Adaptive Genetic Algorithm Model for Intrusion Detection

K. S. Anil Kumar
Assistant Professor
Department of Computer Science
Sree Ayyappa College
Eramallikkara, Chengannur, Kerala (India)
ksanilksitm@gmail.com

Dr. V. Nanda Mohan
Professor Emeritus in Knowledge Management
Department of Futures Studies
University of Kerala, Thirvananthapuram-34(India)
nandamohanv@gmail.com

Abstract
Intrusion detection systems are intelligent systems designed to identify and prevent the misuse of computer networks and systems. Various approaches to Intrusion Detection are currently being used, but they are relatively ineffective. Thus the emerging network security systems need be part of the life system and this is possible only by embedding knowledge into the network. The Adaptive Genetic Algorithm Model - IDS comprising of K-Means clustering Algorithm, Genetic Algorithm and Neural Network techniques. The technique is tested using multitude of background knowledge sets in DARPA network traffic datasets.
Keywords: Intrusion Detection System (IDS), K-Means Clustering, Neuro - Genetic, DARPA dataset

1. Introduction

The proliferation of computer networks in the contemporary period, in particular the contribution of e-commerce to the world economy, has necessitated the security of computer networks an international priority. Intrusion detection has become an inevitable area of research, since it is technically infeasible to build a system that can offer total resistance to attacks. Intrusion as generally described is an act of trespassing or infringing the integrity, confidentiality or preventing the availability of a resource [1]. Intrusion Detection Systems detects unauthorized or malicious attacks over a computer system that occurs primarily through network. These attacks can compromise the security and trust of a system.

Researchers have developed Intrusion Detection Systems (IDS) capable of detecting attacks in several available environments. Categorized broadly based on their patterns of detection, IDSs can be classified as misuse detectors or anomaly detectors. Misuse detectors rely on comprehending the patterns of known attacks [2, 3], while anomaly detection exploits user profiles as the basis of detection, and brands the characteristics of the deviant from the normal ones as intrusion [2, 3, 4, 5].

We have used K-Means to cluster normal and intrusion packets, Genetic Rule Algorithm for rule generation from the perceived traits of normal and abnormal clusters, and Neural Networks for detecting abnormal packets similar to the ones given during training sessions and for an artificial intelligence that detects anomalies not presented during training [6]. The study has been based on the selected fields of DARPA dataset.

2. Rational of the Study

This section gives a brief introduction of the techniques used in the proposed work.

2.1 Architecture of the GAM

The Genetic Algorithm Model of intrusion detection system is developed with a combination of K- Means clustering and Neuro - Genetic Algorithm. K-Means clustering has been used for creating intrusion and non-intrusion clusters. Then Genetic Algorithm based rule got generated from the perceived traits of normal and abnormal clusters, and Neural Networks for detecting (by learning) abnormal packets similar to the ones given during learning sessions and for an artificial intelligence that detects anomalies not presented during learning. The schematic diagram of architecture of GAM IDS is depicted in Fig.2.1

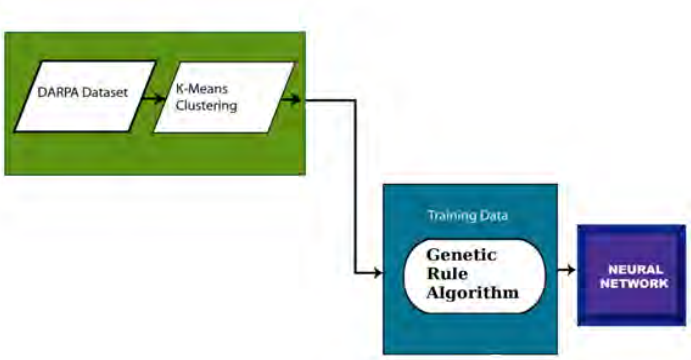

Figure 2.1: Architecture of GAM IDS System

2.2 K- Means Clustering Algorithm

K-means clustering is a technique that classifies objects in to K number of groups based on their attributes or features. Obviously, K is a positive number. The cluster centroid is calculated first. Then the grouping is done by minimizing the sum of squares of distances between the data and the corresponding cluster centroid. The object of K means clustering is to classify the data by analyzing the traits and then organizing them in accordance to their attributes. The reason why K-Means clustering has been chosen for the algorithm is that the packets we analyze needs to be categorized in to just two clusters, normal and intrusion. Hence, the value of K is defined as '2' [7].

Since the real-world databases normally used are highly susceptible to noisy, missing and inconsistent, it can be initially processed so as to improve the quality and ease of the mining process. Low quality data will lead to low quality mining results. Clustering can reduce the data size and thus to improve the quality of the mining process.

2.2 Genetic Rule Algorithm

Genetic algorithm is a family of computational models based on principles of evolution and natural selection. These algorithms convert the problem in a specific domain into a model by using a chromosome-like data structure and evolve the chromosomes using selection, crossover, and mutation operators. The range of the applications that can make use of genetic algorithm is quite broad [7], [8]. In computer security applications, it is mainly used for finding optimal solutions to a specific problem. Denning [9] and Javitz et al. [10] have used Genetic algorithm in IDS.

The process of a genetic algorithm usually begins with a randomly selected population of chromosomes. These chromosomes are representations of the problem to be solved. An *evaluation function* is used to calculate the "goodness" of each chromosome. During evaluation, two basic operators, *crossover* and *mutation*, are used to simulate the natural reproduction and mutation of species. The selection of chromosomes for survival and combination is biased towards the fittest chromosomes [11].

Genetic Rule Algorithms can be used to evolve simple rules for network traffic [7], [12]. These rules are used to differentiate normal network connections from anomalous connections. These anomalous connections refer to events with probability of intrusions. The Genetic Rule Algorithm is used for creating the training dataset in GAM. In this algorithm, a population is created with a group of individuals created randomly and the individuals in the population are then evaluated. Using an evaluation function, a score based individuals are created based on how well they perform at the given task. For effective and efficient learning in a real time environment, dual learning of the characteristic of the abnormal verses normal is also explored in the system.

2.3 Artificial Neural Network

The Neural networks can learn, differentiate, and extract the underlying correlations and sort out the patterns of the data fed. Generally executed in conventional computers, neural networks typically are software simulations. They focus on structuring the connections between the processing elements termed as neurons by the terminology rather than manipulating zeroes and ones for computation as performed in the conventional digital modal. These structures and weights determine the output. Hence we factored in this aspect of the neural network mechanism; and we assigned weights to the genetic rules generated. These weights denote the degree to which the presence of a particular attribute has influenced the conclusion of abnormality of the intrusion packet.

Neural networks designate these neurons to manipulate the inputs mathematically and generate the output. Here is where we perform the analysis of the genetic rules under scrutiny, with reference to their calibrated weights.

This analysis is performed in every neuron in the network until the results are obtained. The system can now be expected to initiate its own activity in response to external stimuli not exactly matching the ones presented during the training sessions [13]. Neural Networks can deploy a myriad of approaches for making determinations that include genetic algorithm, fuzzy logic, gradient based training, and Bayesian methods [14]. After a comprehensive analysis, we opt to make use of genetic algorithm for rule generation. Based on the complexity of the requirement, layers (called occasionally as knowledge layers) in neural networks can be organized in varying quantities and the system can be graded accordingly [15]. The solution uses the built in facility of Mat lab to feed forward the learned relationship to higher knowledge layers.

3. Methodology

The proposed solution attempts to exploit the potential of K-means Clustering, Genetic Rule Algorithm and Neural Network techniques for Intrusion Detection. Efficiency of the proposed solution is analysed by sampling DARPA datasets. The fields in the datasets are scrutinized and categorized as intrusion and normal packets. This classification is done by applying K-means clustering technique.

The Genetic rule is used to generate the training set for the neural network. A rule consists of two parts: the antecedent and the consequent. In general, the antecedent part is a conjunction of a number of attribute values. Each categorical attribute in the conjunction is represented by 10 bits, where 10 is the number of possible logical operators between the attributes. If a particular value is present then the corresponding bit is set to 1; otherwise, it is 0. Multiple bits can be set to 1 representing a logical OR among the corresponding attribute values. Multiple attributes in the antecedent are chained by the logical AND operator. A representation of a solution might be an array of bits, where each bit represents a different logical operation, and the value of the bit (0 or 1) is substituted to logical operator 'AND' or 'OR'. The sample rules generated using Genetic rule algorithm is illustrated in table 3.1

Sample query with rules are given below
{'select count(*) from normal where protocoltype = "icmp" or land= 0 and wrong_fragment= 52 or synflood= 69 or num_comp= 0 or same_srv_rate= 0 or diff_srv_rate= 8 and count= 699 or srv_count= 160 and dst_host_count= 218 and dst_host_srv_count= 290';}

Table-3.1: Sample query with rules

The consequent part consists of the attribute of DARPA dataset and its value only. If the output (bit pattern) of genetic algorithm has been transformed to logical operators, the rules can be created using the logical operators (AND/OR). The generated rule is used in SQL query for creating the training data set to the neural network. The weights are calculated based on records retrieved as the result of SQL query and it is stored as the 12^{th} attribute of the record as shown in Table 3.2. If no record is generated from the result of SQL query, create another rule for generating the records for learning the system. For better results training set is created from both Intrusion data as well as Non-Intrusion data.

Protocol Type	Land	Wrong Fragment	Syn flood	Num Comp	Same SRV rate	Diff SRV rate	Count	SRV Count	Dst host Count	Dst host SRV count	Weights derived
1	0	21	81	0	0	0	793	968	824	342	2.065031
1	0	39	4	0	0	4	626	969	1000	345	841.5
2	0	72	93	0	0	6	920	37	498	313	2.0625
1	0	55	21	0	0	7	105	11	483	435	76.5
2	0	73	23	0	0	1	242	51	521	117	25.1194
1	0	22	61	0	0	3	407	972	788	513	1.835333
1	0	44	89	0	0	7	661	610	28	992	420.75

Table- 3.2: Weights generated using genetic rule algorithm

The flow-chart, which reveals the sequence of operations for generating training dataset using genetic rule algorithm is shown in the figure 3.1

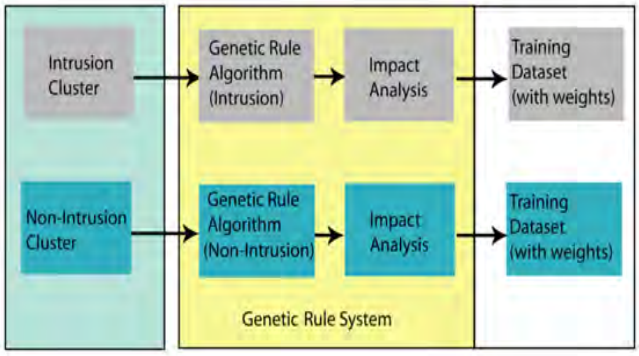

Figure 3.1: The training dataset generated using genetic rule system

The learned neural network is used for interpreting the data for anomaly. When there is an anomaly, then the neural network classify it as intrusion otherwise classify it as normal. The RMSE test has been used for validation of the output. For validating the model, the goodness of fit test using Chi-square has been used. As the RMSE follows the Chi-square distribution, the use of Chi-square test has been justified. The RMSE value estimated (0.001) being less than the table value of Chi-square, the model got validated.

5 Result and Discussion

The optimization algorithmic model, GAM is used for looking into the optimum characteristics of the other Models. The table 5.1 demonstrated the average parametric figures of the experimental results.

Analysis	Percent of Result
True Positive Rate	99.32390
True Negative Rate	99.36260
False Positive Rate	0.63741
False Negative Rate	0.67609
Probability of Detection	99.9990
Overall accuracy	99.350

Table 5.1: GAM result Analysis with 10000 records

The Genetic Algorithm model provides intrusion detection inference system that can be adapted on a real time basis. The algorithm is designed to perform dual learning (Intrusion and Non-Intrusion) to identify the characteristics of intrusion verses normal is to overwhelm the possibility of indecisiveness and the subsequent wrong prediction due to nuances that exist in their patterns.

The experiments showed a significant reduction in false alarm rate of 0.68 percent and achieved a very high average accuracy of 99.35 percent in GAM. McHugh and John [16] reported that the difference of accuracy in learning and test data is almost less than 1 % in case of detection of attack, however in case of normal the difference is around 5 % which suggest that detection of new/novel attack by IDS needed to be improved further. Here in this study the difference of accuracy in learning and testing in GAM is less than 1% and such IDS inference systems are well suited to detect new/novel attacks. The GAM also offers a significant reduction in false alarm rates and eliminates the need for interference by a human analyst.

5. Conclusion

The solution crafted with an object to create a powerful intrusion detection approach proved worthwhile. Convergence of K-Means, Genetic Rule Algorithm and Neural Network has helped to achieve a robust architecture. The inherent deficiencies perceived in these individual techniques towards attaining an effective intrusion detection algorithm were rectified by blending them appropriately. Comprehensive analysis of the

characteristics of the abnormal and even the normal packets helped recognition of their patterns and discrimination efficiently. The adaptability of the model is tested and concludes that improvements in detection rate, false positive rate and accuracy were achieved. Thus it is demonstrated that the system is adaptive since it exhibited improvements after being trained with new data.

6. Acknowledgements

The University Grants Commission, New Delhi, has granted financial assistance to this research project. We are also grateful to the Principal, Sree Ayyappa College for permitting to do this research project.

7. References

[1] Heady R., Luger G., Maccabe A., and Servilla M. 1990. The architecture of a Network level intrusion detection system, Technical Report, CS90-20, University of New Mexico, Albuquerque, NM 87131.
[2] Denning D. (1987) “An Intrusion-Detection Model,” IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, pp.222-232
[3] Kumar S., Spafford E. H. (1994) “An Application of Pattern Matching in Intrusion Detection,” Technical Report CSD-TR-94-013. Purdue University.
[4] Ryan J., Lin M-J., Miikkulainen R. (1998) “Intrusion Detection with Neural Networks,” Advances in Neural Information Processing Systems, Vol. 10, Cambridge, MA: MIT Press.
[5] Terran lane, Carla E. Brodley, Temporal Sequence Learning and Data Reduction for anomaly Detection, Vol. 2, No. 3, August 1999, pp. 295- 331.
[6] Masayuki Murakami, Nakaji Honda. A study on the modeling ability of the IDS method: A soft computing technique using pattern-based information processing, International Journal of Approximate Reasoning, Volume 45 , Issue 3 August 2007, Pages 470-487.
[7] Marimuthu, A. Shanmugam.A, Intelligent progression for anomaly intrusion detection, 6th International Symposium, PP: 261-265, Jan 2008.
[8] Ren Hui Gong, Mohammad Zulkernine, purang Abolmaesumi, “A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection” , Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 246-253, 2005.
[9] Denning D E, “An Intrusion Detection model”, IEEE transactions on Software Engineering”, Vol. SE-13, No-2, 1987, PP 222-232.
[10] H.S. Javitz, A. Valdes, “The NIDES Statistical Component Description and Justification,” Technical report, SRI International, Menlo Park, CA, March 1994.
[11] Marin. J, Ragsdale.D, Sirdu. J, “A hybrid approach to the profile creation and intrusion detection”, DARPA Information Survivability Conference & Exposition II, Vol 1, pp: 69-76, 2001.
[12] Susan M. Bridges, Rayford B. Vaughn, “Fuzzy data mining and genetic algorithms applied to intrusion detection”, National Information Systems Security Conference (NISSC), pp: 131-134, 2000.
[13] Ganesh Kumar, P. Devaraj, D. Network Intrusion Detection using Hybrid Neural Networks, Signal Processing, Communications and Networking, 2007. ICSCN '07. International Conference. pp: 563-569, ISBN: 1-4244-0997-7. Feb. 2007
[14] K.S. Anil Kumar and Dr. V. NandaMohan," Novel Anomaly Intrusion Detection Using Neuro-Fuzzy Inference System", International Journal of Computer Science and Network Security, vol.8, no.8, pp.6-11 , August 2008.
[15] S. Selvakani Kandeeban and Rengan S Rajesh, “Integrated Intrusion Detection System Using Soft Computing”, International Journal of Network Security, Vol.10, No.2, pp. 87-92, Mar. 2010
[16] McHugh,John “Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA IDS evaluations as performed by Lincoln Laboratory”, ACM Transactions on Information and System Security, vol.3, No.4, Nov. 2000.