ARTIFICIAL INTELLIGENCE IN THE CYBERSECURITY FIELD

DOI 10.20535/2411-1031.2024.12.1.306275
UDC 004.056.53

VOLODYMYR ONISHCHENKO,
OLEKSANDR PUCHKOV,
IHOR SUBACH

INVESTIGATION OF ASSOCIATIVE RULE SEARCH METHOD FOR DETECTION OF CYBER INCIDENTS IN INFORMATION MANAGEMENT SYSTEMS AND SECURITY EVENTS USING CICIDS2018 TEST DATA SET

Automated rule generation for cyber incident identification in information management and security event systems (SIEM, SYSTEM, etc.) plays a crucial role in modern cyberspace defense, where data volumes are exponentially increasing, and the complexity and speed of cyber-attacks are constantly rising. This article explores approaches and methods for automating the process of cyber incident identification rule generation to reduce the need for manual work and ensure flexibility in adapting to changes in threat models. The research highlights the need for utilizing modern techniques of Intelligent Data Analysis (IDA) to process large volumes of data and formulate behavior rules for systems and activities in information systems. The conclusion emphasizes the necessity of integrating multiple research directions, including analyzing existing methods and applying IDA algorithms to search for associative rules from large datasets. Key challenges addressed include the complexity of data modeling, the need to adapt to changes in data from dynamic cyber attack landscapes, and the speed of rule generation algorithms for their identification. The issue of the "dimensionality curse" and the identification of cybersecurity event sequences over time, particularly relevant to SIEM, are discussed. The research objective is defined as the analysis and evaluation of various mathematical methods for automated associative rule generation to identify cyber incidents in SIEM. The most effective strategies for enhancing the efficiency of associative rule generation and their adaptation to the dynamic change of the cybersecurity system state are identified to strengthen the protection of information infrastructure.

Keywords: Intelligent Data Analysis, associative rules, SIEM, cyber incident, cyber threat, cyberspace, data classification, information infrastructure

Problem Statement. Automated rule generation for identifying cyber incidents in SIEM systems is a critical task for protecting modern cyberspace. As data volumes grow exponentially and the speed and complexity of cyber-attacks continually increase, one of the main objectives of deploying cybersecurity systems is to detect relevant behavior patterns or anomalies within the vast data streams processed by SIEM. This requires the integration of various IDA methods to create effective and practical rules. However, traditional methods for rule creation and tuning in cyber incident detection often demand significant manual effort and expert knowledge, leading to delays in threat detection and increasing the risks to the cybersecurity of information infrastructure assets.

The primary goal of the research is to develop new methods and enhance existing ones that enable the automated generation of rules for detecting cyber incidents, thereby reducing the need for constant human intervention and increasing the flexibility of SIEM systems in adapting to new cyber threat models. This involves leveraging modern IDA techniques and improving the efficiency of algorithms that analyze historical and real-time data to formulate rules for identifying anomalous behavior of entities within Information and Communication Systems (ICS) and detecting suspicious activities within them.

The main cyber threats are characterized by great variety, complexity of the data required for their detection and the need for their rapid processing. It is important that the selected and implemented methods can not only effectively detect known types of cyber incidents/attacks but also be capable of adapting to new, unknown types and identifying them in near real time. Additionally, ensuring transparency and interpretability of the automatically generated rules is crucial so that cybersecurity analytics experts can understand and trust the results of their work.

To achieve this goal, it's crucial to integrate multiple research directions. Firstly, a detailed analysis of existing methods for automated rule generation in cyber incident detection should be conducted to identify their weaknesses and potential for improvement. Secondly, the algorithms of these methods should be applied, which will be based on methods of finding rules with the subsequent use of existing big data datasets to identify patterns that may indicate anomalies or potential threats.

The solution to the main problems of applying mathematical methods can be achieved through automated rule generation in SIEM systems. Representing them as mathematical methods illustrates key challenges such as insufficient accuracy, complexity in data modeling, and the need for model adaptability to changes in data and the speed of rule identification.

The curse of dimensionality often arises due to the increase in the number of variables, leading to a sharp increase in the space that needs to be explored. Methods must be capable of adapting to changes in data, which often requires constant parameter updates. The problem of detecting sequences over time is particularly relevant for SIEM systems, as cyber threats often unfold gradually through a series of events occurring over a certain period of time. This aspect can be explained using mathematical formulas associated with time series and methods for their analysis.

The necessity of this research is underscored by the complexity of detecting and analyzing sequences over time, which may contain crucial information about cyber threats in SIEM systems. Specifically, challenges are associated with determining the correct order of rule application, proper noise handling, and accurate detection of change points, all of which are critical for precise identification and response to cyber incidents.

The aim of this work is to analyze and evaluate various mathematical methods for automated rule generation for the identification of cyber incidents in SIEM systems, with a particular focus on using data mining methods, especially sequential analysis. This research aims to study and improve approaches that effectively detect and classify data sequences characteristic of cyber incidents in the high-dynamic conditions of modern cyberspace. The main goal of the study is to determine the most effective strategies for increasing the efficiency of rule generation, responsiveness and adaptability of cyber security systems to strengthen the protection of information resources.

Problem statement for analyzing existing rule generation methods. The task involves researching and analyzing methods of automated rule generation that effectively detect and classify sequences of events indicative of cyber incidents in security monitoring systems (SIEM systems) [1]. Methods should be capable of adapting to changes in threat models and identifying complex and dynamic attack patterns based on analysis of large volumes of data from various sources. The main goals of the research include:

1. Research on the methodology of automated identification and updating of cyber incident identification rules, which are based on the intelligent analysis of data using the Apriori method.
2. Validation and assessment of the advantages and disadvantages of the investigated methods using existing data, assessing their ability to detect new and known types of cyberattacks.
3. Analysis of the ability of methods to effectively detect new threats and attack pattern behaviors, including minimizing false positive and negative detection results.
4. The results of this work should include:
5. A systematic review of existing methods and approaches to automated rule generation.
6. Identification of the advantages and disadvantages of existing methods and algorithms that enhance the accuracy of cyber incident identification through automated rule generation.
7. Proposing ways to address the shortcomings of existing methods and approaches.

A systematic review of existing approaches provides a deep understanding of the current state of automation technologies in the field of cyber incident identification, uncovering their key limitations and opportunities for improvement. By evaluating existing methods and algorithms based

on state-of-the-art principles of statistical and sequential analysis, it was possible to significantly enhance the accuracy and speed of cyberattack identification.

Assessing the practical applicability and effectiveness of the investigated methods in real-world conditions allows not only for the examination of existing methods but also their enhancement for effective integration into the everyday operations of organizations utilizing SIEM systems. Consequently, organizations will not only be able to react to threats but also forecast potential cyber incidents, preempting attackers' moves [2]. This is critically important for creating an adaptive, flexible cybersecurity system capable of effectively responding to dynamic changes in cyber threats within the technological environment. The adaptability of such systems will not only elevate the level of cybersecurity in organizations but also foster the development of adaptive cybersecurity systems capable of confronting continuously evolving threats.

Algorithm for searching association rules using the Apriori method. One of the most well-known methods for discovering association rules is the Apriori data analysis method. This method is used for the automated generation of rules to identify cyber incidents in SIEM systems. The method is employed to detect frequent patterns in large databases or event logs occurring within the perimeter of an information and communication system's security. The core idea of the algorithm revolves around determining the frequency of event identifiers. If a specific event identifier or combination of identifiers appears frequently, then all of their subsets should also appear frequently. The graphical representation of the method can be illustrated as follows:

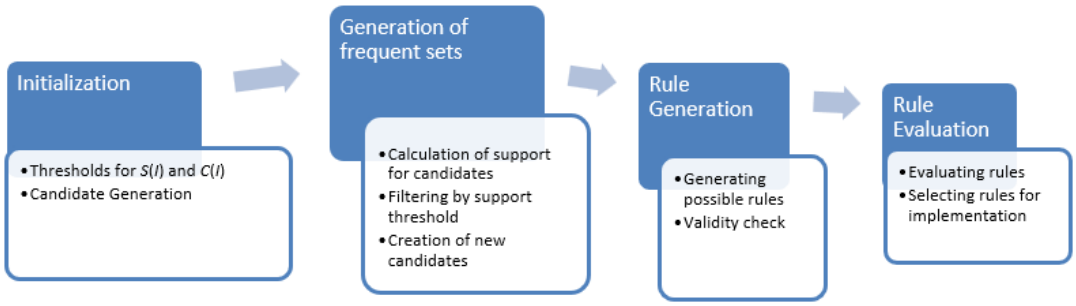

Figure 1 – Structural diagram illustrating the sequence of steps involved in utilizing the Apriori algorithm for automated rule generation in the context of SIEM systems

During the initialization stage, the expert sets the primary parameters that will govern the process of mining association rules. The formation of rules from frequent itemsets in SIEM system databases using the Apriori method can be considered a two-step process [3]: first, frequent sets of compromise event identifiers are found, and then associative rules are generated from these sets. We define frequent compromise identifiers in system logs or transactions in the SIEM system database. The main mathematical concepts used in the method include Support and Confidence[3]–[4]. Let D be a set of transactions, and $S_{\min}$ be the minimum support threshold. Let $I = \{i_1, i_2, \ldots i_n\}$ denote the set of all elements (compromise indicators) that may appear in transactions. We define the minimum thresholds for *Supp* (Support) and *Conf* (Confidence). Candidate sets $C = \{c_1, c_2, \ldots c_k\}$ and rules must satisfy the minimum support and confidence values to be in the set of significance. C_k is the set of all candidate sets containing k elements. The generation of the candidate set occurs by generating initial lists of sets I, consisting of all possible one-element sets i_n. At the stage of generation of frequent sets $F = \{f_1, f_2, \ldots f_k\}$, filtering and selection of elements that occur frequently in the data takes place. For each candidate c in C_k, its support (frequency of occurrence) is calculated. c remains in C_k only if every subset s of $k-1$ elements contained in c is frequent. This means that each such subset s must be present in the set F_{k-1}, where F_{k-1} is the set of all frequent itemsets consisting of $k-1$ elements. Sets that do not meet the support threshold are discarded. This property is called "apriori" and it is critically important for the efficiency of the Apriori algorithm

because it significantly reduces the number of candidates that need to be checked at each step. Essentially, if a subset of a set is not frequent, there is no point in checking a larger set that contains this subset because it will not be frequent either. The efficiency of the method is improved by the Apriori algorithm, which prohibits candidates containing infrequent subsets from forming new larger sets. Mathematically, this can be represented as follows [5]:

$$C_k = \left\{ c \mid \forall s \subseteq c, |s| = k-1 \rightarrow s \in F_{k-1} \right\}.$$

Where c represents the set of candidates with k elements, and s represents subsets of c with $k-1$ elements. The condition $s \in F_{k-1}$ imposes the restriction that every subset must be frequent, meaning its support value must exceed the specified threshold.

The support of a security event identifier or a set of identifier elements is determined as the ratio of the number of transactions containing this element or set to the total number of event transactions in the database. Support is a measure of the frequency (or relative frequency) with which a set of event identifiers appears in the transactional database of a SIEM system. Mathematically, the support of a set of items [6]:

$$Supp(c) = \frac{\left| \left\{ t \in D : c \in t \right\} \right|}{|D|}. \quad (1)$$

Where $|D|$ is the total number of transactions in the database collected by the SIEM system, and $\left| \left\{ t \in D : c \in t \right\} \right|$ is the number of transactions containing the set of identifiers C, where C is the set of event indicators. $c \in t$ means that the candidate set c is a subset of transaction t, i.e., all elements of set c are present in transaction t – representing an individual transaction in the dataset. Using the formula (1) it is possible to determine the frequency and relative frequency at which the set of event indicators occurs in the dataset. It is measured as the fraction of transaction event log entries that contain this set of cyber incident characteristics, relative to all transactions in the dataset. To reduce the volume of data for analysis, less significant itemsets are removed, and focus is placed on those that occur more frequently, increasing the chances of finding significant associative rules by filtering items based on support [7]:

$$F_k = \left\{ c \in C_k \mid Supp(c) \geq Supp_{\min} \right\}$$

Where $Supp_{\min}$ is the setting of the support threshold, which is the minimum support value for a set of items to be considered frequent. This threshold is set based on experience, experiments, or the specifics of the task. All itemsets with support lower than $Supp_{\min}$ are discarded. Only those sets that satisfy the support condition are considered frequent and taken for further analysis. Using the already identified frequent sets, new, larger sets consisting of combinations of frequent elements are generated. For each new larger set, the support is recalculated, and filtering based on $Supp_{\min}$ is performed. This process continues until it is no longer possible to generate new larger frequent sets that satisfy the support threshold, or until F_k becomes empty, indicating that there are no more frequent sets of greater length that can be formed.

After generating frequent itemsets, the second stage involves generating rules. A rule takes the form $A \Rightarrow B$, where A and B are non-intersecting itemsets, meaning $A \cap B \neq 0$. Rules are evaluated using confidence and support (1). In automated rule generation, confidence is an important metric. Assessing the quality of rules directly depends on the task at hand in the context of SIEM systems, especially when rule tuning and selection occur through automated rule generation for detecting cybersecurity incidents. The "quality of rules" refers to the ability of rules to accurately detect real threats without triggering many false positives. Assessing the quality of rules in the Apriori method is based on metrics such as confidence, support, precision, and recall. Let's consider confidence. To do this, we'll define the confidence of a rule $A \Rightarrow B$ in the Apriori method as the ratio of the number of transactions containing both A and B to the number of transactions containing A. Mathematically, this is expressed as follows [8]:

$$Conf(A \Rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)}. \quad (2)$$

Where $A \cup B$ defines the union of sets of compromise indicators A and B, meaning the cases where events A and B occur together. $Supp(A)$ is the proportion of transactions containing A. $Supp(A \cup B)$ is the proportion of transactions containing both A and B. The confidence of the rule $A \Rightarrow B$ indicates how often transactions containing A also contain B. This is expressed by the ratio of the support of $A \cup B$ to the support of A and forms the confidence score of the automatically generated rule. High confidence means that when event A occurs, event B is also very likely to occur. This is crucial for determining the reliability of detection rules in the context of SIEM systems and responding to cybersecurity incidents. This information helps security analysts focus on the most likely and significant alerts, reducing the number of false positives, which adds resilience and reliability to information systems. It's worth noting that when forming automated rules using the Apriori method, a weighting coefficient is not considered [4], which provides the significance and priority of rule application, and this is one of the drawbacks of this method.

The process of forming a rule occurs only if certain constraints are met and considered:

$$Supp(A \Rightarrow B) \geq Supp_{\min}$$

$$Conf(A \Rightarrow B) \geq Conf_{\min}$$

Where $Conf_{\min}$ is the minimum confidence threshold set by an expert in the field of the cybersecurity, considering the imposed constraints on the rule generation process and defining $x \in F$ where $X = \{x_1, x_2, \ldots x_k\}$ is a frequent item set denoted as an element of the set F. Thus, X belongs to the set of all frequent item sets, defined as those satisfying the specified support threshold. X represents a specific frequent itemset from F, which is considered for further subdivision into subsets and rule generation from it. Set X is used for rule generation by analyzing its subsets and determining those that satisfy the criteria of support and confidence. X : Initially, one of the frequent itemsets identified in the previous step of frequent itemset generation is selected. These itemsets satisfy the established support threshold. The set X is divided into all possible non-overlapping subsets X_1, X_2, where $X = X_1 \cup X_2$ and $X_1 \cap X_2 = \varnothing$. This means that the elements in X_1 do not appear in X_2, and vice versa. For each pair (X_1, X_2), it is checked whether the conditions of minimum support and confidence are met. The rule $X_1 \Rightarrow X_2$ is considered valid if it satisfies these thresholds. The support of the rule is defined as the ratio of the number of transactions containing $X_1 \cup X_2$, to the total number of transactions. The confidence of the rule is defined as the ratio of the support of X to the support of X_1, indicating the frequency with which X_2 occurs in transactions also containing X_1. This process is repeated for all frequent itemsets, leading to the formation of a complete set of rules. These rules can then be used to analyze associations between different elements in transactions. Let's apply this method to investigate its performance on a test dataset and generate automated rules for identifying cyber events.

For the test dataset, CICIDS2018 was chosen [9]. The CICIDS2018 dataset consisted of network connection examples, each described by a set of attributes. Here is a typical data structure reflecting the dataset:

– Protocol indicates the protocol number used in the flow (e.g., TCP or UDP).

– Tot Fwd Pkts determines the total number of packets transmitted from the source to the destination.

– Flow Byts/s indicates the average data transmission rate in the flow, measured in bytes per second.

Here is a partial list of cyber-attack indicators that will be used for further analysis of association rules to review the performance of this method. The implementation of this method was done using the PYTHON programming language and the Itertools library. To implement the Apriori algorithm,

a program was written. As a result of running this program, rules were generated based on the CICIDS2018 dataset.

RULE 1: {'Flow Duration < average (Flow Duration)'} → {lable=DDos} When Flow Duration is less than the average, block the source IP address and notify the administrator.

RULE 2: {'Protocol ≠ 6 (TCP) or 17 (UDP) || backward_packets < threshold_packets'} → {'lable=PortScan'} When the Protocol is not equal to 6 or 17, block the connection and record information for further analysis.

RULE 3: {'Tot Fwd Pkts > threshold || total_backward_packets == 0'} → {'lable=Bot'} When the Tot Fwd Pkts exceeds a specified threshold or there are no packets in the reverse direction, the logging level is increased, additional analysis is triggered on the host, and the IP address is blocked.

RULE 4: {'Flow Byts/s or Flow Pkts/s > threshold || flow_bytes > threshold_bytes'} → {'lable=Infiltration'} When Flow Byts/s or Flow Pkts/s is greater than the threshold, often set speed limiting rules for this connection and notify the administrator.

RULE 5: {'RST Flag Cnt || total_fwd_packets > threshold_packets '} → {'lable=Brute Force'} When the "RST Flag Cnt" or other TCP flags indicate a type of attack, or the packet count exceeds a specified threshold, the source IP address is blocked, and the administrator is notified.

RULE 6: {' method == "GET" || method == "POST" and path = .* (<|> | script | alert | onerror |onload).* '} → {'lable=XSS'} When the method is GET or POST and the path matches a specific value, we raise the logging level, initiate additional analysis on the host, and block the IP address.

RULE 7: {' method == "GET" || method == "POST" and path= .*(UNION | SELECT | INSERT | UPDATE | DELETE | DROP | EXEC | OR | AND).*'} → {'lable=SQLInjection'} When the method is GET or POST and the path matches a specific value, we raise the logging level, initiate additional analysis on the host, and block the IP address.

RULE 8: {'RST Flag Cnt || total_fwd_packets > threshold_packets and dst_port = 21 and protocol = 6'} → {'FTP Patator'} When the "RST Flag Cnt" or other TCP flags or Port set 21, indicate a type of attack, or the packet count exceeds a specified threshold, the source IP address is blocked, and the administrator is notified.

RULE 9: {'RST Flag Cnt || total_fwd_packets > threshold_packets and dst_port = 22 and protocol = 6'} → {'SSH Patator'} When the "RST Flag Cnt" or other TCP flags or Port set 21, indicate a type of attack, or the packet count exceeds a specified threshold, the source IP address is blocked, and the administrator is notified.

RULE 10: {'Flow Duration < average (Flow Duration) and port = 80 || port = 443'} → {lable=DDos Slowloris} When Flow Duration is less than the average, block the source IP address and notify the administrator.

In the given rules, the cyber-attack indicators mentioned (Flow Duration, Protocol, Flow IAT) are derived from the context of network traffic analysis and are parts of the CICIDS2018 dataset, which was used for researching and detecting network intrusions. Here's the decryption of the indicators: Protocol – This indicator denotes the type of protocol used in the connection between systems (For example, TCP or UDP). This type of connection typically indicates more stable and legitimate communication between two systems. Flag Cnt: A marker indicating Echo Reply. In the network context, this is part of ICMP (Internet Control Message Protocol) traffic, where a host responds to an ICMP Echo Request (commonly used in ping operations). An Echo Reply indicates that the server is active. ICMP: A protocol used for transmitting control and diagnostic messages in IP-based networks (such as messages indicating host unreachable or no route to the host). It is important for the operation of network diagnostic tools like ping and traceroute. The calculated values of Support, Confidence, and Precision for the discovered rules are listed in the table 1.

Table 1 lists the rules with corresponding values of Support, Confidence, and Precision. Using these metrics, such as Support, Confidence, and Precision, allows for the evaluation of rules and decision-making regarding their use in SIEM systems. The ability to assess enables the evaluation of the effectiveness of each rule within the context of selected characteristics and helps identify the most suitable rules for specific data analysis tasks [10]–[11].

Table 1 – Table of model hyperparameters

	Support	Confidence	Precision
RULE 1	0.570	0.744	0.93
RULE 2	0.570	1.000	0.86
RULE 3	0.570	1.000	0.89
RULE 4	0.570	0.992	0.95
RULE 5	0.574	0.749	0.99
RULE 6	0.574	1.000	0.87
RULE 7	0.570	0.744	0.83
RULE 8	0.570	1.000	0.80
RULE 9	0.570	0.992	0.92
RULE 10	0.570	1.000	0.93

However, a crucial part of finding associative rules is parameter tuning, especially Support and Confidence. These parameters directly affect the quantity and quality of discovered associative rules. Setting Support too high or too low can lead to loss of useful information or excessive computational workload accordingly. Therefore, understanding how these parameters impact algorithm results is important. The dependencies are illustrated in the following figures.

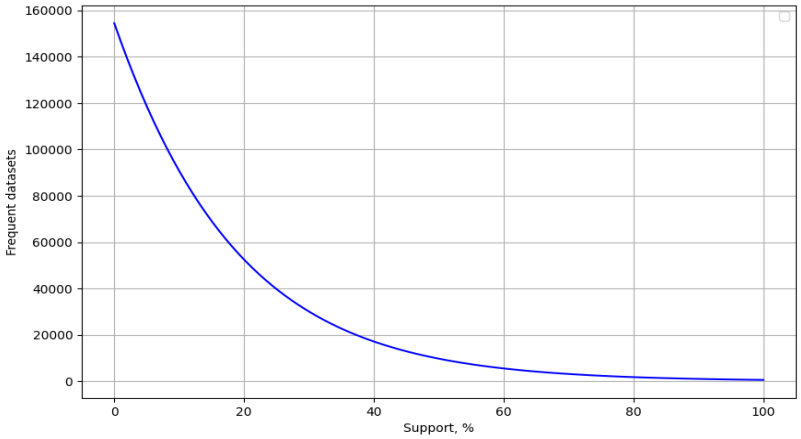

Figure 2 – Dependencies of the number of found frequent itemsets on Support

This dependency shows that if Support is set at a high level, many potentially frequent itemsets will not meet this criterion and therefore will not be included in the results. This reduces the number of found frequent itemsets. With a low level of Support, more specific and rare combinations of elements can be identified, which can be useful for deeper data analysis.

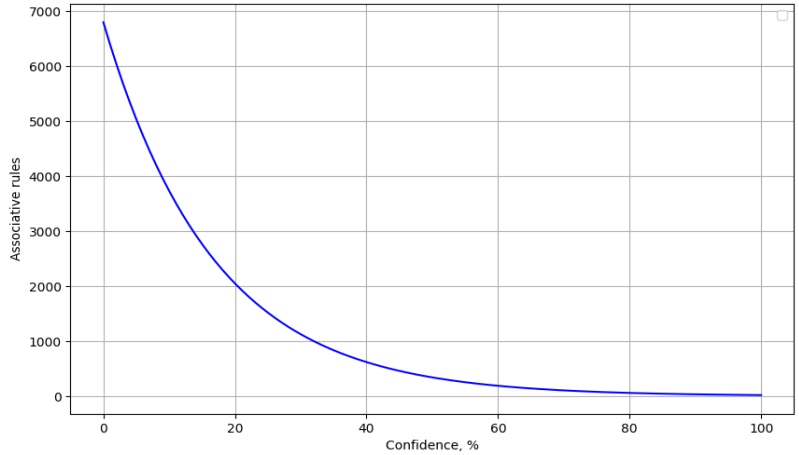

Figure 3 – Dependencies of the number of found frequent itemsets on Confidence

The dependency in Figure 3 helps to understand how frequently different combinations of elements form reliable rules, allowing for the identification of structures and patterns in the data. Rules that remain at high Confidence values are more reliable. The smooth change in the graph of the number of discovered rules indicates a more even distribution of rules across confidence levels, suggesting stability and diversity in associations within the data.

After analyzing Apriori method of automated rule generation, it can be concluded that the Apriori method is one of the basic algorithms for searching associative rules in large databases and performs reasonably well in automatically generating rules. However, according to research, it has several significant drawbacks that may limit its effectiveness and speed, especially when working with very large datasets. Some of these drawbacks include:

Exponential Growth of Candidate Sets: The number of candidates sets that Apriori has to evaluate can grow exponentially with the increase in the size of the dataset and the decrease in the $Supp_{\min}$ threshold. If I is the set of all elements and $|I| = n$, the number of all possible subsets of elements is $2^n - 1$. In practice, this can require significant computational resources.

Sparse data sensitivity: In databases where items rarely appear together (high sparsity), Apriori may struggle to identify significant frequent itemsets, especially if $Supp_{\min}$ is set too high.

A large number of rules with low $Conf_{\min}$: A low $Conf_{\min}$ threshold may lead to the generation of a large number of rules, many of which may be insignificant or random. This can result in noise in the results, making interpretation more difficult.

One of the significant drawbacks of the Apriori method concerning rule formation over time is its inefficiency when dealing with event sequences or time series, where the order of events is crucial. Apriori evaluates the frequency of co-occurrence of items in transactions or datasets but does not consider the temporal sequence in which these items appear. This means that if events A and B occur together in a time interval, but A always occurs before B, Apriori still considers them as two elements that can occur in any order. This lack of consideration for temporal relations between events makes Apriori unsuitable for analyzing time series or event sequences where the order of events needs to be taken into account. This becomes particularly critical in applications where the order of events directly influences the results.

These drawbacks need to be considered when choosing methods for analyzing large datasets and may encourage the use of more advanced or modified methods that address these limitations. Let's consider the results of the CICIDS2018 dataset investigation resulting from the application of association rules.

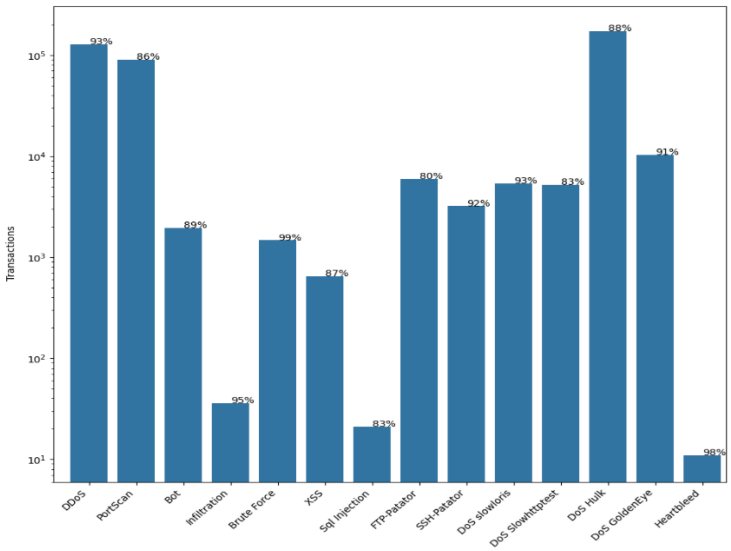

Figure 4 – Diagram for determining the accuracy of cyber-attack identification

Figure 4 shows the quantitative values of detected attacks resulting from the applied rules. The accuracy of these rules is presented in percentage and displayed on the columns for different types of attacks. This diagram allows analyzing the quantitative indicator of transactions identified as malicious. Upon analyzing this graph, it can be concluded that the majority of applied rules have a Precision rate exceeding 90%, which is sufficient for use in cyber incident detection and cyber incidents identification. Rules operating at an 80% Precision level can also be applied but require additional analysis.

Conclusion. The conducted research involved the analysis and application of the Apriori automated rule generation method for identifying cyber incidents in SIEM systems, emphasizing the use of data mining methods. The analysis confirmed that employing data mining techniques significantly enhances the efficiency of detecting and classifying cyber incidents. Both advantages and drawbacks of the Apriori data mining method were identified through its application on the CICIDS2018 dataset. The research demonstrated that the associative rules derived using the Apriori method exhibit sufficient Precision to be effectively applied in the detection and identification of cybersecurity incidents. Utilizing the Apriori algorithm and the methodology for obtaining associative rules contributes to the development of efficient SIEM systems capable of countering contemporary cyber threats.

REFERENCE

[1] B.M. Herasymov, and I.Iu. Subach, “Indicators of the quality of information support and their influence on the effectiveness of the use of decision support systems”, *Bulletin of KNU named after T.G. Shevchenko*, iss. 20, pp. 27-29, 2008.

[2] B.M. Herasymov, I.Iu. Subach, P.V. Khusainov, and V.O. Mishchenko, “Analysis of the tasks of monitoring information networks and methods of increasing the efficiency of their functioning”, *Modern information technologies in the field of security and defense*, no. 3 (3), 24-27, 2028.

[3] C. Islam, M.A. Babar, R. Croft, and H. Janicke, “SmartValidator: A framework for automatic identification and classification of cyber threat data”, Journal of Network and Computer Applications, 202(9):103370, 2022, doi: https://doi.org/10.1016/j.jnca.2022.103370.

[4] [E. Ficke, and S. Xu, “Apin: Automatic attack path identification in computer networks”, in *Proc. 2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Arlington, pp. 1-6, 2020, doi: https://doi.org/10.1109/ISI49825.2020.9280547.

[5] Z. Li, X. Li, R. Tang, and L. Zhang, “Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules”, *Front. Psychol*., 11:582480, 2021, doi: https://doi.org/10.3389/fpsyg.2020.582480. ().

[6] K. Nalavade, and B.B. Meshram, “Finding frequent itemsets using apriori algorithm to detect intrusions in large dataset”, *International Journal of Computer Applications & Information Technology*, vol. 6, iss. 1, pp. 84-92, 2014. [Online]. Available: http://www.ijcait.com/ IJCAIT/61/611.pdf. Accessed on: Mar. 19, 2024.

[7] A.E. Ibor, F.A. Oladeji, and O.B. Okunoye, “A survey of cyber security approaches for attack detection prediction and prevention”, *International Journal of Security and its Applications*, 12(4), 15-28, 2018, doi: https://doi.org/10.14257/ijsia.2018.12.4.02.

[8] N.A. Azeez, T.J. Ayemobola, S. Misra, R. Maskeliūnas, and R. Damaševičius, “Network intrusion detection with a hashing based apriori algorithm using Hadoop MapReduce”, *Computers*, 8(4):86, 2019, doi: https://doi.org/10.3390/computers8040086.

[9] CSE-CIC-IDS2018 on AWS. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html. Accessed on: Mar. 11, 2024.

[10] A. Alsanad, and S. Altuwaijri, “Advanced Persistent Threat Attack Detection using Clustering Algorithms”, *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 640-649, 2022, doi: https://doi.org/10.14569/IJACSA.2022.0130976.

[11] H.N. Mohsenabad, and M.A. Tut, “Optimizing Cybersecurity Attack Detection in Computer Networks: A Comparative Analysis of Bio-Inspired Optimization Algorithms Using the CSE-CIC-IDS 2018 Dataset”, *Applied Sciences*, 14 (3):1044, 2024, doi: https://doi.org/10.3390/app14031044.

Стаття надійшла до редакції 18.06.2024.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Б.М. Герасимов, та І.Ю. Субач, “Показники якості інформаційного забезпечення та їх вплив на ефективність застосування систем підтримки прийняття рішень”, *Вісник КНУ ім. Т. Г. Шевченка*, Вип. 20, с. 27-29, 2008.

[2] Б.М. Герасимов, І.Ю. Субач, П.В. Хусаінов, та В.О. Міщенко, “Аналіз задач моніторингу інформаційних мереж та методів підвищення ефективності їхнього функціонування”, Сучасні інформаційні технології у сфері безпеки та оборони, № 3 (3), с. 24-28, 2008.

[3] C. Islam, M.A. Babar, R. Croft, and H. Janicke, “SmartValidator: A framework for automatic identification and classification of cyber threat data”, Journal of Network and Computer Applications, 202(9):103370, 2022, doi: https://doi.org/10.1016/j.jnca.2022.103370.

[4] E. Ficke, and S. Xu, “Apin: Automatic attack path identification in computer networks”, in *Proc. 2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Arlington, pp. 1-6, 2020, doi: https://doi.org/10.1109/ISI49825.2020.9280547.

[5] Z. Li, X. Li, R. Tang, and L. Zhang, “Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules”, *Front. Psychol*., 11:582480, 2021, doi: https://doi.org/10.3389/fpsyg.2020.582480. ().

[6] K. Nalavade, and B.B. Meshram, “Finding frequent itemsets using apriori algorithm to detect intrusions in large dataset”, *International Journal of Computer Applications & Information Technology*, vol. 6, iss. 1, pp. 84-92, 2014. [Online]. Available: http://www.ijcait.com/ IJCAIT/61/611.pdf. Accessed on: Mar. 19, 2024.

[7] A.E. Ibor, F.A. Oladeji, and O.B. Okunoye, “A survey of cyber security approaches for attack detection prediction and prevention”, *International Journal of Security and its Applications*, 12(4), 15-28, 2018, doi: https://doi.org/10.14257/ijsia.2018.12.4.02.

[8] N.A. Azeez, T.J. Ayemobola, S. Misra, R. Maskeliūnas, and R. Damaševičius, “Network intrusion detection with a hashing based apriori algorithm using Hadoop MapReduce”, *Computers*, 8(4):86, 2019, doi: https://doi.org/10.3390/computers8040086.

[9] CSE-CIC-IDS2018 on AWS. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html. Accessed on: Mar. 11, 2024.

[10] A. Alsanad, and S. Altuwaijri, “Advanced Persistent Threat Attack Detection using Clustering Algorithms”, *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 640-649, 2022, doi: https://doi.org/10.14569/IJACSA.2022.0130976.

[11] H.N. Mohsenabad, and M.A. Tut, “Optimizing Cybersecurity Attack Detection in Computer Networks: A Comparative Analysis of Bio-Inspired Optimization Algorithms Using the CSE-CIC-IDS 2018 Dataset”, *Applied Sciences*, 14 (3):1044, 2024, doi: https://doi.org/10.3390/app14031044.

ВОЛОДИМИР ОНІЩЕНКО,
ОЛЕКСАНДР ПУЧКОВ,
ІГОР СУБАЧ

ДОСЛІДЖЕННЯ МЕТОДУ ПОШУКУ АСОЦІАТИВНИХ ПРАВИЛ ДЛЯ ВИЯВЛЕННЯ КІБЕРІНЦИДЕНТІВ СИСТЕМАМИ УПРАВЛІННЯ ІНФОРМАЦІЄЮ ТА ПОДІЯМИ БЕЗПЕКИ НА ПРИКЛАДІ ТЕСТОВОГО НАБОРУ ДАНИХ CICIDS2018

Автоматизоване формування правил для ідентифікації кіберінцидентів у системах управління інформацією та подіями безпеки (SIEM) відіграє важливу роль у кіберзахисті сучасного кіберпростору, де об'єми даних зростають експоненційно, а складність та швидкість кібератак постійно збільшуються. У статті розглянуто підходи та методи для автоматизації процесу формування правил ідентифікації кіберінцидентів, для зменшення потреби в ручній роботі та забезпечення гнучкості адаптації до змін у моделях загроз. Проведене дослідження висвітлює потребу у використанні сучасних технік інтелектуального аналізу даних (ІАД) для опрацювання великих обсягів даних і формування правил поведінки систем та активності в інформаційних системах. Зроблено висновок про необхідність інтегрування кількох напрямків досліджень, включаючи аналіз існуючих методів та застосування алгоритмів ІАД для пошуку асоціативних правил з даних великого обсягу. Основні виклики, які висвітлюються, включають складність моделювання даних, необхідність адаптації до змін у даних з динамічного ландшафту кібератак та швидкодії алгоритмів формування правил їхньої ідентифікації. Розглянуто проблему "прокляття розмірності" та виявлення послідовностей подій кібербезпеки у часі, які є особливо актуальними для SIEM. Визначено мету дослідження як аналіз та оцінку різних математичних методів автоматизованого формування асоціативних правил для ідентифікації кіберінцидентів у SIEM. Визначено найбільш ефективні стратегії для підвищення ефективності процесу генерації асоціативних правил та їхньої адаптації до динамічної зміни стану системи кібербезпеки для зміцнення захисту інформаційної інфраструктури.

Ключові слова: інтелектуальний аналіз даних, асоціативні правила, SIEM, кіберінцидент, кіберзагроза, кіберпростір, класифікація даних, інформаціна інфраструктура.

Volodymyr Onishchenko, junior researcher, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0009-0000-1355-9178, v.o.onishchenko@ukr.net.

Puchkov Oleksandr, PhD in philosophy, professor, head of the Institute of special communication and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0000-0002-8585-1044, iszzi@iszzi.kpi.ua.

Subach Ihor, doctor of technical science, professor, head at the cybersecurity and application of information systems and technologies academic department, Institute of special communications and information security of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0000-0002-9344-713X, igor_subach@ukr.net.

Оніщенко Володимир Олександрович, молодший науковий співробітник, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

Пучков Олександр Олександрович, кандидат філософських наук, професор, начальник Інституту спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

Субач Ігор Юрійович, доктор технічних наук, професор, завідувач кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв’язку та захисту інформації Національного технічного університетуУкраїни “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.