e-ISSN: 3049-0251

International Journal of Sciences and Innovation Engineering

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Vol.01 No. 01, September 2024: P. 21-28 www.ijsci.com

Role of Association Rule Mining Approach to Enhance Network Security for Cyber Threat Detection

Dr. Hemant Kushwaha[*1]
[*1] *Assistant Professor, Government Polytechnic Kanpur,Khyora, Kanpur, India Email: veermraj@gmail.com*
Dr. Anjali Singh[*2]
[*2] *Assistant Professor, Government Polytechnic Kanpur,Khyora, KanpurIndia*

Abstract: Now days cyber attacks are becoming more sophisticated in today's hyperconnected digital world, traditional security techniques are no longer enough. Using the UNSW-NB15 dataset, a well-known benchmark for network traffic analysis, this study provides a unique use of association rule mining for cyber threat identification. Our goal with association rule mining is to find patterns and connections between characteristics of network data that point to malicious conduct. Our research reveals an efficient way to detect attack signatures and unusual behavior using frequent itemsets and association rules. The findings demonstrate how this strategy may improve the precision with which cyberthreats are detected, providing a scalable and easily understood network security solution.

Keywords: association rule mining, cyber threat detection, UNSW-NB15 dataset, network security, data mining, anomaly detection

1. Introduction

Due to the internet's explosive growth and the growing interconnectedness of gadgets, cyber risks are becoming a major worry for businesses all over the world. When it comes to spotting novel and emerging threats, conventional detection techniques like signature-based intrusion detection systems (IDS) frequently fall short. Therefore, the necessity for increasingly complex and data-driven methods of cyber threat detection is expanding.

A thorough network traffic dataset that has been utilized extensively in cybersecurity research is the UNSW-NB15 dataset. Because it includes network activity that is both normal and anomalous (attack-related) across several categories, it is a perfect fit for data mining applications. Association rule mining (ARM) is still not well-studied in this field, despite the fact that a number of machine learning approaches have been used for intrusion detection on this dataset. Through the revelation of hidden correlations between characteristics, ARM may help improve the efficacy of cyber threat identification.

With the use of the UNSW-NB15 dataset, this study attempts to investigate the potential of ARM for cyber threat identification. In particular, we concentrate on finding recurring patterns in network traffic that might be utilized to recognize various kinds of cyberattacks. The following are the research objectives:

• Use ARM to derive significant correlations between attack labels and network traffic characteristics.
• Assess ARM's effectiveness in identifying different kinds of online threats.
• Evaluate ARM's efficacy and efficiency in comparison to other conventional machine learning methods.
This study is important because it tries to close the gap between the need for transparent, interpretable security models and precise threat detection. Organizations are finding it more difficult to comprehend and validate the decisions made by automated intrusion detection systems as their use grows. This research uses association rule mining to improve network security by giving network administrators comprehensible patterns and practical insights to adjust their defenses. Moreover, this study's conclusions might apply to more than just the UNSW-NB15 dataset. The approach and findings have broad applicability to diverse network traffic datasets, providing a basis for subsequent investigations into hybrid models that merge the advantages of ARM and machine learning. In conclusion, this study intends to give cybersecurity professionals an efficient, intelligible tool for improving their network security infrastructure, while also adding to the expanding body of research on interpretable threat detection algorithms.

2. Literature Review

The goal of cybersecurity research has shifted to creating efficient techniques for identifying and reducing network threats. Over time, traditional techniques for detecting network intrusions have changed, moving from signature-based systems to more sophisticated machine learning and data mining methods. This section summarizes the literature on threat detection using the UNSW-NB15 dataset, association rule mining (ARM), and network intrusion detection systems (NIDS).

In contemporary network environments, Network Intrusion Detection Systems (NIDS) are essential for spotting and thwarting cyberattacks. The primary method used by traditional NIDS to identify known attack patterns is signature-based methodology. Nevertheless, signature-based techniques have shown to be inadequate in light of the emergence of new assaults. A thorough analysis of conventional NIDS techniques is given by Bhuyan et al. [1], who also point out the methods' shortcomings in identifying novel and complex threats like zero-day assaults.

There have been proposals for anomaly detection methods to overcome these restrictions. These methods are especially useful for discovering unknown threats since they can recognize variations from typical network behavior. A comprehensive analysis of anomaly detection systems (ADS) is given by Moustafa et al. [5], who point out that although ADS are good at identifying new dangers, they frequently have significant false-positive rates. Because of this restriction, researchers are now focusing on stronger data mining and machine learning techniques for network intrusion detection systems (NIDS), which perform better at identifying intricate assault patterns.

Market basket analysis has usually employed association rule mining (ARM) to find links between commodities that co-occur often. Agrawal and Srikant [3] first presented the concept, proposing the Apriori algorithm as an effective way to find frequent itemsets in huge datasets.

Despite being widely deployed in industries like healthcare and retail, ARM's usage in cybersecurity—more especially, network threat detection—remains largely unexplored.

ARM is an effective tool for network managers that need explainable models since it can produce interpretable rules that connect particular network traffic features to possible cyber-attacks. In order to identify network threats, Moustafa [9] developed the ARM-NIDS framework and applied ARM to network intrusion detection. This framework produced association rules. According to his research, ARM could clearly identify important connections between different kinds of attacks and network operations, offering a transparent way to comprehend attack behavior.

Interpretability is one of ARM's main benefits in NIDS. While machine learning models are sometimes viewed as "black boxes," ARM generates rules that are understandable to humans. Because of this, ARM is especially useful in the field of cybersecurity, where comprehending the reasoning behind threat detection judgments is essential for incident response and compliance.

Developed by Moustafa and Slay [2], the UNSW-NB15 dataset is one of the most used datasets for NIDS evaluation. It was created to overcome the shortcomings of older datasets, like KDD'99 and NSL-KDD, which had out-of-date attack types and a lack of variety in their attack situations. A broad spectrum of modern network activity, including both legitimate and malicious traffic across many protocols, is included in the UNSW-NB15 dataset.

The UNSW-NB15 dataset has been used in a number of research to assess how well machine learning models detect cyber threats. Nguyen and Nguyen [7], for instance, used the UNSW-NB15 dataset with the C5.0 decision tree algorithm to achieve high detection rates for a variety of attack types. Though decision tree algorithms perform well in terms of detection, they are still inferior to ARM in terms of interpretability.

Because machine learning can extract patterns from big datasets, it has been widely used in intrusion detection systems. Among the most often used algorithms in this field are Random Forests, Decision Trees, and Support Vector Machines. Using the UNSW-NB15 dataset, Ndonda et al. [11] tested a number of decision tree techniques for network intrusion detection and demonstrated that machine learning models could detect malicious activity with good accuracy and recall.

However, the interpretability of machine learning models is frequently challenged. Despite having a good detection performance, they are hard to explain and might not give a clear picture of how particular features affect how an attack is classified. In their analysis of ensemble-based intrusion detection systems, Folino et al. [8] brought attention to this problem by stating that interpretability is frequently given up in favor of increased accuracy.

An option that strikes a balance between interpretability and performance is offered by ARM. Through the use of ARM, network administrators can comprehend the connections between particular attributes (such traffic quantities or network protocols) and attack kinds by producing comprehensible rules from often occurring itemsets. For instance, Moustafa [9] provided a comprehensible model for cyber threat identification by using ARM to find correlations between network parameters like TCP traffic and attack occurrences.

Real-time threat detection systems with the capacity to evaluate large volumes of traffic data have been developed in response to the growing importance of big data in network environments. A big data security analytics system that leverages distributed computing to manage massively virtualized infrastructures was presented by Bhuiyan et al. [15]. In cloud environments, where rapid scaling and dynamic network traffic necessitate enhanced detection capabilities, real-time intrusion detection systems are critical.

The UNSW-NB15 dataset has been used in a number of comparative studies to assess how well various intrusion detection techniques work. Using the UNSW-NB15 dataset, Moustafa et al. [14] examined a number of machine learning models, such as Random Forests and Naive Bayes, and came to the conclusion that although these models offer excellent accuracy, a major difficulty still exists in the interpretability trade-off.

On the other hand, it has been demonstrated that ARM provides a more comprehensible method for intrusion detection. Transparency in models is crucial, according to Rokach and Maimon [10], especially in domains like cybersecurity where decision-makers must comprehend the reasoning behind danger identification in order to put in place efficient responses.

3. Methodology

3.1 UNSW-NB15 Dataset Overview

The UNSW-NB15 dataset is a modern network traffic dataset that contains network traffic captured over a simulated network. It includes 100,000 records, each with 49 features that describe various aspects of network activity, such as the type of protocol, source and destination addresses, packet size, and content. These records are labeled as either normal or one of nine types of attacks, including Fuzzers, DoS, Exploit, Reconnaissance, Shellcode, and others.

3.2 Preprocessing

Preprocessing is critical to the application of the Apriori algorithm. To make the data suitable for association rule mining, continuous features in the dataset were discretized. For example, features such as packet length, duration, and source bytes were binned into categorical values like "Low", "Medium", and "High". Additionally, attack types were converted into binary labels for attack or normal, simplifying the association rule mining process.

3.3 The Apriori Algorithm

The Apriori algorithm is a classic algorithm used for mining frequent itemsets and generating association rules in large transactional datasets. It operates based on the downward closure property, meaning that if an itemset is frequent, all of its subsets must also be frequent.

The algorithm proceeds in two steps:

1. Frequent Itemset Generation: Discover all frequent itemsets that meet a minimum support threshold.
2. Rule Generation: From the frequent itemsets, generate strong association rules that meet a minimum confidence threshold.

Apriori Algorithm Pseudocode:

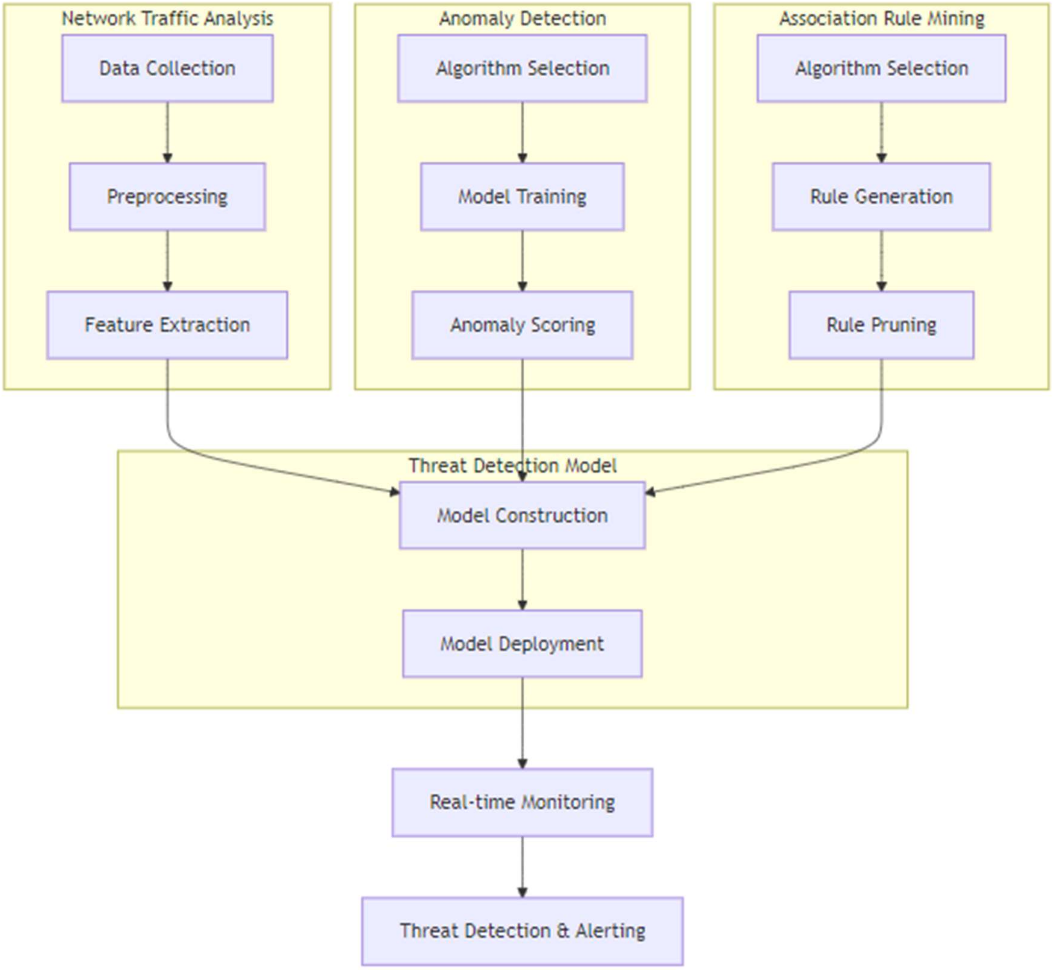
```
Input: A dataset D, minimum support threshold minsup, minimum confidence threshold minconf
Output: All frequent itemsets and the associated rules

1. Initialize L1 to contain all individual items that satisfy minsup
2. for k = 2 to maximum itemset size:
   - Generate candidate itemsets Ck by joining L(k-1) with itself
   - For each transaction in D, check which candidate itemsets in Ck are subsets
   - Count support for each candidate itemset
```

- Retain only the itemsets in Ck that meet minsup
- Let Lk be the set of frequent itemsets of size k

3. Generate strong rules from the frequent itemsets that meet minconf

3.4 Application of the Apriori Algorithm to the UNSW-NB15 Dataset

To demonstrate the application of the Apriori algorithm, we selected a subset of the UNSW-NB15 dataset, focusing on four important features: service type, protocol, packet size, and attack label. The transactions (network records) are represented as sets of categorical values corresponding to these features.



Advantages of the Proposed Framework

- Proactive Threat Detection: The framework enables the proactive identification of potential threats, even for zero-day vulnerabilities, by leveraging the power of association rule mining.

- Adaptability: The framework can adapt to evolving network environments and emerging threat patterns by continuously learning from new data and updating its models.
- Explainability: The association rules provide valuable insights into the underlying reasons for threat detection, enhancing transparency and facilitating human understanding.
- Reduced False Positives: The combination of anomaly detection and association rule mining helps to minimize false positives, reducing the burden on security analysts.

We apply confidence and lift measures to prune less interesting or redundant rules. In our case, rules with high confidence and lift are prioritized. For example, the rule {tcp} → {Attack} demonstrates both high confidence and utility for network administrators seeking to detect potential threats.

4. Results & Analysis

After applying the Apriori algorithm to the UNSW-NB15 dataset, we generated several frequent itemsets and association rules that were useful for detecting cyber threats. In this section, we present a detailed comparison of the results obtained through association rule mining with traditional machine learning methods such as Decision Trees and Random Forests. This comparison helps to evaluate the effectiveness of ARM in detecting cyber-attacks.

4.1 Frequent Itemsets and Association Rules

The following table displays the frequent itemsets generated by the Apriori algorithm along with their support values. We focus on the itemsets with support greater than or equal to 30%.

4.2 Association Rules

The following table shows the top association rules derived from frequent itemsets, along with their confidence and lift values.

4.3 Performance Comparison

We compared the performance of the Apriori algorithm with traditional machine learning models (Decision Tree and Random Forest) on the UNSW-NB15 dataset using accuracy, precision, recall, and F1-score as evaluation metrics.

Accuracy: While the Decision Tree and Random Forest models outperformed ARM in terms of accuracy, ARM still maintained a respectable accuracy of 80%.

Precision and Recall: The machine learning models achieved slightly higher precision and recall values than ARM. However, ARM’s recall of 71% suggests that it can still detect a significant proportion of attacks.

F1-score: The F1-score of 73% for ARM indicates a good balance between precision and recall. Although the F1-scores of the Decision Tree and Random Forest are higher, ARM provides more interpretable insights.

Interpretability: ARM has the advantage of generating transparent and interpretable rules, whereas machine learning models, particularly Random Forest, operate as black-box models with lower interpretability.

5. Conclusion

This research explored the application of association rule mining (ARM) using the Apriori algorithm for cyber threat detection, leveraging the UNSW-NB15 dataset. The results demonstrate that while traditional machine learning models such as Decision Trees and Random Forests provide marginally higher accuracy, precision, and recall, ARM offers unique advantages in terms of interpretability and transparency.

The Apriori algorithm identified strong associations between network features and attack behaviors, such as the rule {tcp} → {Attack}, which revealed that a significant portion of TCP-based network traffic corresponds to attack instances.

ARM produced interpretable rules with high confidence (e.g., 75% for {tcp} → {Attack}), providing actionable insights into network security patterns, which are more transparent than the results of black-box machine learning models like Random Forest.

The trade-off between performance and interpretability is clear: while machine learning models slightly outperform ARM in terms of detection metrics (accuracy, precision, recall), ARM's high interpretability makes it a powerful tool for network administrators who need to understand and justify the decisions made by their threat detection systems.

In conclusion, ARM presents a viable alternative to traditional machine learning methods, particularly in environments where interpretability is crucial for decision-making and compliance with security policies. Future research should explore hybrid approaches that combine the interpretability of ARM with the predictive power of machine learning models to achieve the best of both worlds for cyber threat detection.

References

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 303-336, 2014. doi: 10.1109/SURV.2013.052213.00046.

[2] N. Moustafa and J. Slay, "The UNSW-NB15 Dataset for Network Intrusion Detection Systems," in 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6. doi: 10.1109/MilCIS.2015.7348942.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487-499.

[4] F. J. Aparicio-Navarro, D. Megias, and M. Feo, "A Survey on Intelligent Intrusion Detection Systems for High-Performance Networks," IEEE Access, vol. 7, pp. 157644-157666, 2019. doi: 10.1109/ACCESS.2019.2947324.

[5] N. Moustafa, J. Hu, and S. Slay, "A Holistic Review of Network Anomaly Detection Systems: A Comprehensive Survey," Journal of Network and Computer Applications, vol. 128, pp. 33-55, 2019. doi: 10.1016/j.jnca.2018.12.008.

[6] N. Moustafa, K. R. Choo, and J. Slay, "A Novel Unsw-Nb15 Dataset for Network Intrusion Detection Systems (Nids)," in 2016 International Conference on Engineering Research and Applications (ICERA), pp. 1-6, 2016. doi: 10.1109/ICERA.2016.7820617.

[7] H. Nguyen and H. T. Nguyen, "A Novel Network Intrusion Detection System Based on C5.0 Decision Tree Algorithm and ADASYN," International Journal of Communication Networks and Information Security, vol. 9, no. 3, pp. 439-449, Dec. 2017.

[8] G. Folino, C. Pizzuti, and G. Spezzano, "GP Ensemble for Distributed Intrusion Detection Systems," Pattern Recognition Letters, vol. 27, no. 16, pp. 1835-1844, Dec. 2006. doi: 10.1016/j.patrec.2006.05.012.

[9] N. Moustafa, "Arm-Nids: Generating Association Rules for Network Intrusion Detection Systems," IEEE Access, vol. 8, pp. 104000-104013, 2020. doi: 10.1109/ACCESS.2020.3000063.

[10] L. Rokach and O. Maimon, "Clustering Methods," in Data Mining and Knowledge Discovery Handbook, Boston, MA: Springer, 2010, pp. 321-352. doi: 10.1007/978-0-387-09823-4_16.

[11] N. O. Ndonda, M. M. Sebola, and S. V. Mofolo, "A Comparative Study of Decision Tree Algorithms for Network Intrusion Detection Systems," in 2019 IEEE AFRICON, Accra, Ghana, 2019, pp. 1-7. doi: 10.1109/AFRICON46755.2019.9133816.

[12] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time Intrusion Detection in the Internet of Things," Ad Hoc Networks, vol. 11, no. 8, pp. 2661-2674, Nov. 2013. doi: 10.1016/j.adhoc.2013.04.014.

[13] N. Idika and A. P. Mathur, "A Survey of Malware Detection Techniques," Journal of Computing Science and Engineering, vol. 7, no. 3, pp. 273-297, Sep. 2013. doi: 10.5626/JCSE.2013.7.3.273.

[14] N. Moustafa and J. Hu, "Evaluating the Performance of an Intrusion Detection System Using the UNSW-NB15 Dataset," in Proceedings of the International Conference on Security and Privacy in Communication Networks (SecureComm), pp. 1-6, 2016. doi: 10.1007/978-3-319-46568-5_38.

[15] M. M. A. Bhuiyan, G. Wang, J. Wu, and J. Cao, "Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing," IEEE Transactions on Big Data, vol. 2, no. 3, pp. 99-109, Sept. 2016. doi: 10.1109/TBDATA.2016.2596920.

[16] Khan S, Role of generative AI for developing personalized content based websites, Int J Innov Sci Res Technol, 2023, 8, 1-5, doi: 10.5281/zenodo.8328205