DOI 10.37882/2223-2966.2024.11.33

# ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ АССОЦИАТИВНЫХ ПРАВИЛ ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ДАННЫХ СЕТЕВОГО ТРАФИКА В ЦЕЛЯХ ОБЕСПЕЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ

## RESEARCH ON THE EFFECTIVENESS OF ASSOCIATIVE RULES FOR EXTRACTING KNOWLEDGE FROM NETWORK TRAFFIC DATA FOR CYBERSECURITY PURPOSES

***O. Sabinin***
***R. Turusov***
***R. Chuprina***

*Summary*. This paper investigates the effectiveness of using associative rules in analyzing network traffic data to discover new knowledge that can be useful for cybersecurity purposes. Associative rules allow the discovery of hidden dependencies and patterns in data, which can carry information missed by other machine learning techniques. The paper presents key approaches and algorithms for generating associative rules, their advantages, and limitations. The results of an experimental study demonstrating the effectiveness of associative rules in analyzing network traffic are presented. The necessity of associative rules integration into traffic analysis systems is substantiated and directions for further research in this area are suggested.

*Keywords*: associative rules, knowledge extraction, traffic analysis, machine learning, data analysis, big data, network security, traffic optimization.

***Сабинин Олег Юрьевич***
*Кандидат технических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого*
*olegsabinin@mail.ru*
***Турусов Роман Андреевич***
*Санкт-Петербургский политехнический университет Петра Великого*
*turusov97@yandex.ru*
***Чуприна Роман Владимирович***
*Санкт-Петербургский политехнический университет Петра Великого*
*romanchuprina55@gmail.com*

*Аннотация*. В данной статье проводится исследование эффективности использования ассоциативных правил при анализе данных сетевого трафика для выявления новых знаний, которые могут быть полезны для целей кибербезопасности. Ассоциативные правила позволяют обнаруживать скрытые зависимости и закономерности в данных, что может нести в себе информацию, упущенную при использовании других методов машинного обучения. В статье приводятся ключевые подходы и алгоритмы генерации ассоциативных правил, их преимущества и ограничения. Представлены результаты экспериментального исследования, демонстрирующие эффективность ассоциативных правил в анализе сетевого трафика. Обосновывается необходимость интеграции ассоциативных правил в системы анализа трафика и предлагаются направления для дальнейших исследований в этой области.

*Ключевые слова*: ассоциативные правила, извлечение знаний, анализ трафика, машинное обучение, анализ данных, большие данные, сетевая безопасность, оптимизация трафика.

## Введение

Анализ сетевого трафика является одним из ключевых аспектов обеспечения кибербезопасности в современных информационных системах. В условиях постоянно растущего объема передаваемых данных и увеличения числа конечных пользователей, необходимость в эффективных методах анализа становится все более актуальной. Цель такого анализа — выявление аномалий в контролируемых параметрах сетевого трафика, обнаружение попыток несанкционированного доступа и предотвращение атак, что позволяет обеспечить защиту сетевой инфраструктуры и данных. Традиционно для анализа сетевого трафика используются методы машинного обучения, направленные на классификацию и кластеризацию данных [7, с. 9].

Классификация в рамках анализа сетевого трафика используется для присвоения классов новым образцам данных на основе обучающего набора. Это позволяет выявлять аномалии, фишинг, вторжения и предсказывать атаки. Наиболее распространенными алгоритмами являются методы на основе деревьев решений, случайных лесов и метода опорных векторов (SVM), а также различные вариации нейронных сетей. Эти методы эффективны для определения типа трафика и обнаружения аномалий в реальном времени [14, с. 29].

Кластеризация применяется в рамках анализа сетевого трафика для группировки данных на основе их сходства. Алгоритмы кластеризации, такие как k-means, DBSCAN и иерархическая кластеризация, позволяют выявлять скрытые паттерны в сетевом трафике и могут быть использованы для сегментации сети и обнаружения аномальных кластеров данных, которые могут указывать на подозрительную активность [6, с. 25; 13, с. 52].

Анализ временных рядов используется для мониторинга изменений в сетевом трафике с течением времени. Такие методы, как ARIMA, Prophet и рекуррентные нейронные сети (RNN), позволяют предсказывать будущее поведение трафика и выявлять аномалии на основе временных зависимостей [2, с. 122].

Несмотря на безусловную эффективность вышеуказанных методов, они имеют свои ограничения. Одним из ключевых недостатков является потеря значительной части информации в процессе подготовки данных для классификации или кластеризации. Этот процесс включает фильтрацию, нормализацию и другие методы предобработки, которые могут привести к утрате полезных признаков, особенно тех, которые могут быть выявлены только в контексте других данных.

Здесь на помощь могут прийти ассоциативные правила, которые позволяют выявлять скрытые зависимости и закономерности. Ассоциативные правила, широко используемые в анализе потребительских корзин в ритейле, до сих пор мало применяются в контексте анализа сетевого трафика. Эти правила позволяют находить часто повторяющиеся комбинации атрибутов в данных и определять их взаимосвязи, что может быть особенно полезно для выявления сложных паттернов в сетевом трафике, которые не поддаются обнаружению методами классификации или кластеризации [3, с. 53; 8, с. 89].

Применение ассоциативных правил, как по отдельности, так и в совокупности с другими методами машинного обучения может значительно повысить качество анализа сетевого трафика, предоставляя дополнительные знания и улучшая обнаружение аномалий и угроз.

Задача исследования заключается в оценке эффективности ассоциативных правил для анализа данных сетевого трафика с целью выявления скрытых знаний, которые могут быть полезны для улучшения методов кибербезопасности и оптимизации сетевого трафика.

## Теоретические основы ассоциативных правил

Ассоциативные правила выявляют зависимости между элементами данных в транзакциях. Транзакция представляет собой множество элементов, которые происходят одновременно. Теоретически, эти зависимости, также называемые правилами, основываются на концепции ассоциаций, которые предполагают наличие связи между элементами данных. Сами правила формируются по понятной концепции, если происходит событие X, то с какой вероятностью произойдёт событие Y. Алгоритм позволяет предсказывать вероятность события при наличии другого связанного с ним события.

Для оценки полученных ассоциативных правил на практике чаще всего применяют три метрики, которые определяются, как поддержка или support, достоверность или же confidence и так называемый лифт или lift. Первая метрика, а именно поддержка определяет то, как часто набор элементов встречается в базе данных с транзакциями. Например, для правила если встречается элемент X, то встречается элемент Y, данная метрика будет математически рассчитана как отношение количества транзакция содержащих оба элемента X и Y одновременно, к общему количеству всех транзакций. Достоверность определяется как условная вероятность появления элемента Y, как последователями, при условии наличия предшественника X в одной транзакции. В общем виде данная метрика математически обосновывается как отношение числа транзакций, содержащих X и Y, к числу транзакций, содержащих только X. Лифт измеряет зависимость между элементами, показывая, насколько чаще элемент Y встречается при наличии элемента X по сравнению с его собственным случайным появлением. Лифт определяется как отношение достоверности к поддержке элемента Y.

На практике реализовано несколько алгоритмов, позволяющих генерировать ассоциативные правила, каждому из этих алгоритмов присущи свои особенности и математическое обоснование на этапе генерации частых наборов элементов, для последующего ассоциативного анализа.

Наиболее часто используемым алгоритмом является алгоритма Apriori, которые в своем аппарате генерации частых наборов элементов использует итеративных подход. На каждом шаге алгоритм формирует кандидатов путём объединения частых совокупностей из предыдущего шага и отсекает те, которые не удовлетворяют заданному уровню поддержки. Основываясь на теории вероятностей, Apriori предполагает, что все непустые подмножества частых совокупностей элементов также должны быть частыми. После нахождения таковых алгоритм генерирует ассоциативные правила путём деления поддержки объединённого набора на поддержку поднабора, что соответствует вычислению достоверности (confidence).

Алгоритм FP-Growth представляет собой инновационный подход к анализу транзакционных данных, основанный на использовании компактной древовидной

структуры данных — FP-дерева. Эта структура позволяет эффективно хранить информацию о транзакциях, избегая необходимости многократного сканирования исходных данных. FP-дерево аккумулирует в себе все необходимые сведения о транзакциях, что значительно сокращает временные затраты на поиск частых наборов элементов и повышает эффективность процесса обнаружения скрытых закономерностей в данных.

Процесс извлечения ассоциативных правил в FP-Growth включает в себя создание условных баз данных для каждого элемента и построение соответствующих условных FP-деревьев. Эти деревья содержат информацию о взаимосвязях между элементами. Алгоритм анализирует пути в условных FP-деревьях, генерируя ассоциативные правила и оценивая их значимость и достоверность по аналогии с алгоритмом Apriori, но с более высокой эффективностью.

Eclat, другой алгоритм интеллектуального анализа данных, предлагает альтернативный подход к представлению транзакционных данных. В отличие от традиционного горизонтального представления, где транзакции представлены списками элементов, Eclat использует вертикальное представление, в котором каждому элементу соответствует набор идентификаторов транзакций, содержащих этот элемент. Eclat находит пересечения между наборами транзакций различных элементов, выявляя элементы, которые часто встречаются вместе в транзакциях.

Например, если элемент A встречается в транзакциях {1, 2, 3}, а элемент B — в транзакциях {2, 3}, то их пересечение {2, 3} указывает на транзакции, содержащие оба элемента. Генерация ассоциативных правил в Eclat осуществляется путем вычисления показателей поддержки и достоверности для найденных частых наборов элементов, что аналогично подходу, используемому в других алгоритмах поиска ассоциативных правил [11, с. 10; 15].

## Анализ работ по исследованию эффективности ассоциативных правил к задачам кибербезопасности

Исследования в области анализа данных, генерируемых различными сетями, показали, что существуют различные методы и приемы, используемые для повышения уровня сетевой безопасности.

В литературе [5, с. 9] предлагается подход на основе ассоциативных правил для обнаружения сетевых вторжений, который анализирует данные из различных источников журналов. Процесс включает такие этапы, как сбор и предварительная обработка данных, выявление часто встречающихся закономерностей с помощью алгоритма Apriori, генерация ассоциативных правил и применение этих правил для обнаружения аномалий в режиме реального времени. Такой подход позволяет быстро обнаружить отклонения от нормального поведения системы, которые могут свидетельствовать о наличии кибератаки, и значительно повышает безопасность информационных систем.

На конференции ICCSIE было представлено исследование, посвященное использованию ассоциативных правил для обнаружения мошенничества в телекоммуникационных сетях [9, с. 11]. Предложенный подход заключается в анализе данных о звонках и транзакциях клиентов, моделировании ассоциативных правил с помощью методов машинного обучения, оценке корреляции между различными моделями поведения и использовании модели для автоматического обнаружения подозрительных моделей. Благодаря этому подходу поставщики услуг могут быстро реагировать на мошенническое поведение и предотвращать финансовые потери.

В исследовании [12] авторы собрали и проанализировали журналы атак из различных источников и применили алгоритмы, такие как FP-Growth, для выявления часто встречающихся паттернов. Результаты исследования помогают понять взаимосвязь между различными факторами, влияющими на успех атаки, и способствуют разработке более эффективных стратегий предотвращения инцидентов в будущем.

В работе [4] описывается применение ассоциативных правил для обнаружения мошенничества в банковской сфере. Предложенная методика предполагает анализ данных о транзакциях, выявление необычных шаблонов транзакций с помощью алгоритмов корреляционного анализа, оценку риска каждой транзакции в режиме реального времени и автоматизацию процесса обнаружения мошенничества. Такой подход может значительно повысить эффективность борьбы с финансовыми преступлениями и защитить интересы банков и их клиентов.

Приведенное выше исследование демонстрирует широкий спектр применения ассоциативных правил в области кибербезопасности и подчеркивает перспективы дальнейших исследований в этом направлении. Использование математических аппаратов ассоциативных правил открывает новые возможности для защиты информации и предотвращения киберугроз, а также способствует разработке более совершенных и надежных методов обеспечения безопасности в цифровой среде.

## Методология использования ассоциативных правил для анализа сетевого трафика

Качество и подготовка данных играют ключевую роль в обеспечении точности и надежности результатов анализа сетевого трафика с использованием ассоциативных

правил. Прежде чем применять методы ассоциативного анализа, необходимо тщательно обработать данные.

Первым шагом является удаление всех пропущенных значений. Заполнение пропущенных значений допустимо в других методах анализа данных, но в анализе сетевого трафика этот метод может привести к искажению результатов из-за значительных вариаций значений параметров трафика. Поэтому очень важно удалить строки с пропущенными значениями.

Следующий шаг — удаление NaN (нечисловых) значений, которые могут возникнуть из-за ошибок при сборе или обработке данных. Наличие NaN-значений может серьезно повлиять на результаты анализа, снизив его точность и достоверность.

Тщательная предварительная обработка данных, включая удаление пропущенных и NaN-значений, является необходимым условием для получения надежных и достоверных результатов анализа сетевого трафика.

Следует также удалить дубликаты записей. В контексте кибербезопасности дублирующие записи могут быть полезны для обнаружения повторяющихся атак, однако в контексте ассоциативных правил дубликаты могут привести к искажению вероятностей и полученных правил.

После этого необходимо изучить данные и выбрать атрибуты, которые могут нести полезную информацию для построения ассоциативных правил. Важно провести экспертную оценку каждого атрибута с точки зрения его значимости и информативности. Оставшиеся после оценки атрибуты необходимо преобразовать в вид, пригодный для применения ассоциативных правил. Данные могут быть числовыми или категориальными.

Для категориальных атрибутов следует применить метод, который преобразует категориальные данные в бинарные признаки, создавая отдельные столбцы для каждого уникального значения атрибута. Для непрерывных данных проводится дискретизация, заменяя числовые значения на интервалы, представляющие собой определенные диапазоны. Это позволяет преобразовать непрерывные данные в категориальные. После этого все данные приводятся к виду транзакций, где каждая строка отображает наличие того или иного признака из изначальной таблицы.

На следующем этапе необходимо выбрать метод поиска ассоциативных правил. Применение метода Apriori, например, позволяет выявить часто повторяющиеся комбинации параметров. В процессе применения алгоритма Apriori необходимо выбрать минимальную степень поддержки для часто повторяющихся наборов данных. На их базе создаются ассоциативные правила, основанные на метрике уверенности или лифта. Полученные зависимости должны быть экспертно оценены с точки зрения их полезности для целей кибербезопасности. Важно проверить объективность и реалистичность выявленных правил, меняя метрики для достижения наилучшего результата. Таким образом, методология применения ассоциативных правил для анализа сетевого трафика включает в себя несколько этапов подготовки данных, построения правил и экспертной оценки.

## Получение скрытых знаний из данных сетевого трафика

Набор анализируемых данных, использованный в данном исследовании, был получен из репозитория данных Mendeley. Он включает в себя данные о сетевом трафике, собранные в среде программно-определяемой сети (SDN). Набор данных содержит различные атрибуты, описывающие характеристики сетевого потока, такие как IP-адреса источника и назначения, количество пакетов, количество байт, продолжительность сеанса, протокол и скорость трафика. Каждая строка представляет собой запись потока с 23 атрибутами в конце каждой строки (транзакции) имеется метка, которая показывает атака (A) это или же нормальное поведение (O).

Согласно вышеуказанной методологии из набора данных были удалены строки с пропущенными значениями и дубликатами, в целях обеспечения целостности данных и снижения ошибки в процессе расчета вероятностей. На следующем этапе в рамках экспертного анализа атрибутов в каждой транзакции был удален столбец времени вследствие низкой важности для ассоциативного анализа. При анализе столбца с атрибутом скорости передачи пакетов были отфильтрованы транзакции с отрицательным значением этого параметра.

Категориальные признаки, такие как источник, назначение, протокол были закодированы с помощью одноточечного кодирования. Этот процесс заключался в преобразовании категориальных переменных в двоичный формат, создавая новые столбцы для каждой уникальной категории. Этот шаг был необходим для подготовки данных к работе алгоритма Apriori, который требует двоичного ввода.

Для признаков с непрерывными числовыми значениями были подобраны пороговые значения, чтобы разделить их на три диапазона. Используемые пороговые значения были определены на основе знаний о предметной области и распределения данных. Для каждой характеристики были созданы три новых столбца, чтобы указать, является ли значение низким, средним или высоким, а исходные столбцы были удалены. Набор данных был преобразован в двоичный формат, где каждый столбец представляет собой наличие (1) или отсутствие (0) определенного признака.

Данные в двоичной кодировке были преобразованы в транзакционное представление с помощью библиотеки mlxtend на языке Python. Это представление представляет каждую строку как набор бинарных атрибутов, подходящих для поиска ассоциативных правил.

Основной целью данного исследования было применение алгоритма Apriori к предварительно обработанным данным сетевого трафика для обнаружения скрытых знаний, заключающихся в уверенных зависимостях и закономерностях между данными в транзакциях.

Для обнаружения частых наборов элементов в данных поддержка была установлена на 20 %. Затем из частых наборов были сгенерированы ассоциативные правила с минимальной уверенностью 60 %. Для фильтрации результатов были выбраны только те правила, у которых в качестве следствия стоял параметр атаки.

Для экспертной оценки полученных правил опорными параметрами для предшественников атаки были выбраны: высокий объем переданных байт, высокая скорость передачи, длительность сессии, высокое количество пакетов на поток, высокое общее количество пакетов. Согласно исследованиям [1, с. 209; 10; 16], данные параметры могут свидетельствовать о попытках утечки данных или об атаке типа DDoS, при которой злоумышленники стараются перегрузить сеть, могут быть признаком сканирования сети или массовой передачи данных, указывает на возможные устойчивые атаки, при которых злоумышленники пытаются поддерживать длительное соединение с целью сбора данных или дальнейшего проникновения.

В таблице 1 указаны элементы, предшествующие метки атаки (A), и полученные значения метрик.

Таблица 1.

Найденные зависимости в данных

| Условие | Следствие | support | confidence | lift |
|---|---|---|---|---|
| высокий объем переданных байт, высокая скорость передачи, высокое количество пакетов на поток, высокое общее количество пакетов | A | 0.2076 | 0.6353 | 1.67 |

Анализируя полученные зависимости между параметрами в транзакциях, можно сделать вывод о том, что с вероятностью в 63 % зафиксированным кибератакам предшествуют высокие значения передаваемого объема байт в совокупности с высоким значением скорости передачи данных и высоким количеством пакетов на поток и общим количеством пакетов. В данном случаем значение support в 21 % говорит о том, что данные параметры встречаются вместе в совокупности транзакций не так часто, что свидетельствует о том, что атаки — это редкие включения периодического характера. Данное значение lift, говорит о том, что совокупность данных параметров встречается на 67 % чаще в качестве предшественника атаки, нежели чем нормального поведения. Основным результатом этого практического исследования является выявление сильной зависимости между высокими значениями определенных сетевых характеристик и вероятностью сетевых атак.

При изменении значений поддержки и уверенности алгоритм способен выявлять новые устойчивые зависимости, включающие различные наборы предшественников и последователей.

## Выводы и рекомендации

В процессе приведенного выше исследования проведен анализ эффективности работы ассоциативных правил в рамках обработки данных сетевого трафика с целью нахождения новых знаний, полезных для целей кибербезопасности. Для эффективной работы с данным математическим аппаратом были описаны теоретические основы алгоритмов, входящих в математическую область ассоциативных правил, представлена методология подготовки данных к применению этих методов и шаги по интерпретации полученных результатов с помощью особых метрик.

На основе проведенного практического анализ можно сделать вывод о том, что применение ассоциативных правил для анализа данных сетевого трафика является эффективным мероприятием. В рамках исследования были получены устойчивые зависимости между сетевыми характеристиками и вероятностью возникновения атаки. Данные зависимости или же знания могут служить отправной точкой в исследованиях по совершенствованию систем мониторинга сетевого трафика.

На основе проведенного исследования рекомендуется интегрировать ассоциативные правила с методами машинного обучения, такими как кластеризация и классификация, для более точного предсказания аномалий в сетевом трафике. Также следует обратить внимание на сбор более детализированных и разнообразных данных о сетевом трафике, что позволит улучшить качество анализа и выявления закономерностей. Наконец, рекомендуется регулярно обновлять и адаптировать модели ассоциативных правил в соответствии с изменяющимися условиями сетевого окружения и новыми типами угроз.

ЛИТЕРАТУРА


1. Aleroud A., & Karabatis G. (2018). Queryable Semantics to Detect Cyber-Attacks: A Flow-Based Detection Approach. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48, 207–223. DOI: 10.1109/TSMC.2016.2600405
2. Bian K. et al. (2021). Characterizing Network Traffic Behavior Using Clustering Techniques. IEEE Communications Magazine, 59, 120–126. DOI: 10.1109/MCOM.001.2000987
3. Feng X. et al. (2022). FAFS: Fuzzy Association Feature Selection Method for Network Intrusion Detection. IEEE Transactions on Fuzzy Systems, 30, 50–60. DOI: 10.1109/TFUZZ.2021.3059598
4. InfoSec Writeups (2021). Association Rule Mining for Cybersecurity. InfoSec Writeups.
5. Kim H. et al. (2022). Effective Association Rule Mining for Cybersecurity. Proceedings of the 2022 ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 24, 1–10. DOI: 10.1145/3558819.3558820
6. Kim J. et al. (2018). Multivariate Network Traffic Analysis Using Clustered Data. Computer Networks, 144, 19–30. DOI: 10.1016/j.comnet.2018.07.011
7. Kotenko I. et al. (2020). Machine learning and data processing for cybersecurity data. IEEE Transactions on Systems, Man, and Cybernetics, 50, 1–11. DOI: 10.1109/TSMC.2020.2968487
8. Kumar V. et al. (2006). Theoretical Foundations of Association Rules. Springer, 5, 80–100. DOI: 10.1007/978-3-540-30142-4_5
9. Lee J. et al. (2020). Network Traffic Analysis Using Association Rule Mining. Sensors, 20, 1–12. DOI: 10.3390/s20123456
10. Oh C., Lee S., Xu W., Vora R., & Kim T. (2022). Mitigating Low-volume DoS Attacks with Data-driven Resource Accounting. ArXiv, abs/2205.00056. DOI: 10.48550/arXiv.2205.00056
11. Srikant R., & Agrawal R. (1996). Mining Quantitative Association Rules in Large Relational Tables. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 26, 1–12. DOI: 10.1145/347090.347101
12. TechTarget. Association rules in data mining [Электронный ресурс]. — Режим доступа: https://www.techtarget.com/whatis/definition/association-rules-in-data-mining, свободный. — Загл. с экрана. — Дата обращения: 03.09.2024.
13. Wang L. et al. (2019). Clustering Analysis of Network Traffic Data. IEEE Access, 7, 150125–150135. DOI: 10.1109/ACCESS.2019.2946967
14. Yong C. et al. (2018). Improved Cluster Analysis Algorithm Using Network Data. Journal of Network and Computer Applications, 94, 30–40. DOI: 10.1016/j.jnca.2017.06.013
15. Zhang J. et al. (2020). Cyber Intrusion Detection through Association Rule Mining on Multi-Source Logs. IEEE Access, 8, 180021–180033. DOI: 10.1109/ACCESS.2020.3020475
16. Zhou L., Liao M., Yuan C., & Zhang H. (2017). Low-Rate DDoS Attack Detection Using Expectation of Packet Size. Secur. Commun. Networks, 2017, 3691629:1-3691629:14. DOI: 10.1155/2017/3691629