Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3 ,July 2021 : 2772 - 2779

Research Article

Enhanced Pre-Processing Techniques based Intrusion Detection Systems using Machine Learning

[R1]Roshni Prasad[1], Surya Prakash Sharma[22], Dr. CS Yadav[33]

1,2,3: Noida Institute of Engineering &Technology Gr. Noida (India)

Abstract-
These days, intrusion detection system (IDS) is the most arising pattern in our general public. This basically screen network traffic and will alarm the organization chairman of any unordinary action. IDS System work by one or the other searching for marks of known assaults or deviations of typical movement. While there are a few detriments of IDS, for example, low recognition rate and high bogus caution rate. In this paper a mixture IDS (HIDS) strategy dependent on support vector machine (SVM) and evidence theory (ET) has been proposed too different assault recognition method to limit the low bogus alert rate and improve exactness.

Keywords- IDS, HIDS, PHAD, ALAD, SNORT

1. INTRODUCTION

The malicious activity and policy violations on network of systems is continuously monitored by device or software application called Intrusion Detection System (IDS). The Security Information and Event Management (SIEM) system keeps track of the malicious activity reported either to an administrator or to the central database [1]. An IDS is not introduced to replace prevention-based techniques such as authentication and access control. Instead, it is implemented as a second line of defense to complement the existing security monitoring and the control component of the system. IDS passively monitor the data and unearth any potentially disastrous connection. Technically IDS are aimed at serving three important security functions i.e., monitoring the data, unearthing any potentially harmful transactions and finally responding to unauthorized activity. With the gigantic structure of the Internet, its distributed nature and lack of central security mechanism, the prevention of attacks is not possible and therefore detection and recovery from attacks become indispensable [2]. The IDS does exactly as the name suggests, it detects the possible intrusion. An intrusion can be in simple terms be defined as an attack on any or all of the security characteristics of the system i.e., confidentiality, availability, and integrity. In this section, we provide a detailed discussion about how various IDS's are categorized. A typical network configuration is depicted in Figure 1.1 In this network layout, a network is connected to an external network, there is a firewall to filter the disastrous connection originating from the outer world directed towards the network. The firewall is the first line of defense to block any harmful connections [3]. As can be seen from the figure, gateways are positioned at the entry of the network, so technically they are able to filter out the connections that originate from / directed to a host in the outside world. IDS's are positioned inside the network, rather than blocking the network connections. It is aimed at analyzing the network connections for the possible harmful connections. 8 IDS is aimed at unearthing potentially harmful connections that have somehow sneaked through the firewall. An IDS checks for the possible network attacks and initiates the corrective approach by alerting the system admin [4].

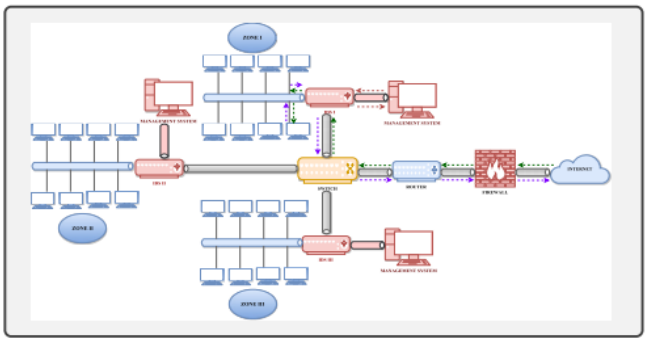

Figure 1: Typical Network Configuration

2. ML BASED IDS

Since the focus of this work is to build Machine Learning-based IDS, we present a block diagram of ML-based IDS shown in Figure 1.6. The process starts with the capturing of traffic records and the captured connections are forwarded to the IDS. The IDS process commences with the forwarding of captured connections to the data pre-processing unit. Here the records are transformed into appropriate form so as to be processed by ML techniques [5]. Once the data has been processed into the suitable format, the next step is the reduction of data. As the amount of data captured can be too large, an appropriate reduction of the data is necessary, so as to leave away the less important variables, without compromising much of the information. After the redundant attributes of the data-set are removed, the next step is to forward the data to the appropriate classifier. There can be a single classifier or a group of them laid in some order. The classifier will result in the model for the normal data. Once the model is ready, it can be tested for the effectiveness, using the test data [6]. Then there is a decision-making component, whose aim is to decide if the connection is normal or disastrous, for the cases where the connection is harmful, the IDS has to generate the alert and initiate the corrective procedure, by informing the system admin [7].

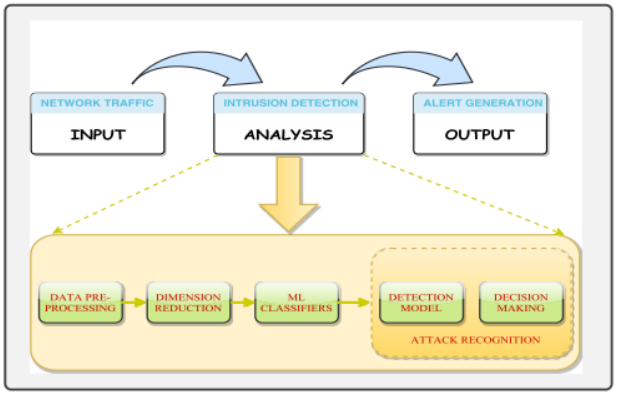

Figure 2: Architecture of ML based IDS

These customary firewalls can't identify a mind-boggling assault, for example, Denial of Service (DoS) and DDoS. Also, conventional firewalls can't differentiate between normal activity and DoS assault movement. Access control fills in as the cutting edge of resistance against interruptions, bolstering both confidentiality and integrity parameters. Intrusion detection is the process of progressively observing the events occurring in a PC or network, examining them for indications of conceivable episodes and often interdicting the unapproved access. Given the complex and fast-moving skyline of cyber security, it is not possible to hard code a machine or component with certain features and expect it to be working effectively at all the times. Rather the focus should be devising a module

which encompasses the previous knowledge and bases 10 the results on the experience that it is having. This makes the field fertile for the application of Machine Learning methods, wherein machines are not explicitly programmed, rather they are placed in some environmental conditions wherein they sense the patterns of interest and the patterns of interest, in this case, are intrusions.

3. LITERATURE REVIEW

Rane et al. (2019), in this paper, aimed at detection of internal intruders in HIDS. Commonly used login ids and passwords may be shared along with co-workers for professional purposes, which can be tampered or used by the attackers as a means of intrusion into the system details. The user was monitored and System Calls (SC) were extracted and the habitual SC pattern based on the habits of the user was taken into account and the profile of the user was stabilized. The forensic technique and other data mining techniques were applied at SC level host IDS to spot the internal attacks. Along with the user login credentials the forensic technique was applied to investigate the computer usage fashion against the collected user profile pattern and thereby check the identity of the user.

Liao et al., (2019), in this paper, With the decision rate threshold of 0.9, the system was able to perform with an accuracy rate of 94%. Nokia Research Center researchers modeled HIDS for mobile devices.

Miettinenet al., (2018), in this paper, along with various protection mechanisms accompanied with mobiles they felt an urge for attack monitor methods as a second line of defense. The framework was designed, taking into consideration the privacy of the mobile user in creating the user profile. The framework had a major share with the host-based intrusion detection inline with the network-based detection system, as researchers felt that mobile requires the monitoring system at both ends. The framework included data collection and IDS modules, the former entrusted with responsibility of monitoring the operating system activities, calculating the system measurements and the data collection at the application level and the later feeding on the collected and pre-processed data performs the actual intrusion detection.

Moon et al. (2018), targeted Advanced Persistent Threat attacks, by analyzing the 30 behavioral pattern of the host user through a 83-dimensional vector, each attribute representing one manner of the user. In order to form the database, they collected 8.7 million features from 4000 malicious and normal programs through the Virtual Machine (VM) environment. The system was designed in such way that frequency of occurrence of each behavior is calculated for each process. C4.5 decision tree was used to build a classifier for the collected information, and each new instance was analyzed against the tree to be segregated as malicious or normal instance. The model had a false positive rate of 5.8% and a false negative rate of 2.0%.

Moskovitch et al. (2018), in this paper, represented a novel HIDS aimed at discovering unknown malware codes. The collection of previous malware codes was taken as repository and each new sequence of behavior was compared with the repository to identify new malware code.

Le et al. (2018), in this paper, centered over detecting intrusion in Routing Protocol for Low Power and Lossy Networks (RPL) attacks. The operations of the RPL were converted into finite state machines through which the network was monitored and any malicious activity was detected. The research was further extended by wherein the simulation trace files were used to model the finite state machines to observe the RPL attacks. The model was further converted into a set of rules to monitor the data transferred between the network nodes. The drawback of the work was that the True Positive Rate was even able to reach 100% but the False Positive Rate was not that low ranging between 0 to 6.78%. Over that it also had an overhead of 6.3% in terms of energy when compared to normal RPL network.

4. PROPOSED METHODOLOGY

Supervised machine learning classifiers can be categorized into multiple types. These types include naïve Bayes, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), generalized linear models, stochastic

gradient descent, support vector machine (SVM), linear support vector classifier (Linear SVC) decision trees, neural network models, nearest neighbours and ensemble methods. The ensemble methods combine weak learners to create strong learners. The objective of these predictive models is to improve the overall accuracy rate. This can be achieved using two strategies. One of the strategies is the use of feature engineering, and the other strategy is the use of boosting algorithms. Boosting algorithms concentrate on those training observations which end up having misclassifications. There are five vastly used boosting methods, which include AdaBoost, CatBoost, LightGBM, XGBoost and gradient boosting.

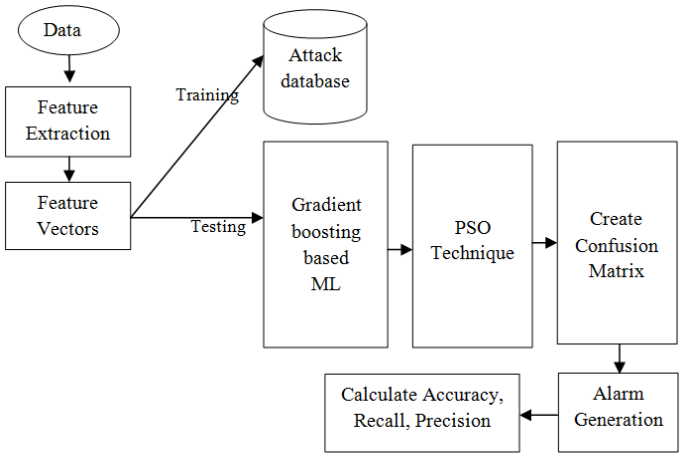

Figure 3: Flow Chart of Proposed Methodology

Gradient boosting (GB) sequentially creates new models from an ensemble of weak models with the idea that each new model can minimize the loss function. This loss function is measured by gradient descent method. With the use of the loss function, each new model fits more accurately with the observations, and thus the overall accuracy is improved. However, boosting needs to be eventually stopped; otherwise, the model will tend to overfit. The stopping criteria can be a threshold on the accuracy of predictions or a maximum number of models created.

Particle Swarm Optimization

PSO calculation was created at first by Kennedy and Eberhart in 1995. This calculation is a nature roused calculation that is propelled from the conduct of flying creature rushes called as a swarm. In this calculation, every arrangement is spoken to as a vector called as a molecule (winged animal). Here, the populace (swarm) may contain any irregular number of introductory arrangements (particles).each molecule begins with its underlying position and speed, at that point moves in the arrangement space to accomplish the ideal outcome. The primary computational strides of PSO incorporate producing beginning and speed of every molecule in populace, refreshing position and speed for a specific number of ages to get the ideal answer for accomplish beyond any doubt goals.

Let us discuss about the mathematical computation of PSO algorithm. Let any particle x_k(solution) in n-dimensional space is represented in equation.

$$x_k = \{x_{k1}, x_{k2}, x_{k3} \ldots\ldots\ldots\ldots x_{kn}\}$$

Where k=1,2,3,…,d and d is the number of particles in the swarm.

Each particle maintains its own velocity, let represented as given in equation 4.2.

$$v_k = \{v_{k1}, v_{k2}, v_{k3} \ldots\ldots\ldots\ldots v_{kn}\}$$

Additionally, in this calculation every molecule keeps up its own best position called as pbest and a best arrangement among every one of the particles called as gbest. In every emphasis or age, the particles move towards ideal arrangement by refreshing their speed and position as indicated by the recipe given

$$v_k(t+1) = w * v_k(t) + c1r1(p_k(t) - x_k(t)) + c2 * r2(g_k(t) - x_k(t))$$

$$x_k(t+1) = x_k(t) + v_k(t+1)$$

Where vk(t+1) speaks to the speed of k^{th} molecule at t+1 emphasis. W is the inactivity weight, vk(t) speak to speed of kth molecule at t cycle. pk(t), gk(t) speaks to the individual best of the molecule and worldwide best of swarm at t emphasis separately. c1 and c2 are two positive genuine constants known as self-assurance factor and swarm certainty factor individually. r1 and r2 are any irregular number produced in the middle of [0,1].from the study I, it has been demonstrated that bigger dormancy weight performs more proficient worldwide inquiry and littler latency weight performs effective nearby hunt. Henceforth this latency weight can be considered as an essential parameter to tune the execution of PSO calculation. This paper proposes a novel methodology to shift the dormancy weight in every cycle to play out the productive worldwide inquiry.

5. RESULT ANALYSIS

The recall or sensitivity gives the true positive rate TPR which is defined as the proportion of the attacks that were correctly identified and calculated using the following formula

$$\text{Re}\,call = \frac{TP}{TP + FN}$$

Precision is also termed as positive predictive value which is defined as the proportion of the predicted attacks that were correct and it can be computed by using the following formula

$$\text{Pr}\,ecision = \frac{TP}{TP + FP}$$

Table 1: Comparison Result

IDS/ attacks types	Total attacked	TP	FP	Recall	Precision
PHAD	10000	3198	6928	0.27	0.21
ALAD	10000	3399	6489	0.37	0.23
SNORT	10000	3608	6288	0.42	0.27
HIDS SVM	10000	3898	6178	0.52	0.32

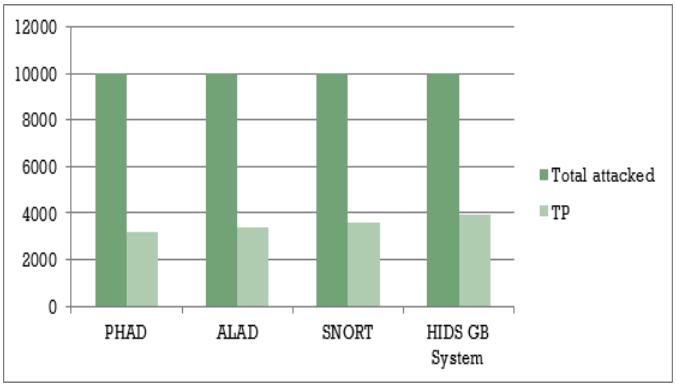

Figure 4: Bar Graph of the Ture Positive Value by HIDS GB System

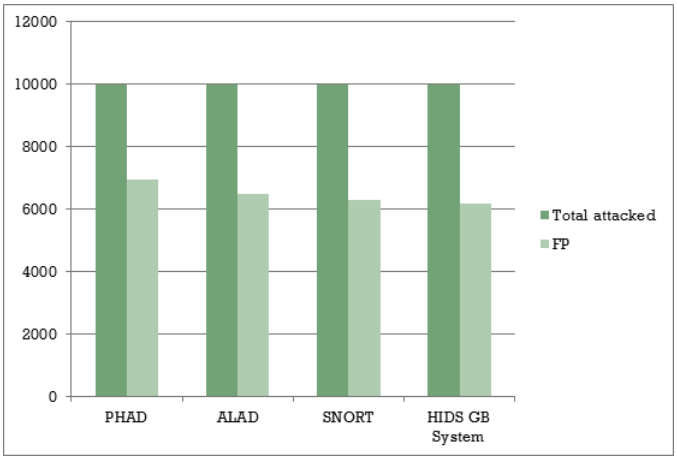

Figure 5: Bar Graph of the False Positive Value by HIDS GB System

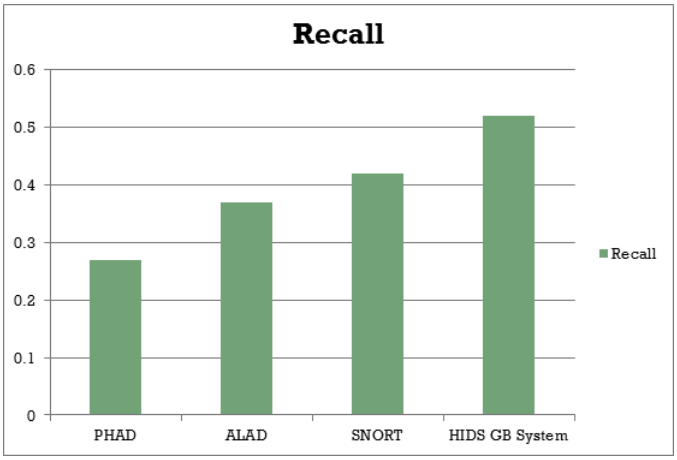

Figure 6: Bar Graph of the Recall Value by HIDS GB System

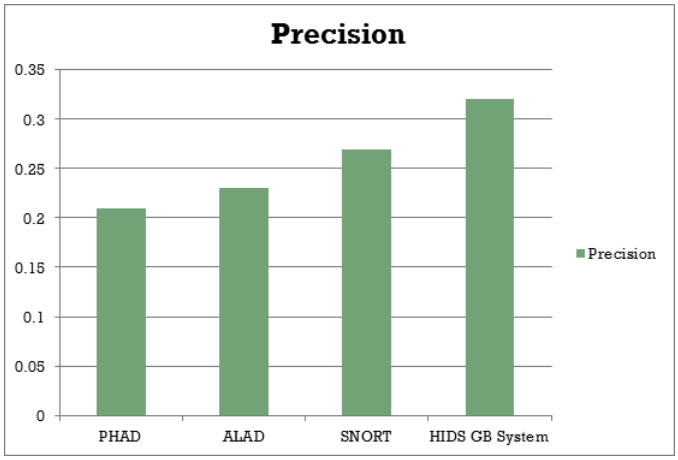

Figure 7: Bar Graph of the Precision Value by HIDS GB System

6. CONCLUSION

The last two decades have witnessed the explosive growth of the Internet and at the same time invited the enthusiasm of destructive users, thereby resulting in rapid growth of Internet-based attacks on a large scale. This emphasizes the need for securing the networks from the possible threats both external and internal. The main motivation of this research work is to monitor and analyze the network traffic for anomaly detection. This research was the first effort of its kind, to implement any of the NLDR techniques for dimension reduction of the network intrusion data-sets. Till now the DR techniques had been used only for data visualization. This research applied the DR techniques for enhancing the detection rate of the classifier. A rich set of DR's was implemented in this research work and was subjected to comparison to traditional linear DR techniques and the comparisons prove the fact that data visualization techniques enhance the detection rate of the classifiers provided adequate cluster magnification occurs.

The trial result shows that the proposed HIDS using gradient boosting are and strong in addressing the information as it was capable to lessen the information and henceforth essentially decreases the time needed to recognize the assaults in the organization traffic.

REFERENCES-

[1] ZHIYOU ZHANG, PEISHANG PAN "A hybrid intrusion detection method based on improved fuzzy C-Means and SVM" at International Conference on Communication Information System and Computer Engineer [CISCE] in 2019.

[2] AFREEN BHUMGARA, ANAND PITALE, "Detection of Network Intrusion Using Hybrid Intelligent System" at International Conferences on Advances in Information Technology in 2019.

[3] RITUMBHIRA UIKEY, Dr. MANARI CYANCHANDANI " Survey on Classification Techniques Applied to Intrusion Detection System and its Comparative Analysis" at 4th International Conference on Communication $ Electronics System (ICCES 2019) IEEE Conference Record #45898; IEEE Xplore ISBN; 978-1-7281-1261-9 in 2019.

[4] R. CHIRAKAR, C.HUANG, " Anomaly Based Intrusion Detection Using Hybrid Learning Approach of Combining K-medoids Clustering and Naïve Bayes Classification, at 8th International Conference on Wireless Communications, Networking and Mobile Computing, Shangai in 2012.

[5] Rane (2019). Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. Computer Communications, 35(7):772 – 783.

[6] Liao. (2019). An Autonomous Labeling Approach to Support Vector Machines Algorithms for Network Traffic Anomaly Detection. Expert Systems with Applications, 39(2):1822–1829.

[7] Dash, S. K., Rawat, S., and Pujari, A. K. (2018). Use of Dimensionality Reduction for Intrusion Detection. In Proceedings of Springer International Conference on Information Systems Security. 306–320.

[8] Datti, R. and Lakhina, S. (2018). Performance Comparison of Features Reduction Techniques for Intrusion Detection System. International journal of Computer Science and Technology, 10(3).

[9] Eesa, A. S., Orman, Z., and Brifcani, A. M. A. (2018). A Novel Feature-Selection Approach based on the Cuttlefish Optimization Algorithm for Intrusion Detection Systems. Expert Systems with Applications, 42(5):2670–2679.

[10] Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., and Fahmy, M. M. (2018). A Hybrid Network Intrusion Detection Framework based on Random Forests and Weighted k-means. Ain Shams Engineering Journal, 4(4):753–762

[11] Farquad, M. and Bose, I. (2018). Preprocessing Unbalanced data using Support Vector Machine. Decision Support Systems, 53(1):226–233.

[12] Fawaz, A. M. and Sanders, W. H. (2018). Learning Process Behavioral Baselines for Anomaly Detection. In Proceedings of IEEE Twenty Second Pacific Rim International Symposium on Dependable Computing. 145–154.

[13] Gharibian, F. and Ghorbani, A. A. (2017). Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection. In Proceedings of IEEE Fifth Annual Conference on Communication Networks and Services Research. 350–358.

[14] Gong, W., Fu, W., and Cai, L. (2017). A Neural Network based Intrusion Detection Data Fusion Model. In Proceedings of IEEE Third International Joint Conference on Computational Science and Optimization. 410–414.

[15] Ham, J., Lee, D. D., Mika, S., and Scholkopf, B. (2017). A Kernel view of the ¨ Dimensionality Reduction of Manifolds. In Proceedings of ACM Twenty-First International Conference on Machine learning. 47–52.

[16] Hamdi, M. and Boudriga, N. (2017). Detecting Denial-of-Service attacks using the Wavelet Transform. Computer Communications, 30(16):3203–3213