International Journal of Computer Networks and Communications Security

VOL. 3, NO. 1, JANUARY 2015, 6–10
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

An Optimized Genetic Algorithm with Classification Approach used for Intrusion Detection

Aliakbar Tajari Siahmarzkooh[1], Saied Tabarsa[2], Ziba Hosseini Nasab[3] and Fareevar Sedighi[4]

[1] Ph.D Student, Department of Computer Sciences, University of Tabriz, Tabriz, Iran

[2, 3, 4] MSc Student, Department of Computer Engineering, University of Mirdamad, Gorgan, Iran

E-mail: [1]tajari1987@gmail.com, [2]tabarsa.saeid@gmail.com, [3]zibahh109@yahoo.com, [4]fareevar.seddighi@gmail.com

ABSTRACT

IDSs which are increasingly a key part of system defense are used to identify abnormal activities in a computer system. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected. To reduce this dependence, various data-mining and machine learning techniques have been used in the literature. During recent years, number of attacks on networks has dramatically increased and consequently interest in network intrusion detection has increased among the researchers. In this paper we have used the terms detection rates and false alarm rates to compare the results of Naïve Bayes algorithm and Support Vector Machine algorithm to find out the results for intrusion detections and by using Naïve Bayes method with Genetic Algorithm technique try to improve the rates for better detection. The proposed algorithm is used for comparative study we have done in this paper on the basis of which we measure the performance and usefulness of particular methods in detecting specific class of attacks. Experimental results performed using the KDD99 benchmark network intrusion detection dataset indicate that it can significantly reduce the number and percentage of false positives and scale up the balance detection rates for different types of network intrusions.

Keywords: *Intrusion Detection System, Naïve Bayes, Support Vector Machine, Genetic Algorithm, KDDCup99 dataset,False alarm Rates.*

1 INTRODUCTION

The Internet is an essential part of our daily activities, and the number of Internet users continues to grow every day to about 1.8 billion in 2010 according to world Internet usage statistics [1]. As the number of Internet users' increases, demand on its services will increase, and more critical data are exchanged over the wires. Therefore, reliable, accurate, and configurable security systems are crucial to provide secure communication links and protect data shared over the Internet. Despite the wide deployment of firewalls, antivirus, and intrusion detection systems (IDSs), cybercriminal activity is on the rise. In 2008, Symantec Corporation reported more than 75,000 active bot-infected computers each day, a 31% increase from 2007 [2]. Finally, according to the World Economic Forum, $1 trillion is the annual estimate for online crime cost [3].

Intrusion detection systems are an essential component of computer security to detect attacks at early stage. They aim to detect intrusions in real time, by monitoring and analyzing the network traffic and looking for attack signatures or deviation from normal behavior. But they fail in processing the enormous data and are unable to provide adequate accuracy and sensitivity. Many algorithms were proposed to mitigate the challenges of increasing traffic, large signature databases, huge behavior profiles, and the difficulty to recognize boundaries between normal and suspicious behaviors. Classification of network traffic to distinguish normal and abnormal behavior is considered one of the main issues in IDS because of the large number of features. In addition, the complex relationships between features are not easy to spot. Therefore, choosing the differentiating features and developing the best machine learning algorithm in terms of high detection rates as well as fast training and testing processes are the main focus of this work.

Anomaly-Based Detection Model, it is used to identify an intrusion when the observed activities in computer systems demonstrate a large deviation from the norm profile built on long-term normal activities and it is able to detect even unknown attacks by comparing the current abnormal events with something that is considered normal.

In such cases, the anomalies may be showing false positives means classifying a normal behavior as an abnormal, and hence as possible attack instances. This discussion points out that the tradeoff between the ability to detect new attacks and the ability to generate a low false alarms rate is the key point to develop an effective IDS. In this, if fuzzy linguistic IF-THEN rules[3] are formulated and a process of fuzzification, inference, and defuzzification leads to the final decision of the system. Although sometimes the fuzzy rules can be directly derived from expert knowledge, different efforts have been made to obtain an improvement on system performance by incorporating learning mechanisms guided by numerical information to define the fuzzy rules. This issue, known as fuzzy rule learning(FRL) which is used to generate results obtained by applying on classification algorithms such as Naïve Bayes & Support Vector Machine in the KDD data set. Several data mining algorithms, such as decision tree, naïve Bayesian classifier, neural network, Support Vector Machines, and fuzzy classification, etc. [4] have been widely used by the IDS community for detecting known and unknown intrusions, from which we used Naïve Bayes and SVM in our work. Data mining based intrusion detection algorithms aim to solve the problems of analyzing the huge volumes of audit data and realizing performance optimization of detection rules [5].

The main emphasis of this paper is to find efficiently the values of detection rates (DR) and false alarms rate (FAR) of the various intrusion attacks detected in the data set and compare them to measure their performances. The experiments and evaluation are performed using the KDD-Cup99 benchmark dataset which contains information on computer networks, during normal and intrusive behaviors. This dataset is available at the University of California, Irvine web site.

2 RELATED WORKS

Many schemes for IDS were proposed to mitigate the issues it suffers from such as accuracy, large datasets unbalanced distribution of data, difficulty to recognize boundaries between normal and abnormal behaviors, and adapting to constantly changing environment [6]. In the succeeding paragraphs, we summarize the major efforts and discuss their advantages and shortcomings.

Jiang et al. [7] proposed two serial and parallel hierarchical neural networks for IDS, based on radial basis function (RBF). This approach integrates both misuse and anomaly-based detection, and takes advantage of the RBF short training and high accuracy. First, the RBF anomaly classifier identifies data as normal or abnormal. When enough data are collected, a C-mean clustering algorithm is used to group intrusions into different categories. Therefore, IDS will automatically use these groups to train a new RBF classifier to detect emergent intrusions. Consequently, it is able to analyze network traffic in real time and train the new classifier automatically.

The approach by Tian et al. [8] combines decision trees and fuzzy logic for misuse detection. They divided the large dataset into subsets, each having its own subdecision tree. All sub-decision tree results are combined using fuzzy integral. This method is suitable for large datasets and outperforms one large decision tree for large datasets. Dhanalakshmi et al. [9] integrated fuzzy logic with data mining and used genetic algorithm (GA) in order to mine fuzzy association rules by extracting the best rules. This method involves two operations: rule generation and detection. The first generates rules from network traffic by using fuzzy data mining and GA. The second operation uses the generated rule subset for intrusion detection.

Kruegel et al. [10] proposed a multisensory fusion approach using Bayesian classifier for classification and suppression of false alarms that the outputs of different IDS sensors were aggregated to produce single alarm. In 2000, Dickerson at al. [11] developed the Fuzzy Intrusion Recognition Engine (FIRE) using fuzzy logic that process the network data and generate fuzzy sets for every observed feature and then the fuzzy sets are used to detect network attacks.

In 2004, Amor et al. [12] conducted an experimental study of the performance comparison between NB classifier and DT on KDD99 dataset. In 2010, Md. Abadeh and J. Habibi, [13] proposed a hybridization of evolutionary fuzzy systems and optimized Genetic Algorithm which is used for Intrusion Detection.

3 PROPOSED METHOD

There has been a flux of work using evolutionary algorithms and swarm intelligence for IDSs because of their distributed nature and inherent parallelism. Those qualities make swarm intelligence a perfect candidate for IDS, which requires vetting through huge amounts of data to distinguish normal and abnormal behaviors. Given the work surveyed in Section 2, there is still room for improvement.

The KDD cup 1999 dataset was used in the 3rd International Knowledge Discovery and Data Mining Tools Competition for building a network intrusion detector, a predictive model capable of distinguishing between intrusions and normal connections [14]. In 1998, DARPA intrusion detection evaluation program, a simulated environment was set up to acquire raw TCP/IP dump data for a local-area network (LAN) by the MIT Lincoln Lab to compare the performance of various intrusion detection methods. It was operated like a real environment, but being blasted with multiple intrusion attacks and received much attention in the research community of adaptive intrusion detection. The KDD99 dataset contest uses a version of DARPA98 dataset. In KDD99 dataset [15], each example represents attribute values of a class in the network data flow, and each class is labeled either normal or attack. Examples in KDD99 dataset are represented with a 41 attributes and also labeled as belonging to one of five classes as follow:

1) Normal, (2) Denial of Service (DOS), (3) Remote to User (R2L), (4) User to Root (U2R), (5) Probing (Probes).

Our aim is to increase level of performance of intrusion detection of the most using classification techniques nowadays by using optimization method like ACO. In this fuzzy IF-THEN rule is used to increase the interpretability and accuracy of intrusion detection model for better results. We choose naive bayes classifier (NB) and Support vector machine (SVM) in our work. Here in this paper we can measure the performance based on train values of these methods by comparing the generated values of DR and FAR respectively. In Naïve bayes classifier using optimizrd Genetic Algorithm technique the values of DR and FAR can be improve for detecting intrusions in the given dataset.

The proposed algorithm can be summarized as follows:

1) Start with to generate input dataset from KDDCup99 to compare the performance of various intrusion detection methods.
2) Apply fuzzy IF-THE rule to increase the interpretability and accuracy of intrusion detection model.
3) Select the training dataset by using train values from the given input dataset to train our model.
4) Choose one of the following classification methods which is used to find out the various intrusion to be detect in test dataset.
    a) Support Vector Machines (SVM)
    b) Naïve Bayes Classifier (NB)
    c) Naïve Bayes with optimizrd Genetic Algorithm
5) Apply it on the test dataset.
6) Generate detection rate (DR) and false alarm rate (FAR) values.
7) Compare these values obtained in previous step and finally obtain the results by measure of performance of the model.
8) Stop.

As we know for better intrusion detection the rate of detection must be higher and rate of false alarm must be lower, in our next section the experimental analysis is shown. To estimate the performance of the system, two important formulas are used to evaluate system accuracy: 1) detection rate (DR) Eq.(1), 2) false alarm rate (FAR) Eq.(2).

$$\text{(1)} \quad \text{DR}=\frac{(\textit{Total number of detected attacks})}{(\textit{Total number of attacks detection})}\times 100\%$$

$$\text{(2)} \quad \text{FAR}=\frac{(\textit{Total number of normal processes})}{(\textit{Total number of misclassified processes})}\times 100\%$$

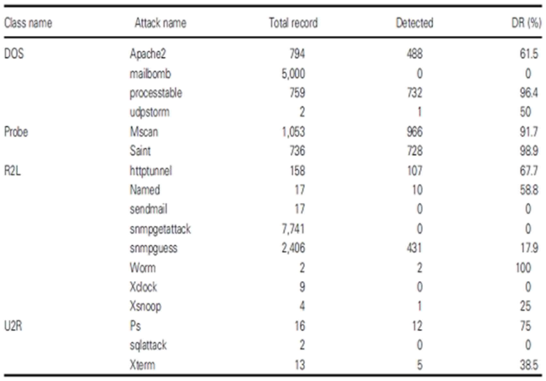
In order to evaluate the performance of proposed algorithm for network intrusion detection, we performed 5-class classification using KDD99 intrusion detection benchmark dataset. All experiments were performed using an Intel core 2 Duo Processor 2.0 GHz processor (2 MB Cache, 800 MHz FSB) with 512 MB RAM, and implemented on a Windows XP Professional operating system.

4 SIMULATION RESULTS

Tables I and II show the experimental results for known and unknown attacks, respectively. we detects all known and unknown attacks such as: “mailbomb,” “sendmail,” “snmpgetattack,” “xclock,” and “sqlattack.” Most of undetected attacks belong to R2L and U2R classes. Attacks such as “mailbomb” and “sendmail” are spam applications that operate above the network layer and are very hard to detect with connection features. They require application layer features and analysis to accurately distinguish them from other traffic.

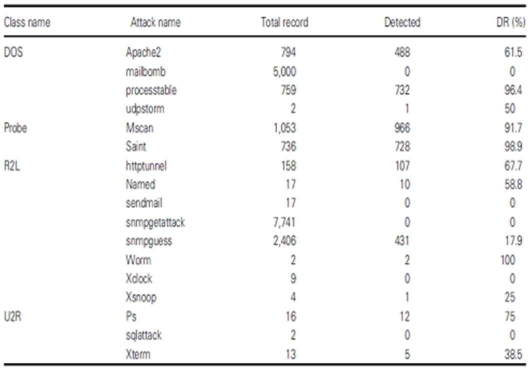
Table III shows the detection rate for known and unknown attack classes. It is observed that our approach achieves excellent detection rates of 99.8% and 92.1% for DOS and probe known attack classes, respectively. Those attacks form the majority of records in the testing dataset. However, the it achieves low detection rate of 35.9% and 2.5% for U2R and R2L known attacks, respectively. That is because those attack classes are underrepresented in the testing dataset.

Table 1: Experimental Results for known attacks



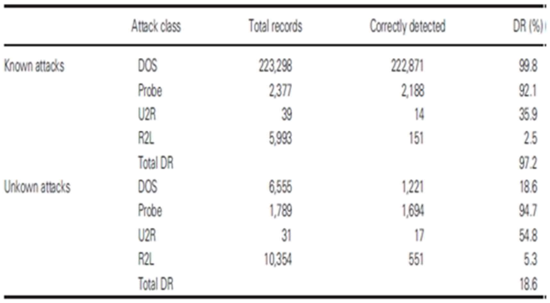
Class name	Attack name	Total record	Detected	DR (%)
DOS	Apache2	794	488	61.5
	mailbomb	5,000	0	0
	processtable	759	732	96.4
	udpstorm	2	1	50
Probe	Mscan	1,053	966	91.7
	Saint	736	728	98.9
R2L	httptunnel	158	107	67.7
	Named	17	10	58.8
	sendmail	17	0	0
	snmpgetattack	7,741	0	0
	snmpguess	2,406	431	17.9
	Worm	2	2	100
	Xclock	9	0	0
	Xsnoop	4	1	25
U2R	Ps	16	12	75
	sqlattack	2	0	0
	Xterm	13	5	38.5

Table 2: Experimental Results for unknown attacks



Class name	Attack name	Total record	Detected	DR (%)
DOS	Apache2	794	488	61.5
	mailbomb	5,000	0	0
	processtable	759	732	96.4
	udpstorm	2	1	50
Probe	Mscan	1,053	966	91.7
	Saint	736	728	98.9
R2L	httptunnel	158	107	67.7
	Named	17	10	58.8
	sendmail	17	0	0
	snmpgetattack	7,741	0	0
	snmpguess	2,406	431	17.9
	Worm	2	2	100
	Xclock	9	0	0
	Xsnoop	4	1	25
U2R	Ps	16	12	75
	sqlattack	2	0	0
	Xterm	13	5	38.5

Table 3: Detection rates for known and unknown arrack classes



	Attack class	Total records	Correctly detected	DR (%)
Known attacks	DOS	223,298	222,871	99.8
	Probe	2,377	2,188	92.1
	U2R	39	14	35.9
	R2L	5,993	151	2.5
	Total DR			97.2
Unkown attacks	DOS	6,555	1,221	18.6
	Probe	1,789	1,694	94.7
	U2R	31	17	54.8
	R2L	10,354	551	5.3
	Total DR			18.6

5 CONCLUSION

In this paper, we compare the performances of different methods used for Intrusion detection on the basis of DR and FAR values. The results obtained as, the value of rates obtained from Naïve Bayes is quite lesser than SVM. And increased efficiency of Naïve Bayes classifier with ACO is increased upto (97% approx.) much better than the two methods which is upto 91% approx.

One more point to be noticed is that with the increase in the train values, there shows slight increase in the DR values in each method which indicates better performance of our work. We conclude that with such a great improvement in percentages of DR and FAR values, our work seems will be useful for more research in this field. Here shows that using optimization technique the results be improve for better detection.

6 REFERENCES

[1] World Internet Usage Statistics News andWorld Population Stats [Online]. http://www.internetworldstats.com/ stats.htm, 2010.

[2] Symantec Corp Threat Report [Online]. http://www. symantec.com/about/news/release/article.jsp?prid=20090413_01, 2010.

[3] Baker M, DeWalt D, Ilube T, Kudelski A, Zittrain J. “Is the Internet at Risk?, World Economic Forum” [Online]. http://www.weforum.org/en/knowledge/KN_SESS_ SUMM_27219?url=/en/knowledge/KN_SESS_ SUMM _27219, 2010.

[4] Han J and Kamber M, "Data Mining:Concepts and Techniques Slides for Textbook — Chapter 7", Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada , 2010.

[5] Su-Yun W and Yen E, “Data mining-based intrusion detectors,” Expert Systems withApplications, Vol. 36, Issue 3, Part 1, April ,pp. 5605-5612, 2009.

[6] Xiaonan Wu S, Banzhaf W. “The use of computational intelligence in intrusion detection systems”. Applied Soft Computing, 10:1–35, 2010.

[7] Jiang J, Zhang C, Kame M. RBF-based real-time hierarchical intrusion detection systems. In Proceedings of the International Joint Conference on Neural Networks (IJCNN’03), vol. 2, pp. 1512–1516, 2003.

[8] Tian J, Fu Y, Xu Y, ling Wang J. Intrusion detection combining multiple decision trees by fuzzy logic. In Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT’05), IEEE Press, pp. 256–258. 2005.

[9] Dhanalakshmi Y, Ramesh Babu I. Intrusion detection using data mining along fuzzy logic and genetic algorithms. International Journal of Computer Science and Network Security (IJCSNS) February, 8(2):27–32, 2008.

[10] Kruegel C, “Bayesian event classification for intrusion detection,” in Proc. of the 19th Annual Computer Security Applications Conference, Las Vegas, NV, 2003.

[11] Dickerson J, Dickerson J, , “Fuzzy network profiling for intrusion detection,” In Proc. of the 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), Atlanta, GA, pp. 301-306, 2000.

[12] Amor N, Benferhat S and Elouedi Z, “Naïve Bayes vs. decision trees in intrusion detection systems,” In Proc. of the 2004 ACM Symposium on Applied Computing, New York, pp. 420-424, 2004.

[13] Abadeh M, Habibi J, "A Hybridization of Evolutionary Fuzzy Systems and Ant Colony Optimization, for Intrusion Detection", Volume 2, Number 1 (pp. 33-46), Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, 2010.

[14] The KDD Archive. KDD99 cup dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.

[15] Tavallaee M, "A detailed analysis of the KDD CUP 99 data set", in Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.

AUTHOR PROFILES:

Aliakbar Tajari is a Ph.D student of Computer Science. Currently, he is an Associate Professor at Mirdamad University. His interests are in System Security and Intrusion Detection Systems.

Saied Tabarsa is MSc student in University of Mirdamad, Gorgan, Iran. His interests are in Parallel Systems and System Security.

Ziba Hosseini Nasab is MSc student in University of Mirdamad, Gorgan, Iran. His interests are in Cloud Computing Networks and Optimization Methods in Scheduling.

Fareevar Sedighi is MSc student in University of Mirdamad, Gorgan, Iran. His interests are in Intrusion Detection Systems and Evolutionary Algorithms in Parallel Computing Systems.