International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6322
***Article Received**: 30 December 2022 **Revised**: 02 February 2023 **Accepted**: 14 February 2023*

Enhancing the Efficiency of Attack Detection System Using Feature selection and Feature Discretization Methods

S. Revathy[1], S. Sathya Priya[2]
[1]Department of Computer Science & Engineering
Hindustan Institute of Technology and Science
Chennai.
E-mail: srevathy@hindustanuniv.ac.in
[2]Department of Computer Science & Engineering
lHindustan Institute of Technology and Science,
Chennai
E-mail: sathyap@hindustanuniv.ac.in

Abstract—Intrusion detection technologies have grown in popularity in recent years using machine learning. The variety of new security attacks are increasing, necessitating the development of effective and intelligent countermeasures. The existing intrusion detection system (IDS) uses Signature or Anomaly based detection systems with machine learning algorithms to detect malicious activities. The Signature-based detection rely only on signatures that have been pre-programmed into the systems, detect known attacks and cannot detect any new or unusual activity. The Anomaly based detection using supervised machine learning algorithm detects only known threats. To address this issue, the proposed model employs an unsupervised machine learning approach for detecting attacks. This approach combines the Sub Space Clustering and One Class Support Vector Machine algorithms and utilizes feature selection methods such as Chi-square, as well as Feature Discretization Methods like Equal Width Discretization to identify both known and undiscovered assaults. The results of the experiments using proposed model outperforms several of the existing system in terms of detection rate and accuracy and decrease in the computational time.

Keywords- Attack detection, Chi-square, Feature selection, Feature Discretization, Hybrid machine learning.

I. INTRODUCTION

Intrusion Detection System (IDS) is an important factor in cyber security domain. IDS analyses data acquired from network devices to discover, determine, and identify intrusions. Hackers or Cyber criminals can steal or modify the data by initiating active security attacks like Phishing, Man in the Middle Attack, Malwares and Denial of Services (DoS), which can be identified using IDS. The existing IDS have a central database of pre-defined signatures where signatures get updated on a given time or when the developer pushes the patch for the vulnerability [11]. The IDS can prevent the attack when it matches the pre-defined signatures but cannot prevent unknown or zero-day attacks.

These zero-day attacks can easily bypass the any type of protection system available in the network [1]. To address the challenge of detecting unknown and zero-day attacks, the anomaly intrusion detection is used,which relies on machine learning algorithms to identify such attacks. Machine Learning plays a major role in detecting attacks in modern era. Machine Learning enables computers to study and raise their experiences without the need for clear programming [16].

Machine learning involves the development of programs that can learn from data [12]. The machine learning algorithms can be generally classified into two main types as supervised and unsupervised learning. Supervised algorithms rely on labeled data from the past to make predictions about future actions. These datasets are used to train the algorithms to identify patterns or forecast outcomes accurately [17]. Over a period of time, a machine learning model can evaluate its own accuracy and improve by learning from its errors. In contrast, unsupervised learning involves the use of algorithms to analyze and cluster datasets that have not been labeled [15]. The unsupervised algorithms are used to discover hidden patterns in data without any manual intervention. The most common supervised algorithms include Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes, and Nearest Neighbors. DBSCAN and K-means anomaly detection are common algorithms used in unsupervised machine learning algorithms. The proposed model employs an unsupervised clustering-based attack detection system that utilizes sub-space clustering, one-class support vector machines, feature selection, and feature discretization. The integration of these techniques, it is possible to reduce

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6322
***Article Received**: 30 December 2022 **Revised**: 02 February 2023 **Accepted**: 14 February 2023*

computational time while achieving both high detection rates and low false positive rates.

The technique of feature selection is used to limit the input variables and improve the efficiency of the machine learning model by removing irrelevant or noisy data. Feature selection is a critical process in machine learning as it helps select the most relevant characteristics based on the type of problem the model is trying to solve. To increase the accuracy of detecting attacks and reduce dimensionality, the proposed model uses the Chi-square feature selection method

II. Related Work

Halimaa et. al. (2019) have investigated a classification methodology for IDS [5]. The accuracy and misclassification rate of intrusions are calculated using the SVM algorithm and Naïve's bayes. These methods are known for solving the classification issues. The NSL-KDD dataset is commonly used to evaluate the effectiveness of intrusion detection methods. The results [5] indicate that the SVM algorithm outperforms the Naïve Bayes method.

Rokade et. al. (2021) have proposed methods to identify intrusions [10]. Synthetic based intrusion dataset NSL-KDD was used to assess anomaly detection accuracy. Enhanced classification and high-class detection are probable because efficient rule structure. SVM, Nave Bayes, and ANN algorithms are demonstrated with diverse data sets in the experimental study, as well as system performance in a actual network situation. Several trials employed experimental analysis to calculate the algorithm's efficiency with a variation of tests, and were getting adequate outcomes.

Singhal et. al. (2021) has analyzed the most effective machine learning technique among the numerous available [13]. SVM, KNN, Naïve Bayes and Decision Tree were employed in the comparative analysis, alongside with the NSL-KDD dataset for Intrusion Detection Model. When tested independently, the accurateness of SVM, KNN, Naïve Bayes and Decision Tree is lesser than the inference detection model. The author concludes the inference detection rate is higher.

Yihunie et. al. (2019) has investigated anomaly intrusion detection systems that are unsupervised are getting better all the time [14]. The author tested five machine learning methods (SVM, Logistics Regression, Random Forests, Sequential Model and Stochastic Gradient Decent) to discover an effective algorithm that identifies abnormality traffic. The RF method clearly overtook the other four algorithms. RF achieves the best recall value, implying the least amount of false negatives.

Chuang et. al (2019) has proposed a hybrid technique that combines two current algorithms with C4.5 and Naïve Bayes which increases training period in IDS as well as the classification model training performance [3]. The proposed algorithm is capable of achieving satisfactory detection results while reducing the amount of training time required.

Ao et. al. (2021) has proposed a model to increase IDS processing efficiency [2]. The author has used Extra Tree, Decision Tree, Random Forest, Ridge and SGD Classifiers. The significant characteristics, and the recursion method uses to exclude unnecessary features, improving the model's accurateness and dependability are identified using NSL-KDD. Experimental results reveal that both the Extra Tree and RF classifiers do well, with the extra tree model providing good stability and accuracy even dealing with challenging difficulties.

Mazumder et. al. (2021) has proposed a model with feature selection for enhanced performance of the IDS [9]. The proposed model combines the unsupervised K-Means algorithm with the supervised Light GBM algorithm. The models is compared with Naïve Bayes, LGB, XGBoost, RF and AdaBoost. The author tested the model on NSL-KDD datasets.

Hakim et. al. (2019) has analyzed feature selection affects the IDS [4]. The ReliefF, Chi-Square, Gain ration and Information Gain methods has been used in the KNN, Naïve Bayes, RF and J48. Findings suggest that feature selection can improve IDS performance significantly, albeit at the cost of a little drop in accuracy.

III. Existing System

Grouping data to a similarity metric in an unsupervised approach is called Clustering technique. The purpose of using clustering technique is to achieve peak within-cluster similarity and small various-cluster similarity. There exist a few options. DBSCAN and K-means machine learning algorithms are two of the methods for grouping the input data.

A clustering technique creates sphere similar clusters from partitioned data. The technique is known as K-Means. It is comparatively effectual and applied to datasets of medium and large size. This method aims to reduce intra-cluster distances while increasing inter-cluster distances. However, it has the disadvantage of requiring amount of K clusters to be mentioned, therefore it is uneasy or obvious operation to be performed. Because form and scale of the delivery of points are so important, correct choice of K is frequently confusing. Furthermore, because the initial centroids are frequently chosen at random, the algorithm is prone to becoming stuck in the local optimal.

A clustering technique based on density creates any shape clusters. The technique is known as DBSCAN. The amount of points inside a radius is known as density. It's especially handy when dealing with spatially clusters. This method is capable of identifying data points that are completely surrounded by other clusters while remaining resistant to outliers. DBSCAN is also

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6322
***Article Received**: 30 December 2022 **Revised**: 02 February 2023 **Accepted**: 14 February 2023*

quite beneficial in numerous real-life applications because that does not need the definition of the amount of clusters, like the K-means method requires.

The detection of attacks is performed using an unsupervised machine learning algorithm known as Subspace Clustering and One Class Support Vector Machine (SSC-OCSVM).The One Class Support Vector Machine combined with Sub Space Clustering is an addition on classic grouping methods which generate clusters after the original dataset's various tiny subspaces. SVM is a pattern recognition and data analysis supervised learning model [6]. The OCSVM method is an addition of the SVM method that is particularly well suited to unlabeled data.

IV. Proposed System

The present clustering approaches are frequently unreliable. Clustering algorithms are influenced by a variety of factors, including their own internal workings. SVM is a pattern recognition and data analysis supervised learning model [7]. The OCSVM approach is an addition of the SVM method, particularly well-suited to unlabeled data. The SVM in OCSVM is trained on data with the regular class. The data gets converted into kernel's feature space and splits them from the beginning with the greatest possible boundary.

The method of Equal-Width Discretization was used to discretize the characteristics. To simplify the categorization process, feature values can be grouped into intervals. The Equal Width Discretization technique is used to make decision-making easier on continuously valued features. This discretization process helps streamline the decision-making process for continuous features. The Chi-Square is a statistical method for comparing the independence of two variables. Tests can also be performed to see whether it has any independence between the variables. Two variables are highly dependent, as indicated by the high-value result.

The steps in the approach of the presented SSC-OCSVM algorithm with feature discretization and selection methods are as follows.

Initialization Phase:

The initialization processes Split the feature space S into N distinct subspaces using a null difference vector V.

Feature Selection phase:

Apply the chi-square test and select the relevant feature.

Feature Discretization Phase:

Apply the equal width discretization method to group the values in the interval.

Learning and clustering Phase:

For each subspace, the dataset is subjected to subspace clustering, and then the OCSVM is applied to each partition

Gathering evidence Phase:

Update the dissimilarity vector V for each partition by accumulating the distances between the multiple outliers detected in subspace Si. This process is based on the concept of evidence accumulation clustering.

Detection of anomalies:

Create a ranked vector rank by ranking vector V. The related sample is considered an attack in Drank, if the unlikeness value is compared against a predetermined threshold value is greater.

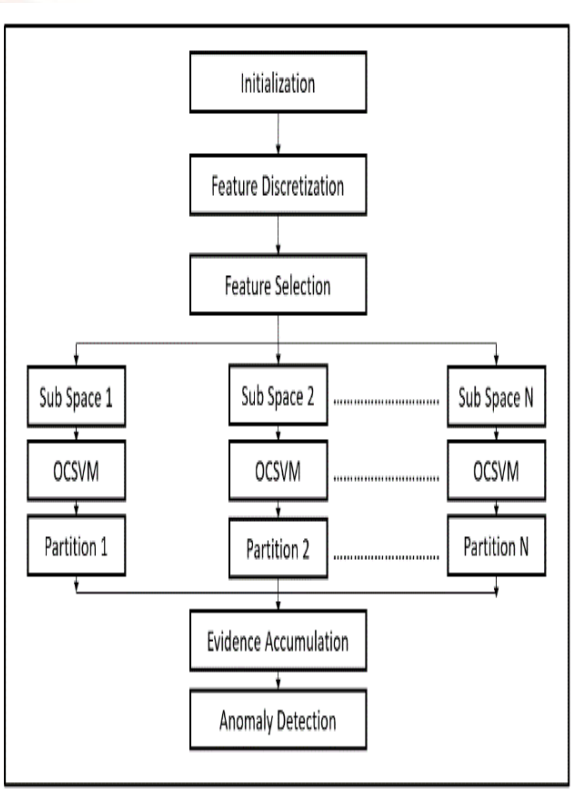

Fig 4.1 System Architecture

V Experimental Analysis

The proposed algorithm performance is compared with SSC-EA. Our results indicate that SSC-OCSVM outperforms the other algorithm in computational time and accuracy with feature selection and feature discretization method.

The Feature discretization method Equal Width Discretization is used to divide the data in to equal width in the bins to cluster the data. The equal width is calculated using the formula

W= (max-min)/N

Where W =width

Max= Maximum range of the value

Min= Minimum range of the value

N= Number of clusters.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6322
***Article Received**: 30 December 2022 **Revised**: 02 February 2023 **Accepted**: 14 February 2023*

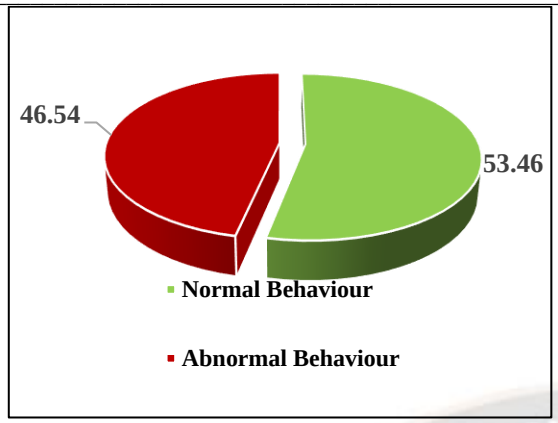

Fig 5.1 Classification of Attacks into normal and Abnormal

Attacks has been classified into normal and abnormal labels in Fig.5.1. In Fig 5.2 the abnormal labels has been classified into different multi class labels. The multi class labels are DoS, R2L, Probe and U2R.

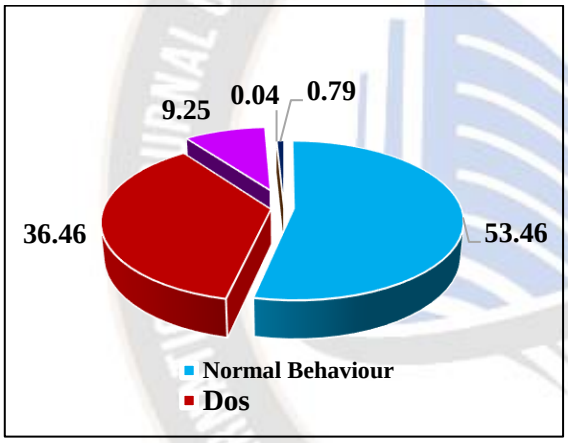

Fig 5.2 Classification of Abnormal Attacks

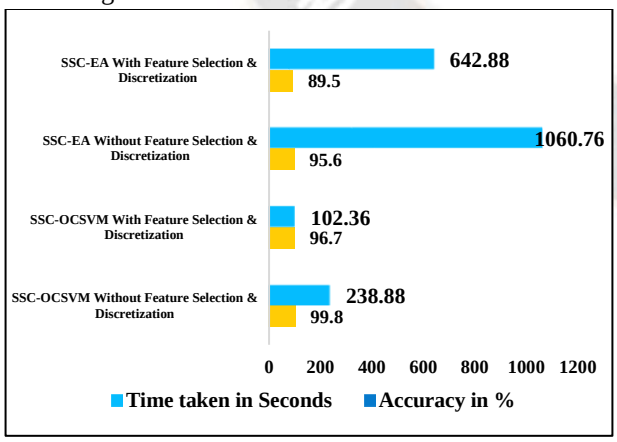

Fig. 5.4 Accuracy & Time Taken of Attack Detection after Feature Discretization & Feature Selection

The accuracy of feature selection method using Chi-Square and Equal width Discretization for different algorithms is shown in Fig. 5.4. The experiments revealed that, while there was a slight decrease in accuracy, the execution time was significantly improved. Furthermore, the proposed algorithm achieved high detection rates and low false positive rates.

VI Conclusion

The proposed hybrid machine learning attack detection algorithm with enhanced feature selection method like Chi-square and feature discretization methods like Equal Width Discretization integrated with SSC-OCSVM algorithm to identify security attacks without any previous familiarity, based on results obtained. The feature selection techniques on the attack detection with accuracy and detection time was investigated. The usage of feature selection and feature discretization in this method flourished in enhancing the quickness of analysis time with high detection rates and low positive rates while somewhat lowering the accuracy. In the future research, the development of a node recovery method by capturing live data is priortized.

References

[1] Al-Qatf, Majjed, Yu Lasheng, Mohammed Al-Habib, and Kamal Al-Sabahi. "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection." Ieee Access 6 (2018): 52843-52856.

[2] Ao, Huilong. "Using Machine Learning Models to Detect Different Intrusion on NSL-KDD." 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE). IEEE, 2021.

[3] Chuang, Po-Jen, and Si-Han Li. "Network intrusion detection using hybrid machine learning." 2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY). IEEE, 2019.

[4] Hakim, Lukman, and Rahilla Fatma. "Influence analysis of feature selection to network intrusion detection system performance using nsl-kdd dataset." 2019 International conference on computer science, information technology, and electrical engineering (ICOMITEE). IEEE, 2019

[5] Halimaa, Anish, and K. Sundarakantham. "Machine learning based intrusion detection system." 2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019.

[6] Liang, Dong, Qinrang Liu, Bo Zhao, Zhihua Zhu, and Dongpei Liu. "A clustering-svm ensemble method for intrusion detection system." In 2019 8th International Symposium on Next Generation Electronics (ISNE), pp. 1-3. IEEE, 2019.

[7] Liu, Lan, Pengcheng Wang, Jun Lin, and Langzhou Liu. "Intrusion detection of imbalanced network traffic based on machine learning and deep learning." IEEE Access 9 (2020): 7550-7563.

[8] Maseno, Elijah M., Zenghui Wang, and Hongyan Xing. "A Systematic Review on Hybrid Intrusion Detection System." Security and Communication Networks 2022 (2022).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6322
***Article Received**: 30 December 2022 **Revised**: 02 February 2023 **Accepted**: 14 February 2023*

[9] Mazumder, AKM Mashuqur Rahman, Niton Mohammed Kamruzzaman, Nasrin Akter, Nafija Arbe, and Md Mahbubur Rahman. "Network Intrusion Detection Using Hybrid Machine Learning Model." In 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-8. IEEE, 2021.

[10] Rokade, Monika D., and Yogesh Kumar Sharma. "MLIDS: A Machine Learning Approach for Intrusion Detection for Real Time Network Dataset." 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2021.

[11] Seo, Wooseok, and Wooguil Pak. "Real-time network intrusion prevention system based on hybrid machine learning." IEEE Access 9 (2021): 46386-46397.

[12] Shaukat, Kamran, Suhuai Luo, Vijay Varadharajan, Ibrahim A. Hameed, and Min Xu. "A survey on machine learning techniques for cyber security in the last decade." IEEE Access 8 (2020): 222310-222354.

[13] Singhal, Abhinav, Akash Maan, Daksh Chaudhary, and Dinesh Vishwakarma. "A Hybrid Machine Learning and Data Mining Based Approach to Network Intrusion Detection." In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 312-318. IEEE, 2021..

[14] Yehudi, Fekadu, Eman Abdelfattah, and Amish Regmi. "Applying machine learning to anomaly-based intrusion detection systems." 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT). IEEE, 2019

[15] Zoppi, Tommaso, Andrea Ceccarelli, and Andrea Bondavalli. "Unsupervised algorithms to detect zero-day attacks: Strategy and application." IEEE Access 9 (2021): 90603-90615.

[16] https://medium.com/cuelogictechnologies/evaluation-of-machine-learning-algorithms-for-intrusion-detectionsystem6854645f9211#:~:text=Anomaly%2Dbased%20intrusion%20detection%20%E2%80%94%20It,behav

[17] https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning