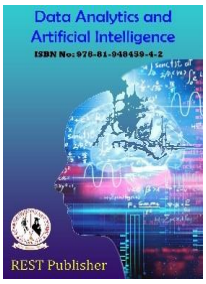
Data Analytics and Artificial Intelligence
Vol: 3(7), 2023
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/



Network Attack Detection Using Machine Learning Approach

***K. Gowri, S. Brindha**
SPIHER, Chennai, Tamilnadu, India.
*Corresponding Author Email: akmgowri@gmail.com

Abstract: *The proliferation of IoT devices in everyday human life has made their security a critical requirement. Currently those devices are not very secure because of several reasons. First, manufacturers do not account much for security, releasing products that are vulnerable to attacks, thus leaving users with security issues that are unlikely to be resolved. Second, many IoT devices do not have enough computing power to run an antivirus or even do not allow one to install an antivirus. Finally, the heterogeneity which characterizes the IoT in terms of applications, hardware, and software, expands the attack surface, while at the same time increases the difficulty of deploying all-encompassing security solutions. Despite some sort of security provided by IoT enabling technologies (e.g., communication protocols), or by intrusion prevention systems (e.g., network firewalls), attackers still find ways to compromise devices, or the communication between them. Due to the importance of system's safety measure the research in data stream mining and Intrusion detection system gained high attraction. In this paper, we represent the mechanism to improve the efficiency of the IDS using Stacked heuristic ensemble classification algorithm (SHECA). Feature selection is done by using Darwinian partical swarm optimization (DPSO).*

1. INTRODUCTION

IDSs may be a piece of hardware or software systems which is used to detect intruders on the network. IDS systems can be distinguished according to where they're installed i.e either on the host or on the network, as well as they differ on how they detect intruders i.e. misuse detection and anomaly detection. While different types of IDS systems exist, each type of IDS has its own benefits and drawbacks. There are two types of IDS i.e Host based IDS and Network based IDS.A host based Intrusion Detection System is a system that monitors a system that it is installed on to detect the misuse or intrusion by notifying the authority or by logging the activity. One can think of a Host based IDS as mediary that monitors whether anything or anyone, whether domestic or foreign, has bypassed the system's security policy. A Host based IDS analyses the traffic directed towards and traffic sent from the specific computer on which the IDS is installed. A host-based system also has the capability to oversee key system files and any attempt to overwrite these files. A Network based IDS consist of hardware sensors located at discrete points on the network, while it may also contain the software that is installed on various computers connected along the network. These types of IDS analyses the data packets both entering and leaving the system and offering real time detection.

2. STREAM DATA MINING

Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.

3. DETECTION METHOD

Signature-based:

Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. This terminology originates from antivirus software, which refers to these detected patterns as signatures. Although signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available.

Anomaly-based:

Anomaly-based intrusion detection systems were primarily introduced to detect unknown attacks, in part due to the rapid development of malware. The basic approach is to use machine learning to create a model of trustworthy activity, and then compare new behavior against this model. Although this approach enables the detection of previously unknown attacks, it suffers from false positives:previously unknown legitimate activity may also be classified as malicious.

4. RELATED WORK

Gisung Kim et.al, [1] proposed a new hybrid method that hierarchically combines a misuse detection and anomaly detection in a decomposed structure. First, the C4.5 decision tree was used to create the misuse detection model that is used to disintegrate the normal training data into smaller subsets. Then, the one-class support vector machine was used to create an anomaly detection model in each decomposed region. Throughout the integration, the anomaly detection model can indirectly use the known attack information to enhance its ability when building profiles of normal behavior. This is the first attempt to use the misuse detection model to enhance the ability of anomaly detection model.C4.5 decision tree does not form a cluster, which can degrade the profiling ability thus abbreviating the efficiency of the system. Shi-Jinn Horng et.al, [2] proposed an intrusion detection system, which combines a clustering algorithm, a simple feature selection algorithm, and the Support Vector Machine (SVM). In this study, in addition to a simple feature selection method, they proposed an SVM-based network intrusion detection system with BIRCH hierarchical clustering for data pre-processing. The BIRCH hierarchical clustering provides a highly qualified and reduced datasets, in place of original large dataset, for SVM training. In addition to reduction of the training time, the resultant classifiers showed better performance than the SVM classifiers using the originally redundant dataset. However, in terms of accuracy, the proposed system could obtain the best performance at 95.72%. This approach provides better performance in terms of accuracy in comparison to the other NIDS (Network based IDS). It only detects Dos and Probe attacks not U2L and R2L attacks. Hong Kuan Sok et.al,[3] presents a paper on using the ADTree algorithm for feature reduction. ADTree also gives good classification performance. In addition, its comprehensible decision rules endows the user to discover the features that heads towards better classification. This knowledge base facilitates to design a smaller dimension of support vectors for suitable classifier. The experiment supports the idea of using this algorithm as both knowledge discovery tool and classification. The classification task has been simplified and the speed increased drastically due to the reduced operations required to implement the classification. Tavallaee et.al, [4] presented a paper on KDD CUP 99 Data Set and after the analysis of the entire KDD dataset it showed that there were two important issues in the data set which affected the performance of evaluated systems, and thus results in a very poor interpretation of anomaly detection approaches. To overcome the issues, NSL-KDD was proposed, which contains selected records of the KDD data set. Although, the proposed data set suffers from some problems and may not be a ideal representative of existing networks, due to the lack of public data sets for networkbased IDSs, they believe that the dataset still can be used as an effective benchmark to help analyst analyze different intrusion detection methods. Abhijit D. Jadhav et.al [6] IDS implementation using machine learning techniques is proposed with Distributed & Parallel approach for simultaneous processing of network traffic data by these different machine learning techniques. This approach enables us to implement this IDS design for Online streamed data for faster detection i.e. increasing efficiency of IDS. Xiaokang Zhou wt.al[7] mitigates the inconsistency between dimensionality reduction and feature retention in imbalanced IBD, we propose a variational long short-term memory (VLSTM) learning model for intelligent anomaly detection based on reconstructed feature representation. An encoder–decoder neural network associated with a variational reparameterization scheme is designed to learn the low-dimensional feature representation from high-dimensional raw data. Three loss functions are defined and quantified to constrain the reconstructed hidden variable into a more explicit and meaningful form. Tran Hoang Hai et.al [8] propose novel architecture of distributed log processing and storage tools to improve N-IDS data processing. Our goal is to improve overall system performance and cost-efficient. In this paper, we recommend the use of distributed processing and storage tools to improve NIDS data processing by Apache Spark and make a comparison with previous works using Hadoop Cluster. Our proposed model introduces a real-time data streaming tool, e.g., Apache Spark Streaming, for near real-time analysis of log processing.

5. OBJECTIVES

To detect the Intrusion from network from NSL KDD dataset using SHECA algorithm.
To Reduce the dimension of features using feature selection methods
To improve the detection accuracy of the proposed model.
To evaluate the performance in terms of detection accuracy, sensitivity,specificity and computational time.

FLOW CHART

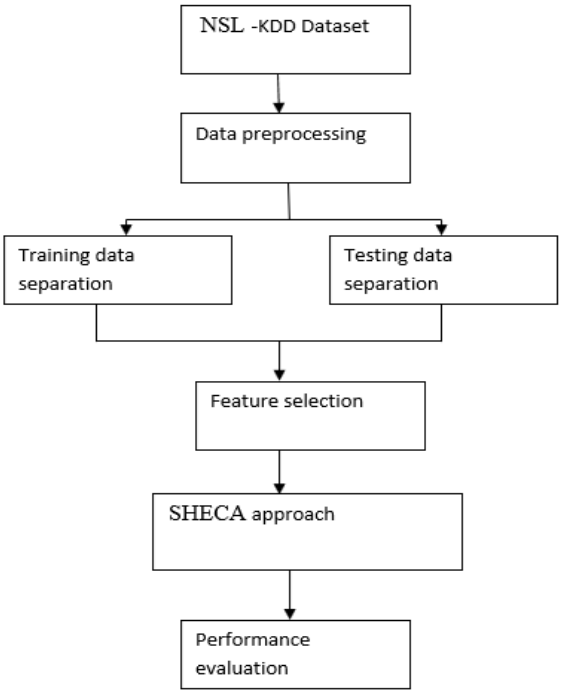


FIGURE 1. Flow Chart

The proposed SHECA technique detects different attacks provided in intrusion detection related dataset. In this work preprocessing like missing value filling, data cleaning is carried out. Whereas, DPSO based feature selection technique is applied to the data for the purpose of selecting informative features. After that, dataset is created by combining all the selected and extracted features. In the end, integrated feature set is used to train the SHECA model, which finally classifies the nine different malware attacks. Block diagram of the proposed technique is shown in above flowchart.

6. RESULTS

Performance metrics	SVM	decision tree	BIRCH	Proposed method
Accuracy	94.5	93.8	95.72	
specificity	89.6	88	92.5	
Sensitivity	91.5	90.99	93.89	

Software used:
MATLAB 2020a.

REFERENCES

[1]. Gisung Kim and Seungmin Lee (2014), A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection With Misuse Detection, ELSEVIER, Expert Systems with Applications vol. 41 pp. 1690 – 1700.

[2]. Shi-Jinn Horng and Ming-Yang Su (2011), "Novel Intrusion Detection System Based On Hierarchical Clustering and Support Vector Machines", ELSEVIER, Expert Systems with Applications. pp. 38 306-313.

[3]. Hong Kuan Sok et.al, "Using the ADTree for Feature Reduction through Knowledge Discovery" Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International ,pp1040 – 1044.

[4]. Tavallaee M, Bagheri E, Lu W, Ghorbani A. "A detailed analysis of the KDD CUP 99 data set", 2009 IEEE Symposium on Computational intelligence for security and defense applications, 2009,pp 1-6.

[5]. F. Amiri, M. Yousefi, C. Lucas, A. Shakery and N. Yazdani, "Mutual Information-Based Feature Selection for Intrusion Detection Systems", Journal of Network and Computer Applications, Vol. 34, 2011, pp.1184–1199.

[6]. Abhijit D. Jadhav, Vidyullatha Pellakuri," INTRUSION DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES FOR INCREASING ACCURACY AND DISTRIBUTED & PARALLEL APPROACH FOR INCREASING EFFICIENCY", 978-1-7281-4042-1/19/$31.00 ©2019 IEEE.

[7]. Xiaokang Zhou , Member, IEEE, Yiyong Hu , Member, IEEE, Wei Liang , Member, IEEE, Jianhua Ma, Member, IEEE, and Qun Jin , Senior Member, IEEE," Variational LSTM Enhanced Anomaly Detection for Industrial Big Data", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 17, NO. 5, MAY 2021.

[8]. Hai, T. H., & Khiem, N. T. (2020). Architecture for IDS Log Processing using Spark Streaming. 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). doi:10.1109/icecce49384.2020.9179188.