ISSN: 2581-3404 (Online) IF: 5.68 (SJIF)
IJIRTM, Volume-9, Issue-3, June-2025

Rough Set Theory Model-based NB Tree-Based Intrusion Detection Approach

[1]Priyanka Tiwari, [2]Pradeep Pandey
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science & Engineering,
[1] SAM College of Engineering and Technology, Bhopal, India.

Abstract. *Bayesian networks are powerful tools for decision-making and reasoning under uncertainty. A simplified form of these networks, known as Naïve Bayes, is particularly efficient for inference tasks due to its computational simplicity. However, Naïve Bayes relies on a strong assumption of feature independence, which may not always hold in real-world scenarios. This dissertation presents an experimental study on the application of the Naïve Bayes algorithm for intrusion detection, incorporating Rough Set Theory to enhance performance. Despite its simple structure, Naïve Bayes demonstrates competitive results in terms of classification accuracy and F-measure. Furthermore, the study compares the performance of Bayesian networks with the Classification and Regression Tree (CART) model using the Kyoto dataset, highlighting the superior or comparable effectiveness of Bayesian approaches. The dissertation also introduces the fundamental concepts of Intrusion Detection Systems (IDS) and Rough Set Theory. An IDS is a crucial security mechanism designed to monitor network activities and alert administrators to potential malicious behavior. Given the increasing frequency and sophistication of intrusion attempts aimed at compromising organizational data, network security has become a critical area of research. Consequently, enhancing intrusion detection capabilities remains a significant focus in the field of cyber-security.*

Keywords: Tree based classifier Rough set theory, signature-based IDS, anomaly-based IDS.

Introduction

An Intrusion Detection System (IDS) is a critical security mechanism designed to monitor computer network activity and report malicious behavior to the network administrator. Intruders frequently attempt to gain unauthorized access to networks, posing significant risks to organizational data. As a result, data security has become a top priority for all types of organizations, making intrusion detection a prominent research topic in the field of cybersecurity. Intrusion Detection (ID) can be defined as a security management system that protects computer systems and networks. It primarily operates through two main approaches: misuse detection and anomaly detection. Misuse detection identifies intrusions by matching observed activity patterns with known signatures of attacks or vulnerabilities. For example, repeated failed login attempts within a short time frame could be flagged as suspicious and trigger an alert.

However, a major limitation of misuse detection is its inability to detect novel or previously unknown attacks, as it relies solely on predefined patterns. On the other hand, anomaly detection identifies deviations from established user behavior profiles. For instance, if a user's session shows a significantly different command usage density compared to their historical average, an alert may be raised. This approach is effective for identifying unknown or zero-day attacks, as it does not require prior knowledge of specific attack signatures. However, it can generate a higher number of false positives, as any significant deviation—whether malicious or benign may be flagged as suspicious. These two techniques offer complementary capabilities in detecting intrusions. In recent years, various machine learning approaches have been explored to enhance intrusion detection accuracy. Techniques such as regression models, nearest neighbor classifiers, Bayesian probabilistic classifiers, decision trees, inductive rule-learning algorithms, neural networks, and online learning methods have been applied to the problem of intrusion classification, offering promising results in improving IDS performance.

I. Rough set theory (RST)

Rough set theory (RST) has emerged as an intelligent tool for knowledge discovery from imprecise, ambiguous datasets through the identification of adapt (feature subset) which represents the maximum information of the system. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others. The rough set theory is an important mathematical tool to deal with imprecise, unpredictable, incomplete information and knowledge. Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated

II. Intrusion detection system (IDS)

An IDS is software which automates the intrusion detection process. So intrusion detection system can be thought of as a security operation which helps in protection of the system e.g., firewalls. It also helps in providing security and prevention against various intrusions which are caused by the attackers. IDS are usually deployed along with some other preventive security controls mechanism, such as access control and authentication, as a second line of defense which protects the information systems. There are two main techniques of intrusion detection misuse & anomaly detection. Anomaly detection detects unusual activity patters in the observed data. It is based on subject's (e.g. a system or a user) normal behavior. Any significant deviation from the usual activity is considered as intrusive. Misuse detection technique recognizes known attack patterns. It is based on signature of known attack. Any action that matches with the pattern of a known attack is considered as intrusive.

IDS Architecture

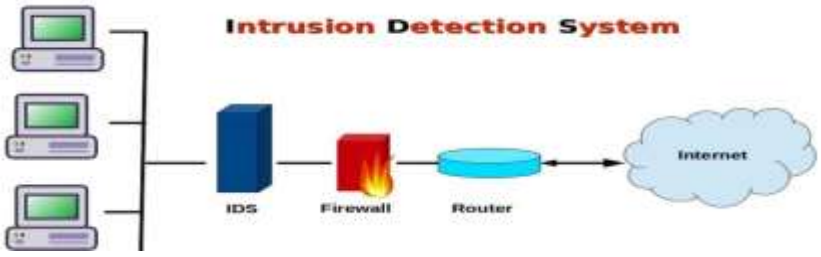

Figure 1: IDS architecture.

III. Neural network

It is one of the most up to date flags preparing innovation. ANN is a versatile, nonlinear framework out how to play out a capacity from information and that versatile stage is regularly preparing stage where framework parameter is change amid operations. After the preparation is finished the parameter are settled. On the off chance that there are loads of information and issue is inadequately reasonable then utilizing ANN model is precise, the non-direct attributes of ANN give it bunches of adaptability to accomplish input yield outline. Fake neural networks, give client the capacities to choose the system topology, execution parameter, learning guideline and ceasing criteria.

IV. NB tree

Bayes networks are one of the most widely used graphical models to represent and handle ambiguous information. Bayes networks are specified by two components:

(i) A graphical component composed of a directed acyclic graph (DAG) where vertices represent event and edges are relations between events.

(ii) A numerical component consisting in an evaluation of different links in the DAG by a conditional probability distribution of each node in the context of its parents. Naive bayes are very simple bayes networks which are possessed of DAGs with only one root node (called parent),representing the unobserved node, and several children, comparable to observed nodes, with the strong assumption of independence among child nodes in the context of their parent.

V. IDS

A number of these aberrations are caused by malicious network attacks like denial-of service or viruses, whereas others are the output of equipment failures and accidental outages [9]. Several strategies have been created by corporations and play very vital roles to protect network infrastructure and communications via the Internet like, through the utilization of firewalls, anti-virus software packages and intrusion detection systems. Present firewalls cannot defend against every category of intrusion, whereby number of intrusions takes advantages of computer system vulnerabilities. Around-the-clock network monitoring is given byan IDS and is also an additional wall to protect the network. The intrusion detection system can also be defined as a process of determining an intrusion into a system through the observation of available information concerning the state of the system, tracking user activities and reporting to a management station.

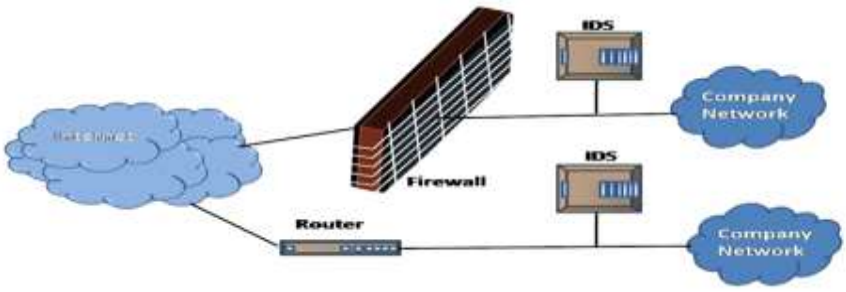

Figure 2: Intrusion detection systems.

Literature Survey

Literature review for different concept and techniques are described below:

Thaseen et al. focused on evaluating different tree-based classification algorithms for classifying network events in IDSs. Their experiments, conducted using the NSL-KDD dataset, involved reducing the dimensionality of the dataset's attributes. The results demonstrated that the Random Tree model achieved the highest classification accuracy and the lowest false alarm rate among the tested models. Additionally, Random Tree was found to outperform other leading intrusion detection models in terms of predictive accuracy.

Ingre et al. developed an IDS based on decision tree models using the NSL-KDD dataset. They proposed four models: two for five-class classification (normal and various attack types), and two for binary classification (normal vs. attack). Feature selection techniques were employed in the five-class models. The CART algorithm, using the Gini index as the splitting criterion, was applied for pattern classification. To reduce dimensionality, they used Correlation-Based Feature Selection (CFS), which contributed to improved model performance.

Zhang et al. provided a comprehensive review of the application of Rough Set Theory (RST), which has been extensively studied over the past three decades. RST has found applications in numerous domains including machine learning, knowledge acquisition, decision analysis, expert systems, and granular computing. It has been recognized as a powerful tool for handling uncertainty in data. The authors emphasize that while RST has broad application potential, several challenges remain in applying it to data mining, such as scalability to large datasets, development of efficient reduction algorithms, parallel computation, and hybrid algorithm integration.

Paul et al. addressed the growing volume of attack data and the need for precise automated classifiers to manage attack registries and improve search engine performance. In their approach, every term and tag on an attack page was treated as a feature. They applied CART, enhanced through a Firefly Algorithm (FA), to identify the most relevant features for attack page classification. The J48 classifier in the Weka data mining tool was used for evaluation, with experiments conducted on the Attack KB and Conference datasets. Their results showed that selecting a feature subset using FA preserved classification accuracy while significantly reducing computational time due to the smaller feature space.

Pant et al. highlighted the importance of feature selection in building effective intrusion detection systems. By reducing the number of input features, feature selection helps avoid overfitting and enhances both efficiency and accuracy. The authors emphasized that Rough Set Theory is a simple yet powerful approach for feature selection, utilizing the concept of reducts to identify the most informative attributes. Their findings suggest that rough set-based feature selection can effectively detect various types of attacks with high accuracy.

Srivastav, Shukla, Kumar, and Muhuri (The authors emphasized the importance of securing Industry 4.0 infrastructures, which rely heavily on interconnected cyber-physical systems and IoT devices, making them vulnerable to sophisticated cyber threats. HYRIDE leverages advanced machine learning algorithms alongside heuristic analysis to identify both known and unknown attacks in real time. Experimental results demonstrated HYRIDE’s superior performance in detecting a wide variety of attacks while maintaining resilience against evasion tactics commonly seen in industrial networks. This hybrid model showcases the potential for comprehensive cybersecurity solutions tailored to the complexity of modern smart manufacturing systems.

Problem Statement

Review several techniques have been proposed to improve the efficiency of classification. User required searching his desire CART techniques through the search engines. There is still a problem of mapping and detection of the tree, to work effectively. Some time they give good results whereas most of the time the result is anonymous because it is required to work on some of the parameters like the accuracy and F-measure. There is still a scope to improve the accuracy and F-measure.

Objective

The main aim of the dissertation work is to improve the performance of intrusion detection is to identify entities attempting to subvert in-place security controls. Various works has been done under this a device or software which monitors network traffic and suspicious activity, if any deviation occurs against normal behavior, then it alerts the system or network administrator, as on the following parameters

1. Accuracy
2. F-Measure.

Result Analysis

Experiment to calculate accuracy between CART and NB-IDS

This experiment is conducted in order to determine the accuracy of the developed IDS with naïve bayes intrusion detection system (NB-IDS) for rough set theory and then comparisons are drawn with CART algorithm for the same test condition. In the experiments both the CART and NB-IDS are trained using Kyoto dataset. This experiment is carried to evaluate accuracy of the CART and NB-IDS. Experiments are performed with variation in dataset from 10% to 50%. Increasing value of dataset indicates how much value of self-data is available for training

Table 1: Accuracy of CART and NB-IDS.

Dataset%	CART	NB-IDS
10	0.87	0.93
20	0.88	0.94
30	0.88	0.95
40	0.89	0.95
50	0.9	0.96

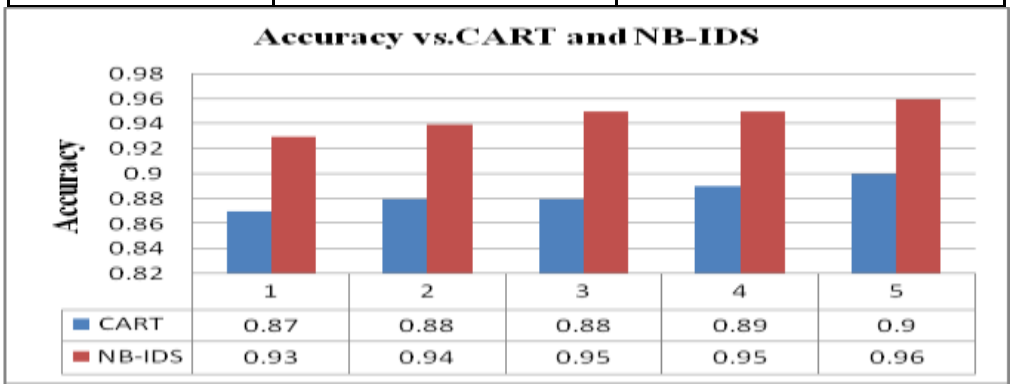

Figure 3: Accuracy comparisons between CART and NB-IDS.

Experiment to calculate F-measure between CART and NB-IDS

This experiment is carried to evaluate F-measure of the CART and NB-IDS. Experiments are performed with variation in dataset from 10% to 50%. Experiments are carried with increasing value of dataset that indicates how much value of self-data is available for training

Table 2: F-measure of CART and NB-IDS.

Dataset%	CART	NB-IDS
10	0.89	0.93
20	0.86	0.94
30	0.84	0.95
40	0.8	0.95
50	0.9	0.96

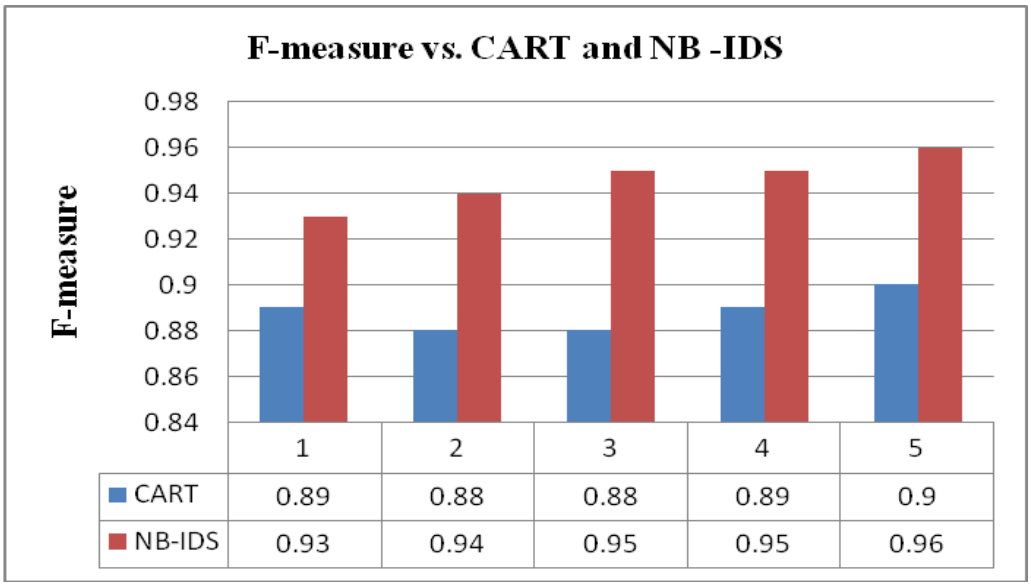

Figure 4: F-measure comparison between CART and NB-IDS.

Conclusion

This dissertation presents the implementation of a Naïve Bayes-based Intrusion Detection System (NB-IDS) within the framework of Rough Set Theory and provides a detailed comparison with existing techniques. The proposed design, developed using the Rough Set Theory model, demonstrates notable improvements in both effectiveness and efficiency compared to prior approaches. Performance evaluation of the proposed system is conducted by comparing it with other established methods, including Self-Organizing Maps (SOM), Classification and Regression Trees (CART), and traditional Naïve Bayes classifiers. These comparisons are based on results reported in the literature.

1. The comparative analysis focuses primarily on accuracy in detecting attacks using the Kyoto dataset as a benchmark.
2. An Intrusion Detection System (IDS) serves as a vital security tool for monitoring inbound and outbound network traffic, aiming to detect and prevent malicious activities such as unauthorized access, misuse, and hacker attacks. While traditional security measures—such as user authentication, encryption, and firewalls—form the initial layer of defense, they do not provide complete assurance of system security. IDS complement these measures by offering a second line of protection.
3. This dissertation highlights the application of Rough Set Theory in conjunction with tree-based classifiers, showing that the proposed approach significantly outperforms conventional methods in terms of detection accuracy.
4. A thorough analysis of the system's performance is provided in the respective section of the dissertation, illustrating the efficiency of the proposed model over existing approaches.

The results clearly indicate that the proposed method delivers superior performance, particularly with respect to accuracy and F-measure, making it a more reliable solution for intrusion detection.

Future scope

This dissertation presents a new intrusion detection model which is the combination of rough set based feature selecting algorithm and NB tree. It provides one of the benefits with regard to the classification itself, to achieve good accuracy regression is used, but as the size of the dataset (either the number of examples or the number of features) grows its execution time becomes expensive which is one of the limitations of this work

(i) The major advantages of the NB-IDS, future work are effective feature selection to reduce the number of decision attributes and size of log data and better classification to enhance the detection accuracy.

(ii) Future works in this direction can be the use of intelligent agents for effective decision making in classification.

(iii) The learning and classifying with naive bayes is generally faster than learning and classifying with decision trees.

References

1. Saurabh, P., Verma,B, Sharma,S.: An Immunity Inspired Anomaly Detection System: A General Framework A General Framework, Proceedings of 7th International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), Springer, pp 417-428, (2012).
2. Saurabh, P., Verma,B.: An Efficient Proactive Artificial Immune System based Anomaly Detection and Prevention System, Expert Systems With Applications, Elsevier, 60, pp 311–320, (2016).
3. Sengupta,N., Sen,J., Sil,J.:Moumita Saha Designing of on line intrusion detection system using rough set theory and Q-learning algorithm,pp.161-168, (2013).
4. Shanmugavadivu, R., Nagarajan, N.: Network intrusion detection system using fuzzy logic, Indian Journal Computer. Science. Engineering,pp.101–111,(2011)
5. Sharma, B., Sharma, L., Lal, C. & Roy, S. Anomaly-based network intrusion detection for IoT attacks using deep learning technique. Comput. Electr. Eng. 107, 108626 (2023).
6. Shen,C.,Chen,Z., Xue,Y.: "Security system construction of land and resources network based on

intrusion detection", 8th International Conference on Biomedical Engineering and Informatics, pp. 795-799 (2015).
7. Srivastav, S., Shukla, A. K., Kumar, S. & Muhuri, P. K. HYRIDE: HYbrid and Robust Intrusion DEtection approach for enhancing cybersecurity in Industry 4.0. Internet Things 30, 101492 (2025).
8. Subramanian,S., Srinivasan,V.B., and Ramasa,C.: Study on Classification Algorithms for Network Intrusion Systems, pp. 1242-1246,(2012).
9. Thaseen,S., Kumar,A.,: "An analysis of supervised tree based classifiers for intrusion detection system", Pattern Recognition, Informatics and Mobile Engineering (PRIME), (2013).
10. Wattanapongsakorn,N., Sangkatsanee,p.,Srakaew,s. and Charnsripinyo,c.: In Classifying Network Attack Types with Machine Learning Approach,pp.c1-c4,(2011).
11. Zanero,S.: In " Network Intrusion Detection System". Proceedings of the 4th annual workshop on Cyber security and information intelligence research, p.17, (2008).
12. Zekri, M., Meslati, L.S.: Immunological approach for intrusion detection, ARIMA Journal, pp.221–240 (2014).
13. Zhang, Q., Xie,Q., Wang,G.: "A survey on rough set theory and its applications", CAAI transactions on intelligence technology,pp.323-333,(2016).
14. Zhong, Y., Wang, Z., Shi, X., Yang, J. & Li, K. RFG-HELAD: A robust fine-grained network traffic anomaly detection model based on heteorgeneous ensemble learning. IEEE Trans. Inf. Forensics Secur. 19, 5895 (2024).