Acta Scientific COMPUTER SCIENCES

Volume 6 Issue 7 July 2024

Research Article

A Random Forest Approach and Principal Component Analysis in Intrusion Detection System Using Machine Learning

Rajavenkatesswaran KC[1]*, Ari Bharathi A[2], Gayathri R[2], Mathan C[2] and Vikram B[2]

[1]*Assistant Professor, Department of Information Technology, Nandha College of Technology, India*

[2]*UG-Final Year, Department of Information Technology, Nandha College of Technology, India*

***Corresponding Author:** Rajavenkatesswaran KC, Assistant Professor, Department of Information Technology, Nandha College of Technology, India.

Received: May 09, 2024

Published: June 11, 2024

Abstract

Malicious behavior on computer networks is monitored and detected by intrusion detection systems, or IDSs. Rule-based systems, which might be hard to maintain and might not be able to identify new attack vectors, are often used by traditional IDSs. IDSs are using machine learning (ML) methods more often since they can learn from data to identify new threats and increase their accuracy over time. To improve network security, the suggested intrusion detection system combines several machine learning approaches with the Random Forest algorithm. When categorizing complicated data, the Random Forests (RF) method yields excellent accuracy results. An established intrusion detection benchmark, the KDD Cup dataset, is used to assess the system's performance. In an ever-changing threat environment, this technique shows tremendous promise in detecting and mitigating harmful actions inside computer networks, improving cybersecurity.

Keywords: Intrusion Detection; Feature Selection; Machine Learning; Random Forest; Detection Rate

Introduction

In today's digital era, ensuring the security of computer networks and data has become of utmost importance. Given the increasing complexity of cyber threats and the interconnected nature of our systems, the necessity for robust network intrusion detection systems (NIDS) has never been more critical [1]. Intrusion detection plays a pivotal role in safeguarding organizations by identifying unauthorized access and mitigating potential threats to information systems. However, traditional intrusion detection methods often encounter difficulties in adapting to the everevolving threat landscape. To overcome these challenges and enhance the effectiveness of intrusion detection, we propose an innovative approach called [2] "Network Intrusion Detection with Two-Phased Hybrid Ensemble Learning and Automatic Feature Selection". This research endeavors to combine cutting-edge techniques from the fields of machine learning [13], data science, and cybersecurity. By integrating the power of ensemble learning and automatic feature selection into a two-phased detection system, we aim to revolutionize the field of network intrusion detection.

Feature selection

In today's ever-expanding digital landscape [3], the security of networks and information systems has become a top priority. With the rise of cyber threats, ranging from sophisticated malware to advanced persistent threats, it is crucial to continuously evolve network intrusion detection systems (NIDS) to prevent unauthorized access and malicious activities. The key to effective NIDS lies in selecting the most relevant data attributes, also known as "features." Feature selection is a critical process in machine learning and data analysis, with the primary objective of identifying and retaining informative attributes while discarding irrelevant or redundant ones. In the context of network intrusion detection, the careful selection of features is essential to improve both the efficiency and accuracy of the detection process.

Related work

As [11] wireless communication has advanced, there are several online security risks. The intrusion detection system (IDS) assists in identifying system attacks and detects attackers. In the past, the IDS has been subjected to a variety of machine learning (ML) techniques to improve its accuracy and performance in detecting intruders. The random forest classification method and principal component analysis (PCA) are currently used in an approach to produce effective IDS. Whereas the random forest will aid in classification, the PCA will aid in dataset organization by lowering the dataset's dimensionality [9]. According to the results, the suggested

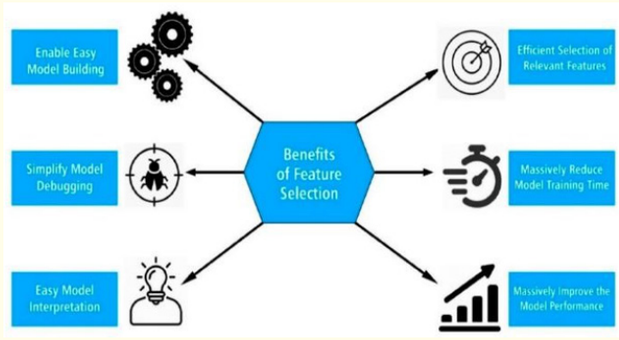

Figure 1: Feature Selection.

strategy outperforms other methods like SVM, Naïve Bayes, and Decision Trees in terms of accuracy.

Problem domain

ML-based IDS can be computationally demanding, requiring significant resources for training and real-time detection. The complex nature of ML algorithms may result in models that are difficult to interpret, hindering a clear understanding of decision-making processes. ML-based IDS systems may be susceptible to adversarial attacks that manipulate input data to deceive the model and evade detection. Implementing ML-based IDS requires expertise in both cybersecurity and machine learning, leading to a potentially high initial setup complexity.

Proposed solution

The proposed system integrates the Random Forest algorithm for intrusion detection within the dynamic and evolving landscape of cyber threats. By first loading relevant data, including the well-established KDD dataset, the system initiates a robust foundation. Subsequent data pre-processing addresses challenges in cyber security data, ensuring the dataset's cleanliness and readiness for analysis. Feature selection categorizes attributes into classes, such as Basic, Content, Traffic, and Host, aiming to improve intrusion detection strategies. The training and testing phases apply the Random Forest algorithm to learn patterns and correlations within the data, subsequently evaluating the model's performance using key metrics like Detection Rate (DR) and False Alarm Rate (FAR). The culmination of these modules results in an Intrusion Detection System (IDS) that leverages machine learning methodologies, specifically the Random Forest algorithm, to effectively enhance cyber security measures against emerging threats in the digital domain.

Methodology

To strengthen network security, the suggested system implements an [4]. Intrusion Detection System (IDS) that uses the Random Forest (RF) method. To implement this system, a variety of

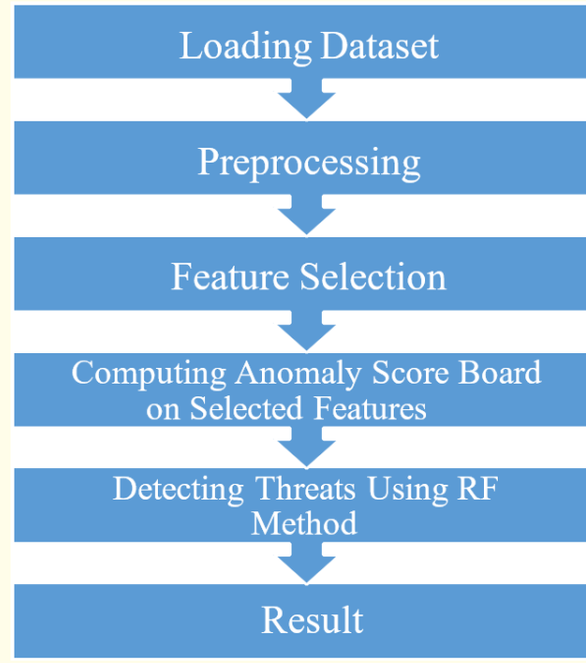

Figure 2: System Architecture.

datasets covering both legitimate and illicit network activity must be acquired. Next, preprocessing operations including feature normalization and management of missing information must be performed. The model's efficiency is increased via [5] feature selection, and a split dataset is used to train a Random Forest classifier [10]. A thorough assessment of the model is conducted with the use of measures including F1 score, accuracy, precision, and recall. Regular updates and a strong framework guarantee the system's adaptability to changing threats, making it a complete proactive network security solution.

Load dataset

Getting the incursion dataset and loading it is the first step. This dataset could include both legitimate and harmful data, as well as information regarding network activity [7]. The intrusion detection model we use is based on the dataset.

Data pre-processing

To make sure the dataset is clean and prepared for analysis, data pre-processing is essential. This might include encoding categorical variables, scaling or normalizing features, and managing missing data. To fully comprehend the properties of the data, you may also need to do some exploratory data analysis.

Training and testing

The dataset should be divided into a training set and a testing set after pre-processing. Your [12] machine learning models are

trained on the training set, and their performance is assessed on the testing set. Typically, 20% is set aside for testing and 80% is used for instruction.

Feature selection

The feature selection [8] module aims to find and select the most relevant characteristics that enhance the performance of the Random Forest (RF)-based intrusion detection system. The procedure entails scrutinizing the dataset to ascertain the importance of every attribute in differentiating between legitimate and malevolent network operations. Feature selection reduces computational complexity, minimizes the chance of overfitting, and concentrates on important features to improve the model's efficiency. To prioritize features for further training, techniques like information gain, recursive feature deletion, or the Random Forest algorithm's feature significance scores might be used.

Predict intrusion with attack type using (RF)

This module uses the Random Forest algorithm to forecast incursions and categorize them according to certain assault types. Using the chosen characteristics, the trained Random Forest model is used to analyze historical network data. The algorithm's capacity to manage both numerical and category variables helps in recognizing different kinds of assaults. The system can discriminate between several assault types and produce precise predictions by using the ensemble learning technique. This module is essential to the Intrusion Detection System's basic operation since it allows the system to classify and identify the types of intrusions that are detected in addition to detecting abnormalities.

Performance and evaluation

To determine how well your intrusion detection system is doing at identifying and categorizing intrusions, you must take the last step in the evaluation process. This thorough evaluation depends on several essential indicators. The model's total correctness of predictions is measured by accuracy, which offers a general indication of the model's performance. Precision measures the percentage of genuine positives among all anticipated positives, examining how accurate forecasts are for certain attack types. Recall, also known as sensitivity, assesses the system's capacity to identify assaults concurrently by calculating the percentage of real positives among all actual positives. The F1 score provides a comprehensive assessment of the model's overall performance by providing a balanced metric that blends recall and accuracy.

Random forest approach algorithm

An ensemble learning technique called Random Forest is well-known for its efficiency in problems involving regression and classification. During training, the algorithm creates many decision trees and merges their predictions to increase overall resilience and accuracy. Every tree starts with a randomly chosen subset of characteristics at each node, and it grows on its own. Each tree is trained on a random subset of the dataset with replacement throughout the training phase, which combines bootstrap sampling with feature randomization to diversify the trees. In regression, the final forecast is based on the average of the individual tree predictions, but in classification, the final prediction is decided by a majority vote of the individual tree predictions.

function Random Forest(X, y, num_trees): for tree in range(num_trees):

subsample_X,subsample_y=random_sample(X, y) tree = build_tree(subsample_X, subsample_y) function build_tree(X, y):

if stopping_condition(X, y):

return create_leaf_node(y) split_feature,split_value = find_best_split(X, y) left_X, left_y, right_X, right_y = split_data(X, y, split_feature, split_value left_subtree = build_tree(left_X, left_y)

right_subtree = build_tree(right_X, right_y) returncreate_decision_node(split_feature, split_value, left_subtree, right_subtree)

Result Analysis

The Modified Random Forest (MRF) technique for empirical analysis of the KDD dataset produces informative results for the Intrusion Detection Systems (IDS) field. The study reveals the unique contributions of each attribute class to the Detection Rate (DR) and False Alarm Rate (FAR) by classifying the dataset into four categories: Basic, Content, Traffic, and Host [6]. Through this detailed analysis, the dataset may be optimized to maximize detection ratio (DR), which measures successful intrusion detection, while decreasing the false alarm rate (FAR) to prevent needless false alerts. The results highlight how crucial attribute class considerations are when creating reliable intrusion detection models and offer insightful information for improving the effectiveness of cyber security measures.

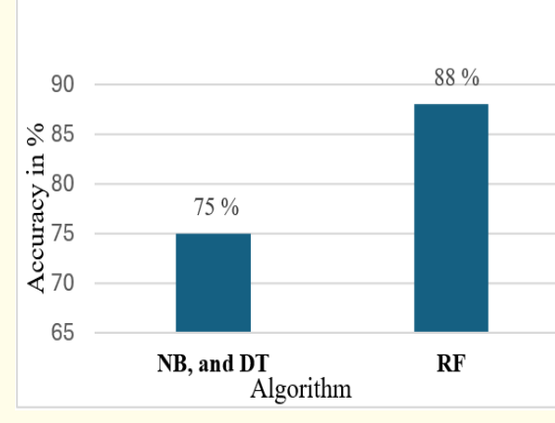

Figure 3: Comparison graph.

In the context of a particular investigation, the table displays the accuracy results of several algorithms, including Naive Bayes (NB), Decision Trees (DT), and the Modified Random Forest (MRF). Interestingly, the combined accuracy of Naive Bayes and Decision Trees is 75%, whereas the Modified Random Forest (MRF) performs substantially better, scoring 88%. These accuracy metrics demonstrate the MRF algorithm's improved performance above its NB and DT counterparts, highlighting its efficacy in the context under analysis. The MRF algorithm is shown in the table as a possible option for the current assignment, highlighting the significance of algorithm selection in reaching improved accuracy rates.

Conclusion

To sum up, adding an intrusion detection system (IDS) based on random forests is a reliable and efficient way to strengthen network security. By use of rigorous feature selection procedures, the system maximizes its capacity to identify relevant features, hence enhancing computational effectiveness and reducing the likelihood of overfitting. The system's adaptability and depth of analysis are enhanced by the Random Forest algorithm's prediction powers, which are especially useful for classifying different sorts of intrusions. The module for Performance and Evaluation guarantees continuous monitoring of the system's precision and flexibility, offering significant perspectives for continuous improvement.

Future Work

The Random Forest-based Intrusion Detection System (IDS) may benefit from further development by using deep learning methods to improve the model's capacity to recognize complex relationships and patterns in network data. Research initiatives can also concentrate on creating more advanced feature selection techniques that are suited to the particular difficulties presented by changing cyber threats. To further strengthen the IDS against changing attack environments, real-time adaptation techniques like online learning and reinforcement learning should be investigated.

Bibliography

1. R Kumar., *et al*. "An intellectual intrusion detection system using hybrid hunger games search and remora optimization algorithm for IoT wireless networks" which was published in the journal Knowledge-Based Systems in November (2022).

2. W Wang., *et al*. "Developed a network intrusion detection system based on representation learning and capturing explicit and implicit feature interactions". Their research was published in the journal Computer Security in January (2022).

3. Satheesh Kumar and M Karthick. "An Secured Data Transmission in MANET Networks with Optimizing Link State Routing Protocol Using ACO-CBRP Protocols". *IEEE Access* (2018).

4. BA Tama and S Lim. "Conducted a systematic mapping study and cross-benchmark evaluation on ensemble learning for intrusion detection systems". Their research was published in the journal Computer Science Review in February (2021).

5. S Lei., *et al*. "Proposed a novel model called HNN for studying intrusion detection based on multifeature correlation and temporal-spatial analysis". Their work was published in the IEEE Transactions on Network Science and Engineering in October (2021).

6. Azath Mubarakali., *et al*. "Optimized flexible network architecture creation against 5G communication-based IoT using information-centric wireless computing". *Wireless Networks* (2023).

7. "Sustainable ensemble learning driving intrusion detection model". IEEE Trans. Dependable Secure Comput 18.4 (2021): 1591-1604.

8. Y Zhou., *et al*. "Building an effective intrusion detection system based on feature selection and ensemble classifier". *Journal of Computer Networks* 174 (2020): 107247.

9. Karthick M., *et al*. "An Efficient Multi-mobile Agent Based Data Aggregation in Wireless Sensor Networks Based on HSSO Route Planning". *Ad Hoc and Sensor Wireless Networks* 57 (2021): 187-207.

10. B A Tama., *et al*. "An enhanced anomaly detection in web traffic using a stack of classifier ensemble". *IEEE Access* 8 (2020): 24120-24134.

11. J Oughton., *et al*. "Explored the comparison between wireless internet connectivity options 5G and Wi-Fi 6". Their findings were published in the journal Telecommunication Policy in June (2021).

12. Y Cheng., *et al*. "Leveraging semisupervised hierarchical stacking temporal convolutional network for anomaly detection in IoT communication". *IEEE Internet Things Journal* 8.1 (2021): 144-155.

13. G Kumar., *et al*. "MLEsIDSs: Machine learning-based ensembles for intrusion detection systems—A review". *Journal of Supercomputer* 76.11 (2021): 8938-8971.