International Journal of Scientific Methods in Computational Science and Engineering 1(1):9-16

Optimizing a Feature Selection Intrusion Detection Algorithm with Data Mining

E.V.N.Jyothi [1*], M.Kranthi[2] , S.Sailaja[3]

[1,2,3]Department of Computer Science and Engineering, RISE Krishna Sai Prakasam Group of Institutions, Ongole, Andhra Pradesh, India

Received: 01 June 2024 • **Accepted:** 02 June 2024 • **Published Online:** 03 June 2024

Abstract: Intrusion detection safeguards computer systems against unauthorized access and malicious activities. Feature selection plays a pivotal role in enhancing the efficiency and effectiveness of intrusion detection algorithms by identifying the most relevant features from vast datasets. In this study, we propose a novel approach to optimize feature selection in intrusion detection algorithms using data mining techniques. We explore various data mining algorithms, including decision trees, genetic algorithms, and particle swarm optimization, to identify the optimal feature subset that maximizes detection accuracy while minimizing computational overhead. Experimental results demonstrate our approach's efficacy in improving intrusion detection systems' performance across different datasets, achieving higher detection rates with reduced computational complexity. Our work advances state-of-the-art intrusion detection by leveraging data mining for efficient feature selection.

Key words: Intrusion detection, Feature selection, Data mining, Optimization, Machine learning, Cybersecurity.

1. Introduction

In today's interconnected digital landscape, computer systems and networks' security is paramount. With the proliferation of cyber threats, ranging from malware attacks to data breaches, organizations face significant challenges in protecting their sensitive information and ensuring the integrity of their operations. Intrusion detection systems (IDS) serve as a frontline defense mecha-

*Correspondence: Associate Professor, Department of Computer Science and Engineering, RISE Krishna Sai Prakasam Group of Institutions, Ongole, Andhra Pradesh, India. Email:jyothiendluri@gmail.com
https://doi.org/10.58599/IJSMCSE.2024.1106
Vol. 1, No. 1, June 2024, pp:9-16

nism, continuously monitoring network traffic and system activities to identify and thwart malicious activities in real time. However, the effectiveness of IDS relies heavily on the accuracy of the features used to characterize normal and anomalous behavior. Traditional intrusion detection approaches often suffer from the curse of dimensionality, where the high number of features in the dataset leads to increased computational complexity and reduced detection accuracy. Feature selection techniques aim to alleviate this problem by identifying a subset of relevant features that capture the essential characteristics of normal and abnormal network behavior. By focusing on the most informative features, IDS can achieve higher detection rates while minimizing false alarms and computational overhead. Data mining, a multidisciplinary field at the intersection of statistics, machine learning, and database systems, offers a rich set of tools and techniques for analyzing large datasets and extracting actionable insights. However, in the context of intrusion detection, the selection of an appropriate data mining algorithm and the optimization of feature selection parameters pose significant challenges. These challenges require careful consideration of factors such as detection accuracy, computational efficiency, and robustness to different types of attacks. Our proposed approach aims to address these challenges and enhance the performance and scalability of IDS. In this paper, we present a novel approach that optimizes feature selection in intrusion detection algorithms using data mining techniques. Our approach is unique in its systematic exploration of various data mining algorithms and evaluation of their effectiveness in identifying the most discriminative features for intrusion detection. By leveraging the synergy between data mining and cybersecurity, our approach aims to enhance the resilience of IDS against emerging threats and improve the overall security posture of organizations operating in today's digital environment.

2. Related Works

When the font name says that you should use Times Roman or Times New Roman, you can utilize one of those fonts. Please use the typeface that most closely resembles Times if neither font is available on your word processor. Thank you for your cooperation. Bit-mapped typefaces should be avoided at all costs. You have to utilize fonts that are Open Type or True-Type 1. For mathematics and other subjects of a similar nature, symbol typefaces should also be incorporated. These systems are incredibly incorrect and inefficient in their operation. A few of the most common threats can affect network resources [1]. These threats include Denial of Service, remote-to-local (R2L), Probe, and user to root attack (U2R). Intrusion detection systems are confronted with several issues, two of the most significant of which are the ability to reliably detect malicious behavior and the ability to keep up with the growing volumes of business traffic [2].They comprehensively explain the process of developing an effective intrusion detection system by utilizing powerful algorithms [3]. Find suspicious activity in the traffic on the network and stop unauthorized users from accessing the resources on the network from being accessed. Although several interruption identification

frameworks have been developed in the past, the current network intrusion detection model has limitations in terms of accuracy and detection time [4].

They have described it. Intrusion detection systems prevent unauthorized access to computer systems, networks, and databases. Harsh Set Trait Decrease Calculation is one of the fundamental ideas utilized to minimize attributes successfully by removing redundancy. They describe the process that should be followed to select the fewest feasible Knowledge Disscovery in Databases (KDD) attributes [5]. The minimal redundancy-maximal-relevance (mRMR) measure and the correlationfeature-selection measure are two fundamental estimations that are taken into consideration in the channel model. Both of these measures are stated as being checked from top to bottom. By weaving the metrics together, we demonstrated that they can be combined into a standard element determination (GeFS) measure [6]. A branch-and-bound algorithm solution to a mixed 0-1 linear programming problem (M01LP) inspired the unique method. They have described it. An algorithm known as an ICRF-based Cuttlefish Feature Selection Algorithm (ICRFCFA) is recommended for effective decision-making on healthcare datasets [7]. It is helpful to incorporate the highlighted determination computation proposed to speed up the anticipated improvement in accuracy [8]. Future work in this field may consist of the creation of innovative criteria for the selection of efficient features [9].

3. Methodology

This specific module consists of an element choice agent and a data set. The data utilized for the KDD Cup'99 was collected via the feature selection module. This set contains 41 features. Using the criteria recorded in the knowledge base, the agent in charge of feature selection chooses the best combination of these 41 features in Figure 1. The word "knowledge base" comes from the fact that it is a compilation of data that includes the features' attributes. Another helpful feature is incorporating the CRF model's freshly created rules into the training phase [10, 11]. The feature selection agent looks up the appropriate characteristics in the knowledge base to determine which laws are most relevant. Users can choose the KDD'99 cup set that comprises the ideal combination of features, which is the most significant benefit of this feature selection module. It accomplishes this by classifying the data according to a predetermined set of rules. The layered approach (LA) is used to construct these criteria all through the course. Identifying common data and attacks is possible by sticking closely to the standards [12, 13]. Also, the requirements help categorize the many kinds of assaults found during testing. This module mainly consists of the training agent and the decision-making agent. Investigative, denial-of-service, R2L, and U2R attack layer framing is the responsibility of the training agent. The agent making decisions can use information and regulations to conclude. This module can only generate two possible outcomes: an average outcome

or an attack. A Probe attack, DoS attack, R2L attack, or U2R attack are all forms of assault. Any one of these kinds of attacks could be used [14]. The dataset used by this agent for training is based on LA and has fewer features than other datasets. The training agent is also responsible for establishing the classification criteria for the knowledge base[15]. The LA, through the use of a four-layer technique, identifies four distinct types of attacks[16].

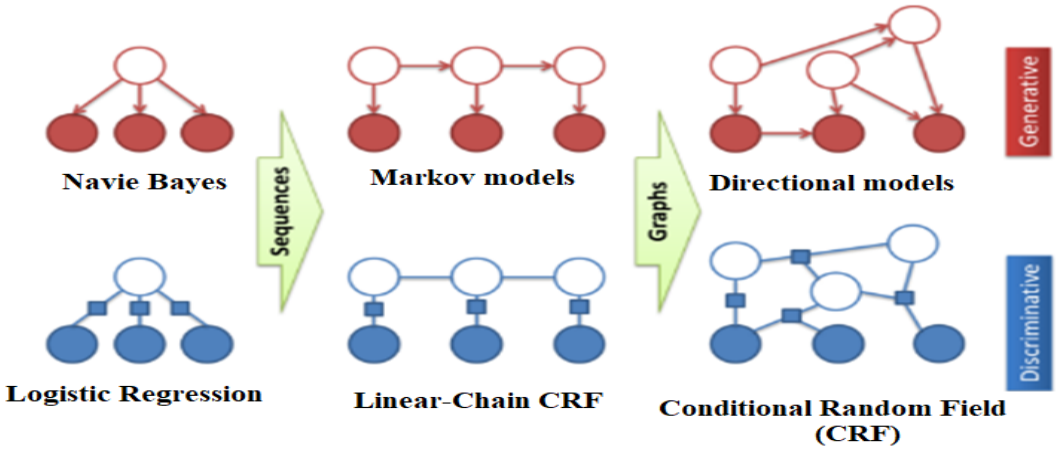

Figure 1. A visual illustration of CRF

Method for Sorting Data Applying LA: Within the scope of this investigation, we utilized the widely-utilized LA classifier in conjunction with the ICRFFSA feature-selecting technique to accomplish accurate classification. During the training phase, the data that has been reduced in characteristics is inputted into the suggested algorithm. Once the computation for element determination has been completed, it is validated by using the rules and facts already present in the knowledge base. This model defines four distinct types of attacks, each derived from the concepts discovered in the knowledge base. When the individuals responsible for the assaults have been identified, this classifier will then determine the nature of the attacks.

4. Results and Discussion

We chose different functions for the other layers in reaction to the type of attack identified by the training layer. This is done about the four assault groups (Probe, DoS, R2L, and U2R) provided by the KDD 99 dataset and alternative attacks. One dedicated module is assigned to each of the four assault groups, and a fifth module is trained on additional assaults that were not included in the four assault groups that were included in the training dataset. To accomplish the goal of teaching the various levels of the framework, we choose different functions in Table 1. Accordingly, we use

Table 1. Intrusion attack performance on the training set

Attack	Precision Rate	Recall Rate
R2L	88.7434	88. 7434
DOS	82. 7434	80. 7434
Normal	86. 7434	85. 7434
U2R	87. 7434	86. 7434
Probe	86. 7434	83. 7434

our domain expertise to select the suitable function for each attack. Afterward, we will dissect the logic underpinning the process of selecting features for each level of the layered architecture.

This type of attack occurs when a malicious actor blocks authorized users from accessing the system or when they make a computer or memory resource too busy to handle the requests they have made. This is Gram. Neptune, Earth, Smurfs, Teardrops, and Pods are all involved in Figure 2 and Figure 3.

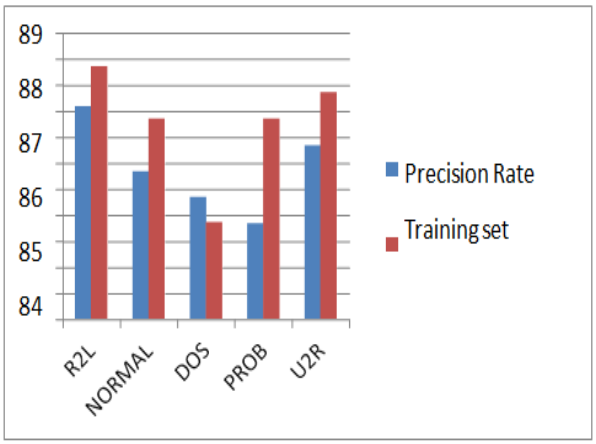

Figure 2. Precision Rate

In a reconnaissance assault, an attacker without a record on the remote computer uses network packets to acquire proximate access as a client of that system by exploiting specific vulnerabilities. Examples include spy, warezclient, warezmaster, write to FTP, and guess password. The choice between prioritizing precision or recall depends on the specific use case and the environment in which the intrusion detection system is deployed. For instance, in a highly secure environment, high recall might be prioritized to ensure no threats go undetected, while in a more resource-constrained environment, high precision might be more important to reduce the burden on security personnel.

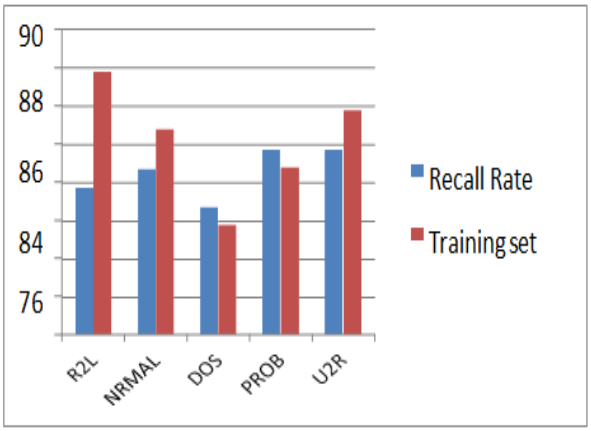

Figure 3. Recall Rate

5. Conclusion

A new interruption location framework was developed to facilitate the creation of intrusion detection systems that are both more effective and produce more precise results. To accomplish this objective, we have created the LAICRF model, a classification method that combines ICRFFSA and LA. This model is designed to detect intrusions efficiently. In this study, rule-based and LA-based classification algorithms are utilized. These approaches significantly reduce the time required for detection while enhancing the accuracy of detection. A strategy for identifying interruptions in the system: This article aims to identify new types of cyberattacks, as indicated in the introduction. In addition, we propose and carry out an additional computation for the steady component choice to successfully select the elements. For inclusion determination, the suggested technique is to combine the calculation of the Cuttlefish Component Choice with the Extended Chi-square algorithm. The findings of the tests indicate that the proposed system is feasible since it successfully detected all types of attacks.

References

[1] Skhumbuzo Zwane, Paul Tarwireyi, and Matthew Adigun. Performance analysis of machine learning classifiers for intrusion detection. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–5. IEEE, 2018.

[2] Shengyi Pan, Thomas Morris, and Uttam Adhikari. Developing a hybrid intrusion detection system using data mining for power systems. *IEEE Transactions on Smart Grid*, 6(6):3104–3113, 2015.

[3] Inadyuti Dutt, Samarjeet Borah, and Indrakanta Maitra. A proposed machine learning based scheme for intrusion detection. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 479–483. IEEE, 2018.

[4] Komal Rasane, Laxmi Bewoor, and Vishal Meshram. A comparative analysis of intrusion detection techniques: Machine learning approach. In *Proceedings of International Conference on Communication and Information Processing (ICCIP)*, 2019.

[5] Pedro Garcia-Teodoro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2): 18–28, 2009.

[6] Venkata Krishna Chaithanya Manam. *Efficient disambiguation of task instructions in crowdsourcing.* PhD thesis, Purdue University Graduate School, 2023.

[7] Hongzhu Tao, Jieying Zhou, and Sen Liu. A survey of network security situation awareness in power monitoring system. In *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pages 1–3. IEEE, 2017.

[8] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, and Pang-Ning Tan. Data mining for network intrusion detection. In *Proc. NSF Workshop on Next Generation Data Mining*, pages 21–30. Citeseer, 2002.

[9] VK Chaithanya Manam, Joseph Divyan Thomas, and Alexander J Quinn. Tasklint: Automated detection of ambiguities in task instructions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 160–172, 2022.

[10] Sugwon Hong, Jae-Myeong Lee, Mustafa Altaha, and Muhammad Aslam. Security monitoring and network management for the power control network. *system*, 2:3, 2020.

[11] V Suresh Kumar, Sanjeev Kulkarni, Naveen Mukkapati, Abhinav Singhal, Mohit Tiwari, and D Stalin David. Investigation on constraints and recommended context aware elicitation for iot runtime workflow. *International Journal of Intelligent Systems and Applications in Engineering*, 12(3s):96–105, 2024.

[12] Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1):e4150, 2021.

[13] Sanjeev Kulkarni, Sachidanand S Joshi, AM Sankpal, and RR Mudholkar. Link stability based multipath video transmission over manet. *International Journal of Distributed and Parallel Systems*, 3(2):133, 2012.

[14] V K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. Taskmate: A mechanism to improve the quality of instructions in crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1121–1130, 2019.

[15] Fadi Salo, Mohammadnoor Injadat, Ali Bou Nassif, Abdallah Shami, and Aleksander Essex. Data mining techniques in intrusion detection systems: A systematic literature review. *IEEE Access*, 6: 56046–56058, 2018.

[16] Sanjeev Kulkarni, Aishwarya Shetty, Mimitha Shetty, HS Archana, and B Swathi. Gas spilling recognition and prevention using iot with alert system to improve the quality service. *Perspectives in Communication, Embedded-systems and Signal-processing-PiCES*, 4(4):34–38, 2020.