Journal of Artificial Intelligence and Computing Applications (2024) 2(1): 35–39

MAIKRON

Journal of Artificial Intelligence and Computing Applications

Exploration Paper

Clustering-Based Cyber Situational Awareness: A Practical Approach for Masquerade Attack Detection

Nelva N. Almanza-Ortega[1], **Joaquin Perez-Ortega**[1], **Sergio M. Martinez-Monterrubio**[2], **and Juan A. Recio-Garcia**[3,*]

[1]National Technological Institute of Mexico, México
[2]International University of La Rioja, Spain
[3]University Complutense of Madrid, Spain

ABSTRACT

Cyber Situational Awareness (CSA) is crucial for detecting and mitigating security threats in evolving digital environments. Traditional intrusion detection systems face challenges related to computational efficiency, scalability, and interpretability, particularly in the detection of masquerade attacks, where attackers mimic legitimate user behavior. This exploratory study conducts a preliminary investigation into a clustering-based approach that integrates OK-Means, an optimized variant of K-Means, with k-Nearest Neighbors (k-NN) to improve intrusion detection. The proposed approach is evaluated using the Windows-Users and Intruder Simulations Logs (WUIL) dataset to assess its feasibility and preliminary performance. Experimental results suggest that this method can achieve up to 99% recall in masquerade attack detection while reducing execution time by 85% compared to conventional k-NN classifiers. Additionally, the integration of explainability mechanisms, such as clustering visualization and attack introspection tools, provides security analysts with interpretable insights into system decisions. As an initial exploration, this study provides early-stage insights into clustering-based CSA methods and lays the groundwork for future research. The findings suggest that this approach can be further developed and extended to other cybersecurity domains, such as phishing and malware detection, contributing to AI-driven security frameworks.

Keywords: cyber situational awareness (CSA), masquerade attack detection, explainable machine learning

1. Introduction

Cyber Situational Awareness (CSA) plays a crucial role in protecting IT systems against evolving cyber threats. As organizations increasingly rely on digital infrastructure, the need for effective intrusion detection mechanisms has become more critical than ever [1]. Traditional intrusion detection systems (IDSs) often struggle with two significant challenges: the complexity of big data environments and the lack of interpretability in decision-making processes [2]. These challenges are especially pronounced in the detection of masquerade attacks, where malicious actors disguise their activity as that of legitimate users. The dynamic and high-volume nature of cybersecurity logs further complicates the real-time identification of such threats, making conventional predictive models less effective in adaptive environments.

To address these limitations, this work explores a clustering-based approach that enhances CSA by

E-mail address: jareciog@fdi.ucm.es
https://doi.org/10.5281/zenodo.14933955

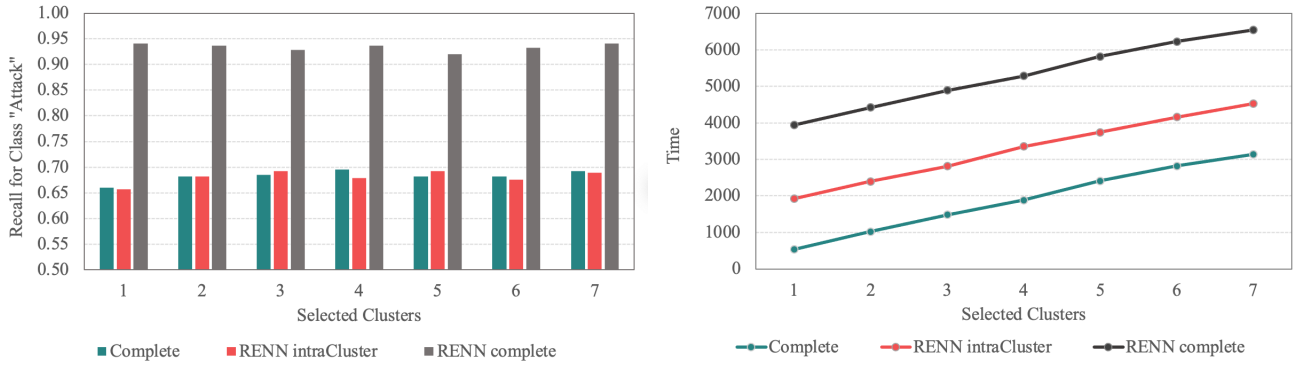

Figure 1. Recall (left) and processing time (right) using different noise reduction approaches (dataset 20%).

improving both efficiency and explainability in masquerade attack detection. Our method combines OK-Means, an optimized variant of K-Means [3], with k-Nearest Neighbors (k-NN) to streamline the classification of potential intrusions. This hybrid approach reduces computational cost while maintaining high detection performance, making it more suitable for real-time threat analysis. Additionally, we integrate explainability strategies to provide security analysts with transparent, interpretable insights into why a particular behavior is flagged as suspicious, improving decision-making in CSA.

This study presents an experimental evaluation of the proposed method using the Windows-Users and Intruder Simulations Logs dataset (WUIL) [4]. We analyze the feasibility and early-stage performance of clustering-based detection in reducing false negatives while maintaining high accuracy. As an exploratory study, this paper provides practical insights into the implementation, optimization, and real-world applicability of AI-driven cybersecurity solutions. These findings serve as a foundation for future research into adaptive and scalable intrusion detection models.

2. Project Description

The proposed approach enhances CSA by improving the efficiency and interpretability of masquerade attack detection. The method leverages a combination of OK-Means clustering [3] and k-Nearest Neighbors (k-NN) to classify potential intrusions while optimizing computational resources. Clustering reduces the search space by grouping similar user behaviors, allowing k-NN to perform instance-based classification on a smaller, more relevant subset of data. This structure improves detection efficiency without significantly compromising accuracy.

OK-Means is a refinement of the traditional K-Means algorithm that optimizes cluster updates, making it better suited for dynamic cyber environments where system behavior constantly evolves. Unlike static machine learning models that require frequent retraining, OK-Means efficiently adapts to new patterns while maintaining clustering quality. Furthermore, k-NN provides an explainable classification process, enabling security analysts to understand why an alert was triggered. To further enhance accuracy, the system integrates Repeated Edited Nearest Neighbor (RENN) [5], a noise reduction technique that filters out inconsistencies and redundant data, reducing false positives and improving model reliability.

The implementation consists of three key steps. First, user activity data is collected from system logs using User Activity Monitoring (UAM) sensors [4]. This data is then preprocessed to extract spatial, temporal, and directional locality features, which help characterize user behavior. Second, the clustering model is periodically updated based on data volume and classification error rate to adapt to evolving attack patterns. Finally, explainability is enhanced through visual analysis tools that allow security analysts to inspect cluster structures and attack feature distributions, making CSA more interpretable and actionable.

3. Implementation and Results

The effectiveness of the proposed clustering-based approach was evaluated using real-world data, with a focus on improving detection efficiency and explainability. This section details the dataset, experimental setup, performance improvements, and the visual tools designed to aid security analysts.

3.1 Dataset and Preprocessing

The evaluation was conducted using the Windows-Users and Intruder Simulations Logs dataset (WUIL) [4], which contains a total of 54,649 instances. The dataset includes both legitimate user activities and masquerade attack attempts, making it well-suited for intrusion detection research. Due to the nature of masquerade attacks, the dataset is highly imbalanced, with 96.77% of instances belonging to legitimate users and only 3.33% corresponding to attacks.

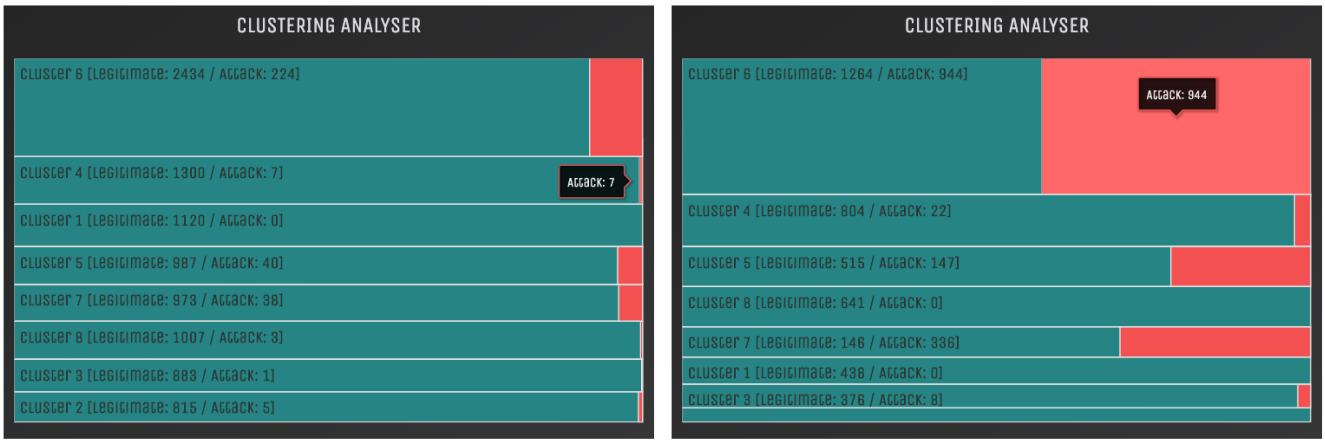

Figure 2. Clustering analyzer tool when visualizing the 20% subsampling (left) and the complete dataset (right).

The dataset consists of multiple behavioral features extracted from user interactions with the file system. These features include:

- File access frequency.
- Event distances between accessed files.
- Directionality of file system navigation.
- Temporal locality patterns in access sequences.

These attributes provide a structured representation of user behavior, enabling clustering and classification models to distinguish between normal and anomalous activities.

3.2 Experimental Setup and Configuration

To analyze the performance of the proposed method, we conducted multiple experiments with different configurations of the clustering and classification components:

- Clustering was performed with $C = 4$ and $C = 8$ clusters.
- For classification, we tested k-NN with $k = 1, 3, 5$ neighbors.
- The number of selected clusters for classification (sC) was varied as $sC = 1, 2, 4$.
- The impact of noise reduction was evaluated using Repeated Edited Nearest Neighbor (RENN) [5], which filters out inconsistent or redundant samples.

The evaluation process included 10-fold cross-validation on five subsamples of the dataset, using 20%, 40%, 60%, 80%, and 100% of the instances. These configurations allowed us to assess the balance between computational efficiency and detection accuracy.

3.3 Performance Gains

The experimental results demonstrated substantial improvements in both computational efficiency and detection performance. Figure 1 presents the recall values and execution times for different noise reduction and clustering strategies.

- **Time Efficiency:** By leveraging OK-Means, execution time was reduced by up to 85% compared to a standard k-NN classifier without clustering.
- **Detection Performance:** The combination of OK-Means and RENN significantly enhanced classification accuracy, achieving up to 99% recall for the detection of masquerade attacks.
- **False Negative Reduction:** The clustering approach minimized the number of undetected attacks, improving the overall reliability of the system.

3.4 Explainability and Visual Tools

A key advantage of this approach is the integration of explainability mechanisms, allowing security analysts to interpret the reasoning behind attack classifications. Two main visual tools were developed:

- **Clustering Analyzer:** This tool provides a hierarchical tree-map visualization of the dataset, showing the proportion of legitimate and attack instances in each cluster. The tool helps analysts assess the risk level of detected attacks by highlighting clusters where anomalies are found.
- **Attack Introspection Tool:** This visualization tool presents a polar chart comparing the detected attack's feature values with those of similar past attacks and the cluster prototype. This allows analysts to understand which behavioral attributes contributed to the classification.

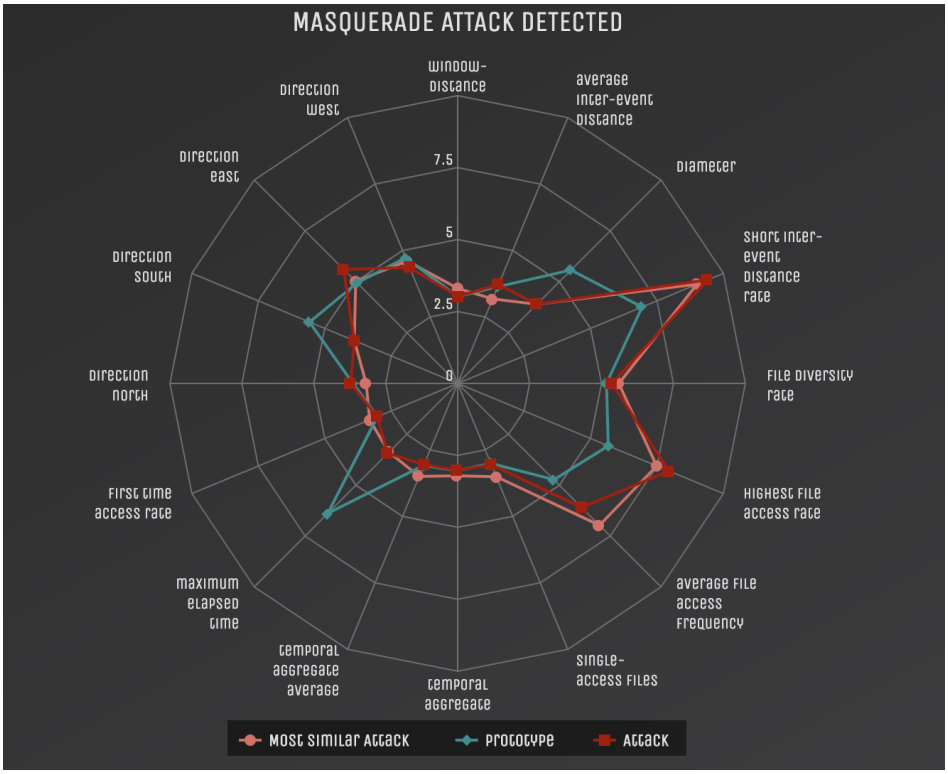

Figure 3. Screenshot of the attack introspection tool showing the features of the attack, the most similar attack that raised the alarm, and the cluster's prototype.

These tools facilitate CSA by enabling analysts to verify whether detected threats are legitimate or false positives, improving trust in automated intrusion detection systems.

4. Discussion and Potential Impact

The proposed clustering-based approach for masquerade attack detection provides a balance between accuracy, efficiency, and interpretability, which are critical components of real-world CSA. Unlike traditional black-box machine learning models, this method offers security analysts actionable insights by leveraging an explainable classification process. The integration of OK-Means and k-NN enhances real-time threat detection while reducing computational overhead, making it suitable for large-scale cybersecurity applications.

While the initial findings demonstrate promising improvements in efficiency and detection performance, further validation is needed to assess long-term scalability and real-world deployment challenges. Future research directions involve integrating adaptive learning techniques to dynamically update clusters based on evolving attack patterns. This would further improve the model's ability to detect previously unseen threats. Additionally, the approach can be expanded to other cybersecurity domains, such as phishing detection, malware analysis, and anomaly detection in industrial control systems. By refining the clustering process and incorporating hybrid AI techniques, this method has the potential to contribute to broader cybersecurity frameworks that prioritize both performance and interpretability.

5. Conclusion

This study explores the feasibility of a clustering-based approach for CSA in detecting masquerade attacks. By combining OK-Means with k-NN, the approach demonstrates a potential balance between computational efficiency, detection accuracy, and explainability. Preliminary results suggest that this method can detect up to 99% of masquerade attacks while reducing execution time by up to 85%. Furthermore, the integration of visual tools enhances interpretability, allowing security analysts to make informed decisions based on attack classifications.

As an exploratory investigation, this work provides early-stage insights into the advantages and limitations of clustering-based CSA methods. Future research should focus on further validation in large-scale, real-world environments, as well as integrating adaptive learning techniques to improve robustness against evolving threats. These findings lay the groundwork for continued innovation in AI-driven cybersecurity solutions.

References

[1] M. Endsley, "Endsley, m.r.: Toward a theory of situation awareness in dynamic systems. human factors journal 37(1), 32-64," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, pp. 32–64, 03 1995.

[2] M. Conti, A. Dehghantanha, and T. Dargahi, "Cyber threat intelligence : Challenges and opportunities," *CoRR*, vol. abs/1808.01162, 2018. [Online]. Available: http://arxiv.org/abs/1808.01162

[3] J. Pérez-Ortega, N. N. Almanza-Ortega, and D. Romero, "Balancing effort and benefit of k-means clustering algorithms in big data realms," *PloS one*, vol. 13, no. 9, p. e0201874, 2018.

[4] J. B. Camiña, C. Hernandez-Gracidas, R. Monroy, and L. Trejo, "The windows-users and -intruder simulations logs dataset (wuil): An experimental framework for masquerade detection mechanisms," *Expert Systems with Applications*, vol. 41, no. 3, pp. 919 – 930, 2014, methods and Applications of Artificial and Computational Intelligence. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413006349

[5] D. Hand and B. Batchelor, "Experiments on the edited condensed nearest neighbor rule," *Information Sciences*, vol. 14, no. 3, pp. 171 – 180, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0020025578900403