in the error rates as compared to only a decision tree classifier based model. This was mainly due to the error rate of the validation set being the fitness function. Hybrids generally tend to fare well in case of decision tree classifier based models.

Nguyen et al. in their paper describe a method to improve over SVM based IDS by employing a fusion of genetic algorithms and SVM.

III. Proposed Architecture

There are mainly two types of NIDSes that are widely prevalent i.e. packet oriented and connection oriented. While connection oriented systems are more suitable for offline analysis because of its large memory requirement, packet oriented systems are require less memory and can detect attacks while it is happening. GANIDS is packet oriented i.e. it analyses every packet as it arrives on the network interface. In the following section we give a high level description of the working of the system.

The architecture of GANIDS is as shown in the figure. The incoming traffic is first captured using a packet capture engine which is then used to extract the payload by removing all the header information present in the packet and the payload is given as input to the genetic algorithm which in the training phase uses it to build profiles.

A. Two Point Crossover

In our system we use a two point crossover scheme where the two parents crossover at two different points producing a total of eight off springs out of which two are replicas of the parents itself which are discarded. The remaining two are then tested for fitness. If they are fit enough then they are added to the population else they are not.

Selection of the parents for crossover is done by finding the fittest chromosome from the existing population and the input data forms the other parent. Since the input data is used to construct profiles the network behavior will be mapped on to the profiles efficiently.

B. Replacement Strategy

There are mainly two types of replacement techniques that are widely used viz. Complete Replacement and Partial Replacement. Complete replacement though easy to implement lose some of the fittest members in the population. However it is desirable for some of the chromosomes that are fit to survive in the population, hence we use a partial replacement

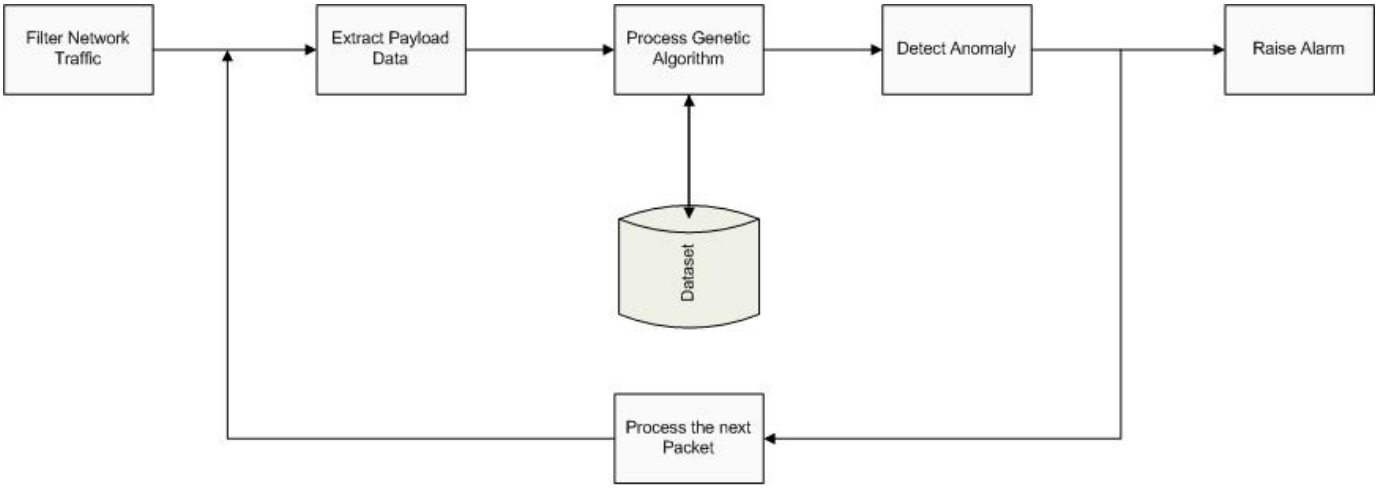

Fig. 1. GANIDS Architecure

technique where only some of the members are replaced and the rest are retrieved as it is.

In our system we use a steady state replacement technique which is a partial replacement technique where the off springs replace the parents in the population. Also in our system parents that are unable to produce an offspring that is fit enough to be added to the population will also be removed form the population.

C. Mutation

Mutation is very necessary in a genetic algorithm because it enables the algorithm to explore the search space more effectively and hence produces better results. In our system we perform mutations based on a mutation probability which varies dynamically during the course of execution.

The mutation probability pm is changed using the equations pmi =

IV. Algorithm

In this section we describe the algorithm that is used by GANIDS.

Problem Definition:

Let x_{i} be the input payload at time instance i, then the problem is to find the Chromosome c which yields the lowest value for the computation manhattan_distance $(c.weight[\,], xi)$

Pseudocode:

In this section we give a brief description of the pseudo code given above. The algorithm given below is used during the training phase i.e. the machine learning phase of the IDS.

If p_{m} is the mutation probability and G the number of generations and nc_i the number of crossover points be two.

Crossover:

Input:
$x_{\mathrm{i}} - payload\ at\ time\ i.$
$fittest - fittest\ chromosome$
Output:
$children\ created\ and\ added\ if\ fit.$

$begin$
$Children[6] = Cross(fittest, xi)$
$for\ all\ c\ \epsilon\ Children$
$find\ the\ fitness\ of\ c$
$if\ fitness(c) > threshold$
$add\_to\_population(c)$
$remove(fittest)$
end

Mutation:

Input:
$c - Chromosome$
Output:
$c1 - mutated\ chromosome$

$begin$
$r = random()$
$ifr > pmthen$
$mutate(c)$
end

Genetic Algorithm:: Input:
xi payload at time i

Output: fittest the fittest chromosome begin for i=0 to G do min_dist = INFINITY fittest = 0 for every c ϵ Chromosome do dist = manhattan_distance(c , xi) if dist $\leq$ min_dist then fittest = i min_dist = dist crossover(fittest , xi) for every c Chromosome mutation(c) end

Overview:

In the training phase, the input from the training data is used to build profiles. The machine learning phase functions as follows. First the input payload is used to find the fittest chromosome. Then the fittest chromosome and the payload itself are crossed to produce a total of eight offsprings out of which two of them are the replicas of the parents itself which are discarded.

The remaining six children are checked for fitness and checked against a threshold value. Only children which are fit enough are added to the population and others are discarded. Also if none of the six children are fit enough to be added to the population then even the parents are also removed from the population.

Then mutation is applied in order to explore the search space better. Mutation is done as follows, a random number r is generated for every chromosome in the population. If the value of r is greater than the mutation probability then some random numbers of weights are changed to some random values.

In the testing phase the fittest chromosome is found as in the algorithm but the crossover and mutation operations are not performed. Instead when the fittest chromosome is found, the minimum distance obtained is checked against a threshold and if it is higher than the threshold then it is flagged off as an anomaly.

V. Performance Analysis

Data still to be tested and analysed.

References

[1] G. O. Young, Synthetic structure of industrial plastics (Book style with paper title and editor), in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 1564.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123135.

[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, An approach to graphs of linear forms (Unpublished work style), unpublished.

[5] E. H. Miller, A note on reflector arrays (Periodical styleAccepted for publication), IEEE Trans. Antennas Propagat., to be published.

[6] J. Wang, Fundamentals of erbium-doped fiber amplifiers arrays (Periodical styleSubmitted for publication), IEEE J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces (Translation Journals style), IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, Infrared navigationPart I: An assessment of feasibility (Periodical style), IEEE Trans. Electron Devices, vol. ED-11, pp. 3439, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Networks, vol. 4, pp. 570578, Jul. 1993.

[12] R. W. Lucky, Automatic equalization for digital communication, Bell Syst. Tech. J., vol. 44, no. 4, pp. 547588, Apr. 1965.

[13] S. P. Bingulac, On the compatibility of adaptive controllers (Published Conference Proceedings style), in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 816.

[14] G. R. Faulhaber, Design of service systems with priority reservation, in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 38.

[15] W. D. Doyle, Magnetization reversal in films with biaxial anisotropy, in 1987 Proc. INTERMAG Conf., pp. 2.2-12.2-6.

[16] G. W. Juette and L. E. Zeffanella, Radio noise currents n short sections on bundle conductors (Presented Conference Paper style), presented at the IEEE Summer power Meeting, Dallas, TX, Jun. 2227, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, An analysis of surface-detected EMG as an amplitude-modulated noise, presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, Narrow-band analyzer (Thesis or Dissertation style), Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, Parametric study of thermal and chemical nonequilibrium nozzle flow, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, Nonlinear resonant circuit devices (Patent style), U.S. Patent 3 624 12, July 16, 1990.

[21] IEEE Criteria for Class IE Electric Systems (Standards style), IEEE Standard 308, 1969.

[22] Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

[23] R. E. Haskell and C. T. Case, Transient signal propagation in lossless isotropic plasmas (Report style), USAF Cambridge Res. Lab., Cambridge, MA Rep. ARCRL-66-234 (II), 1994, vol. 2.

[24] E. E. Reber, R. L. Michell, and C. J. Carter, Oxygen absorption in the Earths atmosphere, Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (420-46)-3, Nov. 1988.

[25] (Handbook style) Transmission Systems for Communications, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 4460.

[26] Motorola Semiconductor Data Manual, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.

[27] (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume (issue). Available: http://www.(URL)

[28] J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

[29] (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

[30] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. IEEE Trans. Plasma Sci. [Online]. 21(3). pp. 876880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar