Intrusion Detection Techniques with Features Classification: Survey & Discussions

Pooja kushwah[1], Prof. Praveen Kataria[2]

[1]M. Tech Scholar, Department of CSE, ASCT, Bhopal (India)

[2]Professor, Department of CSE, ASCT, Bhopal (India)

[1]er.poojakkushwaha@gmail.com, [2] praveenkataria2008@gmail.com

ABSTRACT

Nowadays, the Internet is accessible from all over the place. With the growing number of electronic devices connected to the web, computer network security could be endangered. This problem has raised the question on how to effectively defend the computer network from internal and external attacks, Due to the dynamic nature of wireless network, security and efficient intrusion detection system (IDS) is a challenging task to detect the intruder nodes. The classification algorithm is used to detect the intrusions in an efficient manner. In this paper we present the literature survey and the problem identification for the current intrusion detection system.

Keywords: Anomaly Intrusion Detection, Intrusion Detection System (IDS), Naive Bayes Classifier; KDDCUP, Machine learning.

INTRODUCTION

Information guarantee is a concern of serious global concern. The intricacy and openness of client/server technology pooled with Internet contain fetch about enormous profit to the progressive society; meanwhile, the hurriedly increasing, openness, high complexity, and increasing accessibility of the networks not only escort to exploitation of vulnerabilities in the communication protocol stack but moreover enlarge the risk of existing information security system. Network attack vulnerability depends upon the expensive information hacking by attackers. The attacker intrude the network system or system server and creating network dump, malicious activity, modification, data theft, flood or denial the system process. Network system affecting the lack of attacks and thus require to intellectual intrusion detection model to protect the network system. To builds proficient IDS, data mining techniques used to identify the intrusions and classify the attack patterns. In data, mining learning process demand vast amount of training data and grand complexity and analyze the contrast to recent obtainable methodologies. Attribute Selection, Classification and Accuracy procedure is a four-step procedure for intrusion detection systems [1].

The various network attacks can be categorized as User-to-Root (U2R) attacks, Denial-of-Service (DoS) attacks, Remote-to-Local (R2L) attacks and probe attacks. In DoS attack, the attacker interrupts or denies the user access to the server. Examples are Neptune, Ping of Death, Mailbomb, etc. In U2R, the attacker is allowed privileged access by the extension of the root permissions like those of the administrator, the most common example being the buffer overflow attack. In R2L attack, the assailant intrudes the target system illegally, without any permission from the owner. Last of all, probe attack gathers and analyses information with an aim to map the network system, e.g., scanning software like Satan, Mscan and Nmap collect information from the target system such as hostname, service application, IP address and operating system. Though this attack gathers only data, this information can be employed in attacks of various kinds in the future [2].

Today, the Internet faces threats from intelligent, automated and sophisticated malicious codes that are on the rise. It could be seen in the past that computer worms have the capability to disperse on their own, without human involvement and have the record of launching the worst attacks on computer networks. To provide defense against worms, intrusion detection systems are mostly employed that make use of self-replicating behavior of worms for detecting the signatures and patterns of malicious codes in the network. By the parameters they use for detection, these systems can be categorized as anomaly-based and signature-based systems. IDS or Intrusion Detection System is basically the software for the detection and monitoring of packets in the network traffic. As soon as some abnormal data packets are found that indicates the attack pattern, the system generates an alert. Thus, IDS is becoming a significant tool to secure data the network. Furthermore, it has become one of the most exciting research topics in the research community worldwide.

The implementation of IDS can be as Host-based IDS (HIDS) or Network-based IDS (NIDS) which are the two types. According to authors, HIDS is used to detect intrusions that cannot be detected by NIDS due to the system’s more massive scale and comprehensiveness. Achieving real-time detection and prompt response in NIDS are possible through a collection of information from the network rather than from each host. It can also retain the evidence of the attacks. Although they are different types of IDS, the functionalities are quite similar which is to detect intrusions in a computer network and alert the user accordingly. HIDS acts like a virus scanner, where it scans traffic destined for the host and generates alarm for any sign of malicious activity. As for NIDS, the concept is much simpler than HIDS. It connects the device to the network like a network protocol analyser. NIDS closely monitors all network traffic and generates an alarm to the user upon sensing any form of intrusion. IDS is a system which can help to reduce the risk of losing all the information and data stored on the network as it contributes to strengthening the security of a system. Meanwhile, traditional IDS are known as signature based, which only detect known patterns, and may give rise to problems as the nature of the network varies from time to time. A good system should not be a "black box," which means the inside of a working system should be examinable from outside [2].

Various approaches for feature reduction or classification have been proposed in order to improve the efficiency of IDS in detecting attacks [12]. Many of those are based on machine learning techniques for the purpose of optimizing IDS's feature selection, mainly for better attack classification process. Unfortunately, none of the proposed mechanisms is perfect there are always some shortcomings. Hence, we argue that there is a need for continuous study to improve the performance of IDSs. In fact, sometimes the classification method that might be suitable for a specific problem is not easy to address. This issue is presented in a well known No Free Lunch (NFL) theorem which states that there is no heuristic algorithm best suited for solving all optimization problems. Therefore, after our thorough study, Magnetic Optimization Algorithm (MOA) has been integrated and hybridized with PSO to optimize the performance of ANN in classifying network traffic of IDS.

The main contribution of this work is to achieve a robust classifier that could assist in constructing an accurate IDS, which secures the WANETs. In other words, our proposed MOA-PSO algorithm has improved the detection and classification rates to increase the efficiency of IDS. The IDS is tasked with monitoring and analyzing network activity to differentiate between normal and anomalous activities. If anomalous activity goes undetected, this could potentially cause severe damage to the infrastructure and reliability of a computer system. Therefore, the detection rate of anomalous activity must be maximized. Simultaneous to anomalous activity detection, the

IDS must minimize the false positive rate to avoid undue hassle and confusion. False positives do not put the system at risk but can become a problem if the rate at which they occur is high, which limits the IDS's ability to provide reliable and clear results. The balance between detection rate and false positive rate is the key for an effective IDS [4].

The rest of this paper is organized as follows in the first section we describe an introduction of about the intrusion detection techniques and applications. In section II we discuss about the Parameter optimization, In section III we discuss about the related work, their comparative study. In section IV we discuss about the problem identification related to literature work, Finally in section V we conclude and discuss the future scope.

II PARAMETER OPTIMIZATION

Parameter optimization has a great impact on the performance of classification. Therefore a lot of work has been done for parameter optimization which clearly help to better detect intrusions in networks and information systems. In this survey some of the important techniques for parameter optimization are discussed [9]. Optimization problem is addressed with the help of improved various algorithm in . It also introduced greedy algorithm to improve the efficiency of optimization algorithms. A mathematical model is devised and improved algorithm is also used to optimize the devised model. The proposed system answers the following three questions.

1. How to improve the reliability of the detection system resources constraints
2. How to minimize system resource under reliability constraints
3. How to construct a clear and complete mathematical model to address these two above concerns

In testing activities n-detection agents participated in given detection time. The reliability of each agent and its detection time is known.
Basic assumption:

1. There is no network congestion, transmission interruption and any other unforeseen circumstances.
2. Average detection time is known
3. Detection agent is at start and in working condition.

III RELATED WORK

[1] This paper proposed the hybrid efficient model used to analyze the optimal features in the data, and it improve the detection rate and time complexity effective. This approach deals with high false and low false negative rate issue, first pre-processed data should be correlation based particle swarm optimization with GR-CR (Gain Ratio & Co-Relation) combination of this approach provide learning based some important subset of features and shows progress in the accuracy and time complexity level. Next, the novel approach tested on KDD cup 99, ISCX and ITDUTM dataset. This approach-achieved machine learning methods and it provide better performance in terms of accuracy rate, time taken, precision, and recall of the networks. Proposed approach compared with the following classification methods: Tree, Bagging, Navie Bayes, RBF classifier, Multiclass classifier, Logistic. The simulation results, gave the high detection accuracy (99.7%) in KDD 1999, (98.3 %) in ISCX and (99.3%) in ITDUTM with fewer feature selection. [2] The proposed system has been verified to have high accuracy rate, high sensitivity as well as a reduction in false positive rate. Besides, the intrusions have been classified into four categories as Denial-of-Service (DoS), User-to-root (U2R), Remote-to-Local (R2L) and Probe attacks; and the alerts are stored and shared via a central log. Thus, the unknown attacks detected by other Intrusion Detection Systems can be sensed by any IDS in the network thereby reducing computational cost as well as enhancing the overall detection rate. The proposed system does not waste time by considering and analyzing all the features but takes into consideration only relevant ones for the specific attack and supervised learning neural network is used for intrusion detection. By the application of Snort before back-

propagation algorithm, the latter has only one function to perform detection of unknown attacks. In this way, the time for attack detection is reduced. [3] In this paper, a deep learning binomial classifier for Network Intrusion Detection System is proposed and experimentally evaluated using the UNSW-NB dataset. Three different experiments were executed in order to determine the optimal activation function, then to select the most important features and finally to test the proposed model on unseen data. The evaluation results demonstrate that the proposed classifier outperforms other models in the literature with 98.99% accuracy and 0.56% false alarm rate on unseen data. [4] In this Paper they implemented an Evolutionary General Regression Neural Network (E-GRNN) as a two-class classifier for intrusion detection based on features of application layer protocols (e.g., http, ftp, smtp, etc.) used in simulated network traffic activities. The E-GRNN is an evolutionary search-inspired General Regression Neural Network, which extracts the most salient features to reduce computational complexity and increase accuracy. Our research shows that the E-GRNN classifier was able to achieve a DR of 95.53% and an FAR of 2.11%. The balance between detection rate and false positive rate become more challenging when normal activity and anomalous activity are not static. The activity on the network can change and the IDS must be aware of this change and adapt accordingly. If not, the ability of the IDS to provide accurate and reliable results is greatly diminished. Therefore, an IDS must adapt to different environments, which potentially bring different activity and behavior unseen by the IDS. [5] In this paper, a novel hybrid model was proposed with the purpose of detecting network intrusion effectively. In the proposed model, Gini index is used to select the optimal subset of features, the gradient boosted decision tree (GBDT) algorithm is adopted to detect network attacks, and the particle swarm optimization (PSO) algorithm is utilized to optimize the parameters of GBDT. The performance of the proposed model is experimentally evaluated in terms of accuracy, detection rate, precision, F1-score, and false alarm rate using the NSL-KDD dataset. Experimental results show that the proposed model is superior to the compared methods. In order to validate the performance of our method they performed experiments on the NSL-KDD dataset compared with six baselines. Five evaluation criteria are introduced to conduct fair comparisons, which are accuracy, detection rate, precision, F1-score, and false alarm rate. The experimental results demonstrated that the proposed model performs the best on the whole in comparison with baselines.

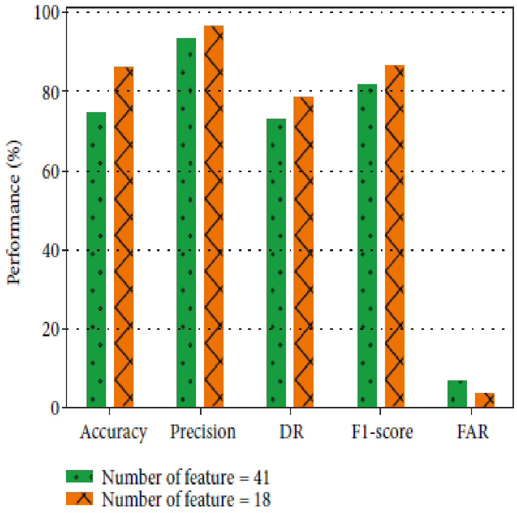

Fig 1: Performance comparison of the proposed methods with different features.

[6] In this paper they propose a fuzzy aggregation approach using the modified density peak clustering algorithm (MDPCA) and deep belief networks (DBNs). To reduce the size of the training set and the imbalance of the samples, MDPCA is used to divide the training set into several subsets with similar sets of attributes. Each subset is used to train its own sub-DBNs classifier. These sub-DBN classifiers can learn and explore high-level abstract features, automatically reduce data dimensions, and perform classification well. According to the nearest neighbor criterion, the fuzzy membership weights of each test sample in each sub-DBNs classifier are calculated. The output of all sub-DBNs classifiers is aggregated based on fuzzy membership weights. Experimental

results on the NSL-KDD and UNSW-NB15 datasets show that our proposed model has higher overall accuracy, recall, precision and F1-score than other well-known classification methods. Furthermore, the proposed model achieves better performance in terms of accuracy, detection rate and false positive rate compared to the state-of-the-art intrusion detection methods. [8] A novel technique is called Simulated Annealing based Naive Bayes classifier (SA-NBC) is proposed to detect Anomaly Intrusion Detection in wireless ad-hoc network. An anomaly-based intrusion detection system is an essential one to observe the network activities and classify whether it either normal or anomalous node. Hence, the accuracy of intrusion detection is enhanced. In proposed SA-NBC technique, simulated annealing is used to choose the optimal feature of the node and thus detect the intrusion in the network. Based on these optimal features, the Naive Bayes classifier is used to classify the malicious node and normal node with the aid of calculating conditional probability. It outlines the vector representation for detecting the network intrusions and observes network behavior and classifying the node as either normal or abnormal (anomalous). The experiments are conducted on different parameters such as anomaly intrusion detection accuracy, execution time and throughput.

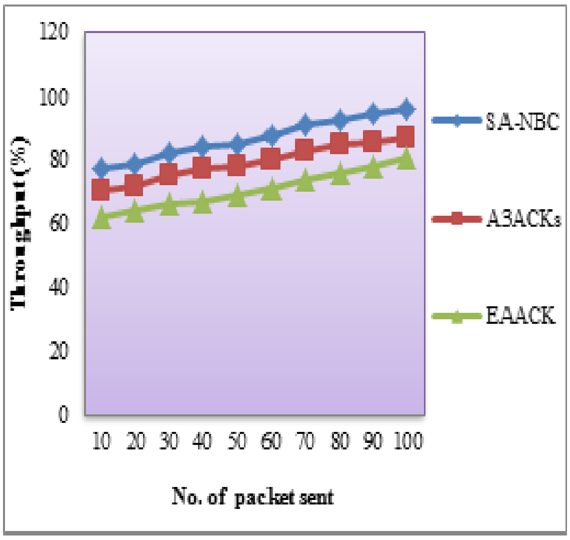

Fig 2: Measure of throughput.

[9] This study proposed a model based on extensive survey to create an efficient hybrid classifier which is jointly based on feature selection, parameter optimization and classification. Feature selection is adapted to refine the area of interest by improving the accuracy of classification, then to optimize the parameters, genetic algorithm (GA) is the most appropriate technique to be used. Parameters optimization using GA also plays a remarkable role to improve classification using support vector machine (SVM). SVM is considered a suitable machine learning technique for classification of intrusions which are detected both in networks and information systems. Finally SVM will classify the observed activity as normal or attack using adopted linear or nonlinear techniques. The proposed solution paves a way to improve accuracy by efficiently detecting the intrusions within real time applications of network and systems. [10] With the rapid increase in the data generation the internet, we need an efficient and real-time intrusion detection system that can cater the growing attacks over the high-speed network. This paper presents a decision tree-based real-time IDS that have the ability to work at high-speed environment. The technique uses the fewer number of flow features i.e. nine best features that are selected amongst forty-one from KDD99 intrusion dataset using FSR and BER techniques. The accuracy of the proposed IDS is evaluated in terms of true positive (TP- more than 99%) and false positive (FP- less than 0.001 %), and efficiency in terms of processing time. The higher accuracy and efficiency make the system to be able to work in a real-time and high-speed environment. [11] In this paper, they proposed a hybrid learning approach through a combination of K-Medoids clustering, Selecting Feature using SVM, and also Naïve Bayes classifier. The KDD CUP’99 benchmark dataset was used for evaluation. The experimental results obtained showed that their proposed approach was an efficient one. In this method, a new training dataset is created by K-Medoids clustering and Selecting Feature using SVM. Then its performance is evaluated by the Naïve Bayes

classifier. The results obtained showed that the proposed method performed well in terms of accuracy, detection rate, and also false alarm rate.
[12] In this paper they propose an efficient IDS based on hybrid heuristic optimization algorithm which is inspired by magnetic field theory in physics that deals with attraction between particles scattered in the search space. Our developed algorithm works in extracting the most relevant features that can assist in accurately detecting the network attacks. These features are extracted by tagged index values that represent the information gain out of the training course of the classifier to be used as a base for our developed IDS. In order to improve the accuracy of artificial neural network (ANN) classifier, they have integrated our proposed hybrid magnetic optimization algorithm-particle swarm optimization (MOA-PSO) technique. Experimental results show that using their proposed IDS based on hybrid MOA-PSO technique provides more accuracy level compared to the use of ANN based on MOA, PSO and genetic algorithm. Updated KDD CUP data set is formed and used during the training and testing phases, where this data set consists of mixed data traffics between attacks and normal activities. [13] This paper introduces a new similarity measure, the covering similarity, which we formally define for evaluating the similarity between a symbolic sequence and a set of symbolic sequences. A pair wise similarity can also be directly derived from the covering similarity to compare two symbolic Sequences. An efficient implementation to compute the covering similarity is proposed which uses a suffix-tree data structure, but other implementations, based on suffix array for instance, are possible and are possibly necessary for handling very large scale problems. They have used this similarity to isolate attack sequences from normal sequences in the scope of host-based Intrusion detection.

IV PROBLEM IDENTIFICATION

However, there are still many problems with intrusion detection systems. First, different types of network traffic in a real network environment are imbalanced, and network intrusion records are less than normal records. As a result, unbalanced network traffic greatly compromises the detection performance of most classifiers. The classifier is biased towards the more frequently occurring records, which reduces the detection rate of small attack records such as U2R and R2L records. Second, due to the high dimensionality of network data, the feature selection methods in many intrusion models are first considered as one of the preprocessing steps, such as PCA (Principal Component Analysis) and chi-square feature selection. However, these feature selection methods rely heavily on manual feature extraction, mainly through experience and luck, and these algorithms are not effective enough. Third, because of the large amount of network traffic and complex structure, the traditional classifier algorithm cannot achieve higher attack detection rate with lower false positive rate. Fourth, the network operating environment and structure in the real world are changing, for example, the popularity of the Internet of Things and the widespread use of cloud services, as well as various new attacks are emerging. Many unknown attacks do not appear in the training dataset.

V CONCLUSION AND FUTURE WORK

The Internet has become a crucial part of everyday communication via social media interaction, e-mail, e-learning, etc. Besides, small and large corporations have extended their consumer base by providing direct customer marketing, internet shopping and inter-company correspondence using basic Internet communication, here we discuss about the types of intrusion detection system and the techniques used to solve the problem of attackers, in future we implement a hybrid techniques and improve the performance of network and ensure the network very safe and secure.

REFERENCES:-

[1] Sivasangari Gopal, Sathya M, “ A Feature Selection for Intrusion Detection System Using a Hybrid Efficient Model”, International Journal of Scientific Research in Computer Science,

Engineering and Information Technology, 2018, Pp 1917-1929.

[2] Rashidah Funke Olanrewaju, Burhan Ul Islam Khan, Athaur Rahman Najeeb, Ku Nor Afiza Ku Zahir, Sabahat Hussain, "Snort-Based Smart and Swift Intrusion Detection System", Indian Journal of Science and Technology, 2018, Pp 1-9.

[3] Malek Al-Zewairi, Sufyan Almajali, Arafat Awajan, "Experimental Evaluation of a Multi-Layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System", International Conference on New Trends in Computing Sciences, IEEE 2017, Pp 167-173.

[4] James Brown, Mohd Anwar, Gerry Dozier, "An Evolutionary General Regression Neural Network Classifier for Intrusion Detection", IEEE 2016, Pp 1-5.

[5] Longjie Li , Yang Yu, Shenshen Bai, Jianjun Cheng, Xiaoyun Chen, "Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO", Journal of Sensors, 2018, Pp 1-10.

[6] Yanqing Yang, Kangfeng Zheng, Chunhua Wu, Xinxin Niu, Yixian Yang, "Building an Effective Intrusion Detection System Using the Modified Density Peak Clustering Algorithm and Deep Belief Networks", Applied Science Journal, 2019, Pp 1-25.

[7] Shadi Aljawarneh, Monther Aldwairi, Muneer Bani Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science, 2017, Pp 1-9.

[8] K. Murugan, P. Suresh, "Efficient Anomaly Intrusion Detection Using Hybrid Probabilistic Techniques in Wireless Ad Hoc Network", International Journal of Network Security, 2018, Pp 730-737.

[9] Asghar Ali Shah, M. Khurram Ehsan, Kashif Ishaq, Zakir Ali, Muhammad Shoaib Farooq, "An Efficient Hybrid Classifier Model for Anomaly Intrusion Detection System", International Journal of Computer Science and Network Security, 2018, Pp 127-137.

[10] M. Mazhar Rathore, Faisal Saeed, Abdul Rehman, Anand Paul, Alfred Daniel, "Intrusion Detection using Decision Tree Model in High-Speed Environment", International Conference on Soft-computing and Network Security, IEEE 2018, Pp 1-5.

[11] L. Khalvati, M. Keshtgary, N. Rikhtegar, "Intrusion Detection based on a Novel Hybrid Learning Approach", Journal of AI and Data Mining, 2018, Pp 157-162.

[12] Ali Safaa Sadiq, Basem Alkazemi, Seyedali Mirjalili Noraziah Ahmed, Suleman Khan, Ihsan Ali, Al-Sakib Khan Pathan, Kayhan Zrar Ghafoor, "An Efficient IDS Using Hybrid Magnetic Swarm Optimization in WANETs", IEEE Access 2018, Pp 29041-29052.

[13] Pierre-Francois Marteau, "Sequence Covering for Efficient Host-Based Intrusion Detection", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, 2019, Pp 994-1006.