RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ПРОГРАММНОЕ И АППАРАТНОЕ ОБЕСПЕЧЕНИЕ РАСПРЕДЕЛЕННЫХ И СУПЕРКОМПЬЮТЕРНЫХ СИСТЕМ

UDC 004.89+004.056
10.25209/2079-3316-2025-16-4-267-285

Использование многоуровневых источников данных для подготовки обучающих наборов для обнаружения кибератак

Дмитрий Дмитриевич **Кононов**[1]✉, Сергей Владиславович **Исаев**[2]
[1,2] Институт вычислительного моделирования СО РАН, Красноярск, Россия
[1]✉*ddk@icm.krasn.ru*

Аннотация. Анализ сетевого трафика является неотъемлемой частью обеспечения безопасности в информационно-телекоммуникационных системах. Использование машинного обучения обеспечивает современным подходам более высокие показатели обнаружения киберугроз.

Предлагается новый подход для формирования обучающих наборов данных, который вводит новую единицу агрегации «сеанс», использует сигнатурный анализ и многоуровневые разнородные источники данных. Сформирован список требований к наборам данных, включающий сохранение первых пакетов соединения, сохранение скрытых областей пакетов, расширенную информацию об источниках трафика (страна, номер автономной системы ASN, тип сети). Дополнительная информация нацелена на выявление атак типа «скрытый канал связи». С использованием предложенного подхода разработан программный комплекс для создания обучающих наборов данных из многоуровневых источников на уровнях L7, L4, L3 модели OSI. В отличие от известных работ, используются реальные данные сетевой активности, а также длительные временные интервалы. Предложенный подход позволяет использовать полученные обучающие наборы для создания более эффективных методов обнаружения и предотвращения вторжения с помощью методов машинного обучения.

Ключевые слова и фразы: Интернет, сетевая безопасность, киберугрозы, анализ сетевого трафика, наборы данных, машинное обучение

Для цитирования: Кононов Д. Д., Исаев С. В. *Использование многоуровневых источников данных для подготовки обучающих наборов для обнаружения кибератак* // Программные системы: теория и приложения. 2025. **Т. 16**. № 4(67). С. 267–285. https://psta.psiras.ru/read/psta2025_4_267-285.pdf

© Кононов Д. Д., Исаев С. В. 2025

Введение

За последние годы сильно возросла роль информационного обмена посредством компьютерных сетей различного уровня, в том числе через Интернет. Государственные и коммерческие информационные системы плотно входят в жизнь современного человека, обеспечивая предоставление различных услуг и сервисов. Корпоративные системы обеспечивают сервисы разного уровня сложности, в том числе возможность полноценной удаленной работы. Большинство подобных систем строятся на основе веб-технологий, предоставляя доступ клиентам на различных платформах.

Часто веб-приложения и веб-сервисы функционируют в публичных сетях и подвергаются различным рискам, связанным с киберугрозами. Задача обеспечения информационной безопасности требует организации и внедрения комплексных мер по защите информации. В зависимости от архитектуры и особенностей веб-приложения для защиты от кибератак используются статистические решения, решения на основе политики безопасности, решения на основе намерений, анализ входов и выходов, статический и динамический анализ приложения [1].

Одним из важных аспектов информационной безопасности является анализ сетевого трафика (NTA – Network Traffic Analysis), который позволяет выявлять и блокировать аномальную активность в сети. Большую угрозу представляют различные ботнеты, которые в автоматическом режиме сканируют системы для выявления уязвимостей. С каждым годом их активность возрастает, также возрастает сложность обнаружения и блокировки, что требует новых подходов, например, комбинирование различных методов машинного обучения [2]. Одной из проблем для анализа трафика является широкое распространение зашифрованных соединений, что не позволяет использовать простые методы идентификации.

Публикуются различные подходы для идентификации и классификации зашифрованного трафика. В [3] приведен обзор существующих работ по анализу зашифрованного трафика, определена классификация по различным направлениям (анализ, безопасность, приватность). В частности, для направления «анализ» выделены категории: идентификация протоколов и приложений, идентификация использования приложений, декодирование потоков, исследование QoS. Для направления «безопасность» выделена категория «Network intrusion detection», которая включает широкий ряд работ от сигнатурного анализа до нейронных сетей.

Активно обсуждается в научных публикациях проблема защиты от кибератак устройств IoT (Internet of Things) различными методами, в том числе с помощью анализа трафика. Авторы [4] приводят древовидную таксономию видов атак на устройства IoT, описывают жизненный цикл

ботнетов, методы обнаружения и рекомендации по их применению в зависимости от профиля и сложности атак.

Анализ сетевого трафика позволяет выявлять аномалии и нештатное поведение веб-приложений и веб-сервисов. Авторы используют различные подходы для идентификации кибератак, которые включают кластерный анализ [5], метод опорных векторов [6], статистические методы [7], а также методы машинного обучения [8]. Как правило, анализ сетевого трафика сопряжен с обработкой источников данных, которые могут иметь большой объем, поэтому для оптимизации времени обработки используются различные методы декомпозиции и многоэтапной обработки. Таким образом, анализ сетевого трафика позволяет получать информацию, которая может быть использована для различных методов обеспечения безопасности и снижения рисков киберугроз.

При идентификации и предотвращении атак существенную роль играет возможность раннего обнаружения с целью минимизации возможных последствий сканирования или вторжения. В работе [9] авторы предлагают многоэтапный анализ трафика с определением самоподобия данных и применением статистических методов, что позволяет обнаруживать кибератаки в реальном или близком к реальному времени. Актуальность минимизации времени обнаружения описана в работе [10], в которой авторы применяют методы глубокого обучения, анализируя содержимое пакетов для идентификации веб-атак в режиме реального времени.

Некоторые авторы [11] приходят к необходимости применения машинного обучения вместо статистических методов для предсказания трафика, что позволяет минимизировать ущерб, однако включает риски ложноположительных срабатываний. Особенно актуально раннее обнаружение для атак вида «отказ в обслуживании» (DoS – Denial-of-service) и «распределенный отказ в обслуживании» (DDoS – Distributed denial-of-service). Методы машинного обучения применяются для защиты от DoS-атак и позволяют получить высокую точность обнаружения [12].

В последнее время более актуальными являются распределенные атаки DDoS, в которых одновременно участвует множество сетевых узлов, что усложняет применение традиционных мер защиты. Другие авторы [13] применяют несколько методов машинного обучения (опорных векторов, байесовские алгоритмы, случайный лес, логистическая регрессия), результаты которых передаются в модуль для голосования, который определяет наличие атаки. Использование этого подхода позволяет блокировать нежелательный трафик без побочных эффектов для информационных систем.

Другими распространенными веб-атаками являются инъекции данных и кода, в том числе SQL-атаки. В работе [14] предлагается методика

обнаружения SQL-инъекций с использованием сигнатурного детектора для формирования обучающих наборов данных, на основе которых с помощью методов машинного обучения достигается высокая точность обнаружения таких видов атак.

Ниже описано применение методики формирования обучающих наборов данных за счет агрегирования нескольких источников на основе реальных данных сетевого трафика Красноярского научного центра (ФИЦ КНЦ СО РАН). Отличие от известных работ заключается в использовании данных реальных сетевой активности на различных уровнях модели OSI (Open Systems Interconnection) на длительных временных интервалах, что позволяет оценивать и сравнивать тренды за различные периоды и использовать полученные обучающие наборы для создания более эффективных методов предотвращения вторжений. Также авторами вводится новый элемент «сеанс», включающий все запросы на прикладном, транспортном и сетевом уровнях за непрерывный период активности и дополнительную информацию об источнике трафика. Сеанс является целостной сущностью, для него формируются атрибуты набора данных.

Целью работы является построение актуального расширенного набора данных сетевого трафика для идентификации вторжений с использованием различных методов анализа данных и машинного обучения. При построении набора данных используются разнородные источники, соответствующие сетевому L3, транспортному L4 и прикладному L7 уровням модели OSI. Разработан программный комплекс для автоматизированной подготовки обучающих наборов данных из разнородных источников. Публикуемая работа является промежуточным этапом в рамках создания автоматизированной системы обнаружения и предотвращения вторжений, архитектура и методика функционирования которой прорабатывается авторами. Созданное программное обеспечение является законченным, однако предполагает возможность дальнейшего развития при решении будущих задач.

Работа состоит из введения, описания источников данных, требований к наборам данных, метода и алгоритма обработки. Описывается использование нового элемента «сеанс». Приводятся смежные работы и их особенности. В конце работы описываются полученные результаты и планы на будущее.

Смежные работы

Множество авторов [15, 16] в своих исследованиях используют только один публичный набор данных на одном уровне модели OSI, что позволяет сравнить свой метод с методами других авторов, однако не учитывает особенности конкретных информационных систем и ограничивает применение

только в рамках заданного уровня. Также общедоступные наборы данных имеют следующие особенности:

(*1*) обученные на HTTP модели плохо работают на HTTPS и наоборот;
(*2*) невозможно использовать признаки прикладного уровня из-за повсеместного применения шифрования (HTTPS, IPSec);
(*3*) в большинстве наборов используются только признаки сетевого соединения (сетевой L3 и транспортный L4 уровни модели OSI);
(*4*) некоторые наборы являются устаревшими и не отражают современные тренды киберугроз (например, KDD Cup 1999).

В отличие от упомянутых работ, предлагаемый авторами метод позволяет формировать актуальные обучающие наборы данных согласно современным трендам и учитывать особенности функционирования гетерогенной сетевой инфраструктуры.

Некоторые авторы используют комбинированные многоступенчатые подходы, что позволяет совмещать преимущества различных методов. В работе [**17**] авторы исследуют предотвращение SQL-инъекций на основе журналов веб-приложения и используют методы сопоставления по шаблону и машинное обучение, что повышает обнаружение атак по сравнению с одноступенчатым методом. Использование одиночных источников данных ограничивает применение методов машинного обучения, поэтому представляет интерес комбинирование нескольких источников данных для формирования обучающих наборов данных. В отличие от указанной, авторы данной работы используют несколько разнородных источников, что позволяет повысить эффективность идентификации атак, в том числе по скрытым каналам связи.

В работе [**18**] предлагается использовать в качестве источников данных журналы приложения и сервиса Datiphy, который выступает посредником между приложением и СУБД MySQL, в результате значительно повысилась точность обнаружения SQL-атак. Помимо однородных источников данных авторы используют разнородные источники, что позволяет насытить обучающие наборы признаками, которые могут свидетельствовать о наличии аномалий в системе при их отсутствии в других наборах данных. Особенность работы – сужение области идентификации до SQL-атак. Предложенный авторами данной работы подход позволяет идентифицировать широкий класс атак на веб-системы.

Работа [**19**] использует подход с агрегацией данных из трёх источников: данные о пакетах в формате PCAP, системные журналы и статистика сетевого узла. Применение данного подхода повысило точность обнаружения веб-атак при увеличении нагрузки на процессор всего на 2.1% по сравнению с одиночными источниками данных. Использование

нескольких источников данных является актуальным в том числе для предотвращения таргетированных атак. Ограничение подхода заключается в использовании редуцированного набора полей из файлов PCAP, также не приводится алгоритм слияния атрибутов из нескольких источников. В отличие от указанной, авторы данной работы собирают расширенную информацию из полей PCAP, а также предлагают метод для слияния разнородных источников согласно модели OSI.

Авторы [20] описывают разработку сетевого фреймворка для анализа атак в системах с большим объемом данных (big data) и предлагают использовать такие источники, как данные NetFlow, системные и прикладные журналы, данные файрволла, антивируса, а также публичные базы данных уязвимостей. Представленный подход позволяет собирать события об аномалиях на различных уровнях и идентифицировать атаки со слабой интенсивностью, используя корреляционный анализ. Однако данные о потоках NetFlow имеют минимальный набор полей, что делает невозможным учет специфических для протоколов полей и флагов, в том числе замаскированных. Также следует отметить, что атаку может характеризовать несколько потоков NetFlow, которые нужно объединять для корректного применения методов машинного обучения.

В работе [21] описан подход с использованием нескольких источников данных для метода опорных векторов (SVM – Support Vector Machine), показывающий более высокую точность по сравнению с традиционными подходами. Между тем для оценки метода в качестве экспериментальных данных авторы указанной работы используют синтетический однородный набор данных и не указывают его структуру и перечень используемых источников.

Источники для наборов данных

В существующих работах авторы используют общедоступные тестовые наборы данных со стандартными веб-атаками и оценивают эффективность своих методов в лабораторных условиях, а не в реальных сетевых инфраструктурах, которые могут иметь свои особенности. Как было сказано выше, публичные тестовые наборы часто содержат ограниченные и устаревшие данные по веб-атакам. Некоторые авторы используют слишком разнородные источники данных, которые косвенно относятся к анализу трафика (например, метрики поведения пользователей) и не всегда могут быть доступны в информационных системах. В других исследованиях применяются короткие временные интервалы для поиска кибератак, что не позволяет сравнивать риски на больших интервалах и оценить направление движения трендов.

Для формирования собственных обучающих наборов данных требуются источники данных, из которых возможно сформировать нужный набор признаков и при идентификации веб-атаки указать ее тип согласно общепринятой классификации. В работах часто применяются два вида классификации: *OWASP (Open Web Application Security Project) TOP 10* и *CAPEC (Common Attack Pattern Enumeration and Classification)*. Как показано выше, актуальным является использование одновременно нескольких источников данных, что позволяет добиться большей точности обнаружения кибератак. В описываемой работе источниками являются данные сетевой активности PCAP пограничного маршрутизатора и журналы веб-сервисов корпоративной сети Красноярского научного центра. Схема сети и точки сбора данных показаны на рисунке 1. Данные

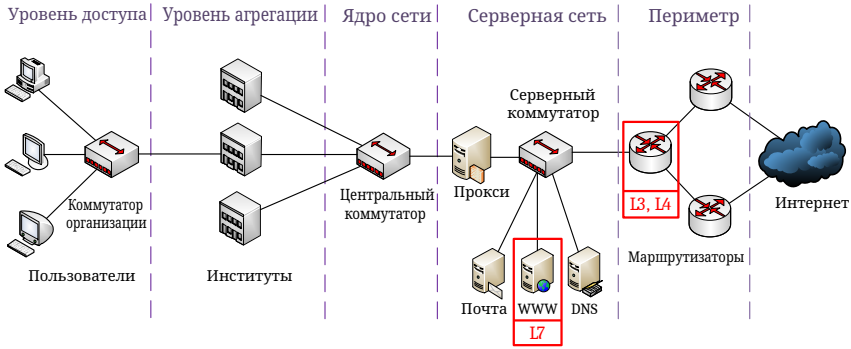

Рисунок 1. Схема сети и точки сбора данных

на уровне L3 и L4 сохраняются в формате PCAP, данные на уровне L7 сохраняются в виде текстовых журналов служб Nginx и Apache. Объем данных за 1 месяц для журналов трафика PCAP составляет 600 Гб и 1 млрд. записей, для журналов веб-сервисов – 4 Гб и 20 млн. записей. Значительный объем данных требует применения высокопроизводительных методов обработки и анализа.

Требования к наборам данных

Авторами работы предложен перечень требований к обучающим наборам данных, сформированных на основе многоуровневых источников данных. Требования включают 2 части: общую и специальную. Первая часть формирует общие требования при сохранении данных для дальнейшего обучения, которые также используются другими авторами, а именно:

1. Необходимо использовать отраслевые стандарты и терминологию для исключения разночтений при обработке и анализе данных.
2. Возможность сравнения метаданных – использование одинакового пространства имен признаков с возможность ретроспективного анализа.
3. Полнота трафика – обучающие наборы должны содержать как легитимный трафик, так и атаки.
4. Двунаправленный трафик – необходимо сохранять как входящие, так и исходящие пакеты.
5. Различные уровни источников по классификации OSI.
6. Необходимо обеспечить единообразную классификацию кибератак для дальнейшего анализа (OWASP TOP10, CAPEC и другие).

Вторая (специальная) часть предложена авторами для формирования расширенного набора признаков с целью обеспечить возможность более точного анализа:

1. Сохранение первых N пакетов соединений (как в системах DPI – Deep Packet Inspection) для использования статистических методов, которые могут определить вероятность появления нелегитимного трафика.
2. Сохранение скрытых областей пакетов. Применяемые в сети Интернет протоколы, как правило, имеют фиксированные заголовки с предопределенными полями пакетов, а также различные расширения. Если пакет имеет небольшой размер или его размер не кратен определенному значению, различные приложения заполняют оставшуюся часть пакета пустыми данными. В некоторых видах атак такие пустые места могут использоваться для передачи информации по так называемому «скрытому каналу связи». В частности, некоторые ботнеты управляются командами, которые отправляются в поле payload протокола ICMP. Традиционные подходы и системы обнаружения вторжений не учитывают такие скрытые поля и не используют их для анализа.
3. Сохранение расширенной информации об источнике трафика (геопривязка, номер автономной системы ASN). Эта информация включает страну, город, числовой номер ASN и другую дополнительную информацию, которая позволяет идентифицировать источник атаки. Предыдущие работы показывают, что риски киберугроз не являются одинаковыми со стороны разных стран-источников [22]. Использование расширенной информации об источниках позволит повысить точность обнаружения кибератак.

Метод формирования наборов и алгоритм его реализации

Для формирования обучающих наборов данных был предложен метод, основанный на сигнатурном анализе. Суть метода заключается в использовании сигнатурного анализа на прикладном уровне L7 для формирования и разметки набора данных.

При создании набора данных формируются дополнительные атрибуты, характеризующие общие закономерности профиля атаки, которые в дальнейшем будут использоваться для выявления атаки без использования сведений о сигнатуре. Обобщающая способность классификаторов на основе данной обучающей выборки позволит выявить атаки с неизвестными ранее сигнатурами.

Упрощенный алгоритм метода представлен на рисунке 2.

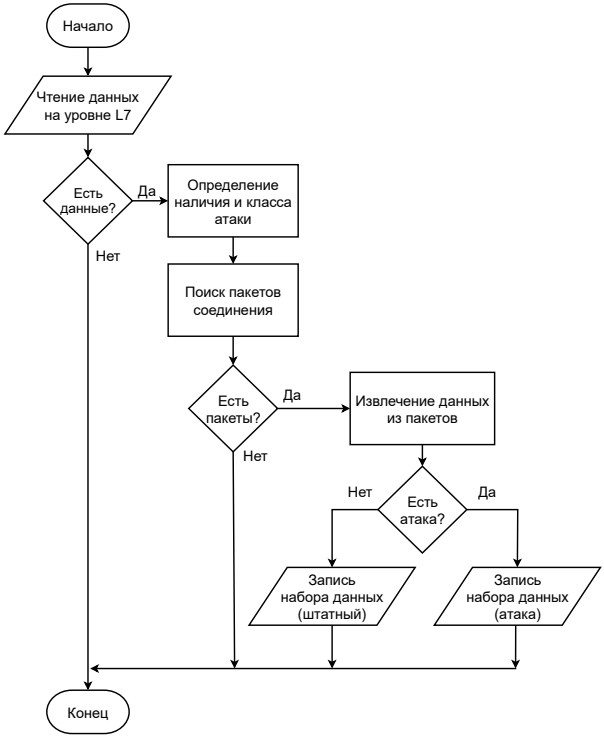

Рисунок 2. Алгоритм формирования обучающих данных

Алгоритм состоит из следующих шагов:

1. Для записей журналов прикладного уровня происходит их агрегирование в сеансы активности. В один сеанс объединяются записи обнаруженного источника по IP-адресу, отстоящие друг от друга не более, чем на заданных интервал времени. Интервал задается на основе предварительного статистического анализа.
2. Для каждого сеанса запускается процесс идентификации веб-атаки на основе сигнатурного анализа с помощью классификатора *OWASP CRS*. При обнаружении атаки сохраняется ее тип согласно классификации OWASP TOP10 или CAPEC, который добавляется в выходной набор данных. Если сигнатурный метод не обнаруживает веб-атаку, признак типа атаки не заполняется, набор данных считается штатным.
3. Для каждого сеанса выполняется поиск пакетов соединения на сетевом и транспортном уровнях (L3 и L4 по классификации модели OSI).
4. Для найденных пакетов извлекаются дополнительные атрибуты, производится их агрегация, в том числе рассчитываются статистические характеристики сеансов, такие как: длительность сеанса, объемы переданной информации, количество пакетов, скорости исходящего и входящего потоков, межпакетный интервал, минимум, максимум и среднее размеров пакетов и сессий, среднеквадратическое отклонение размеров пакетов и сессий и другие. Агрегированные значения добавляются в обучающие наборы данных.
5. Для найденных пакетов извлекаются байты из областей расширения и выравнивания для протоколов IP, TCP, UDP, которые обычно имеют стандартное заполнение. При наличии непустых значений формируются атрибутивные строки.
6. По IP-адресу определяется страна-источник и номер автономной системы ASN, которые также добавляются в признаки. Наличие расширенной информации об источнике позволяет определить не только страну, но и провайдера, тип соединения (проводной, мобильный) и класс пользователя (домашний, корпоративный, хостинг, вредоносный). Данная информация позволит дополнительно учитывать риски киберугроз источника.

Таким образом, разработанный метод позволяет агрегировать данные из источников на различных уровнях модели OSI. Агрегирование происходит на основе сеансов, для которых заполняются атрибуты создаваемого набора данных. Авторы предлагают сохранять скрытые области и расширения протоколов пакетов, что позволит выявлять атаки

типа «скрытый канал связи». Дополнительная информация об источнике трафика (страна, номер автономной системы ASN, тип соединения, класс пользователя) позволит ранжировать их по уровню риска.

Согласно предложенному алгоритму разработан программный комплекс автоматизированной подготовки обучающих наборов данных активности веб-сервисов для машинного обучения. Программный комплекс позволяет осуществлять распределенную обработку данных и состоит из нескольких частей: обработчик, клиент, сервер, брокер сообщений, вычислительный кластер (рисунок 3). Обработчик получает данные

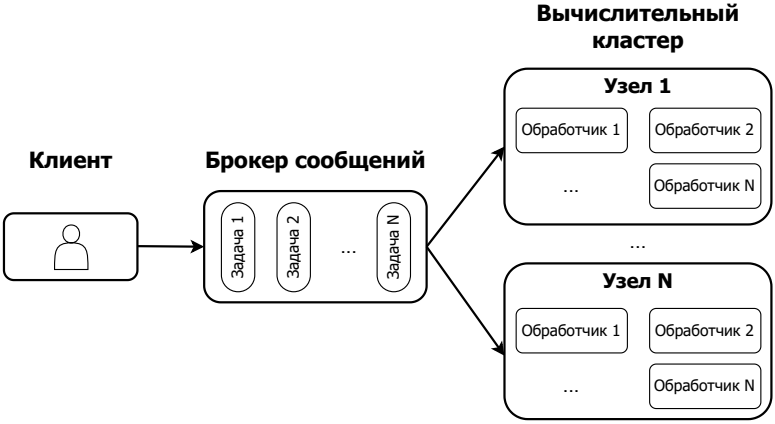

Рисунок 3. Архитектура программного комплекса

из нескольких источников на разных уровнях модели OSI. На уровнях L3 и L4 данные извлекаются из наборов файлов в формате PCAP, на уровне L7 – из журналов веб-сервера. Для выявления атак применяется сигнатурный метод анализа. При обнаружении атаки сохраняются данные о ее типе, пакеты сеанса, из которых извлекаются атрибуты, выполняется их агрегация, рассчитываются статистические показатели. Клиент взаимодействует с брокером сообщений и позволяет запускать на выполнение заранее подготовленные задания, размеченные в формате YAML.

Серверная часть состоит из множества узлов вычислительного кластера, каждый узел может запускать несколько обработчиков параллельно, что обеспечивает параллельную обработку данных и позволяет уменьшить время выполнения заданий. Обработчик обращается к брокеру сообщений, получает новое задание, запускает его на выполнение, сохраняет результаты обработки, затем получает новое задание. Брокер сообщений

распределяет задания для выполнения на серверной части, запущенной на различных узлах вычислительного кластера. Для хранения и обработки очереди заданий используется СУБД Redis.

Основные результаты

На основе разработанного метода агрегации сетевых данных создан программный комплекс, предназначенный для автоматизированной подготовки обучающих наборов данных для машинного обучения [**23**]. Программный комплекс позволяет агрегировать данные на прикладном L7, транспортном L4 и сетевом L3 уровнях модели OSI. Поскольку PCAP-файлы сохраняют данные и на других уровнях, имеется возможность извлекать данные на канальном уровне L2, однако это не является целью данной работы и лишено смысла в контексте удаленных веб-атак. Программа позволяет извлекать признаки из журналов прикладного уровня веб-серверов Nginx и Apache. В качестве источников транспортного и сетевого уровней используются наборы пакетов в формате PCAP с сохранением всех полей (имеется возможность обработки сжатых файлов).

Программный комплекс имеет модуль для сигнатурного анализа журналов прикладного уровня и определения типа атаки по классификациям OWASP TOP10 и CAPEC. Модуль основан на программном коде открытого проекта ModSecurity, межсетевого экрана веб-приложений (WAF – Web Application Firewall). Авторы используют разработанную на языке Си библиотеку ModSecurity, что обеспечивает высокую скорость обработки правил, а также позволяет осуществлять идентификацию веб-атак в режиме реального времени. Модульность проекта позволяет использовать его не только в составе веб-сервера Apache и Nginx, но и для анализа архивных данных из произвольных источников.

База для сигнатурного анализа основана на общедоступной базе данных OWASP CRS (Core Rule Set). Эта база имеет набор сигнатурных правил для обнаружения распространенных видов веб-атак. Преимущества OWASP CRS: идентификация различных классов атак согласно OWASP TOP10, бесплатность, возможность расширения, хорошая документация, активная поддержка сообществом. Поскольку правила CRS ориентируется не на конкретные эксплойты, а на общее описание атак, поэтому они позволяют идентифицировать новые и ранее неизвестные атаки.

Различные коммерческие компании разрабатывают свои базы для межсетевых экранов веб-приложений на основе OWASP CRS. Авторы выполнили модификацию базы CRS с учетом опыта защиты информационных

систем на различных платформах для более эффективного обнаружения веб-атак. Модификации включают обнаружение специфичных атак на IIS (Internet Information Server) платформы Microsoft Windows, PostgreSQL и других коммерческих продуктов, в том числе от российских вендоров (КриптоПро).

Также программный комплекс включает модуль для определения страны-источника для IP-адреса и номера автономной системы ASN на основе публичных баз данных. Архитектура программного обеспечения позволяет добавлять новые признаки с использованием функций преобразования на основе существующих признаков. Имеется возможность агрегирования признаков согласно заданным функциям (количество, сумма, минимум, максимум, различные статистические показатели). Использование в качестве единицы агрегации сеанса связи, а не TCP-сессии позволяет более полно характеризовать поведение источника трафика за длительный период времени.

Программный комплекс написан на компилируемом языке Go и является кроссплатформенным, что обеспечивает возможность его запуска на различных операционных системах, включая Linux, *BSD, Windows. Использование языка Go обусловлено возможностью полноценного встраивания модулей на языке Си и дальнейшей интеграции с другими программными продуктами, разработанными авторами для анализа сетевого трафика.

Выходные данные записываются в широко распространенном формате CSV с указанием имен признаков. Этот формат наиболее часто используется в различных программных продуктах для машинного обучения (PyTorch, TensorFlow, Azure ML), а также в публичных наборах данных. Результаты работы программного комплекса могут быть использованы для формирования обучающих выборок из многоуровневых источников в автоматизированных системах обнаружения кибератак на веб-сервисы на основе методов машинного обучения.

Заключение

В рамках работы авторами предложен новый метод подготовки обучающих наборов данных из многоуровневых источников, которые могут быть использованы в системах обнаружения и предотвращения вторжений на веб-сервисы на основе машинного обучения (соответственно, IDS – Intrusion Detection System и IPS – Intrusion Prevention System).

Авторами сформулированы требования к наборам данных, которые позволяют извлекать и анализировать данные по скрытым каналам связи, недоступные в традиционных подходах, и новая единица агрегации «сеанс». Метод включает использование дополнительных признаков на основе характеристик источников трафика.

Разработан кроссплатформенный программный комплекс для формирования обучающих наборов данных из многоуровневых источников на уровнях L3, L4, L7 модели OSI. Предложенный подход позволит увеличить точность обнаружения кибератак с помощью методов машинного обучения.

Предполагается использование полученных наборов данных для тестирование существующих и разработки новых методов идентификации и предотвращения сетевых угроз. Также планируется обогащение наборов с помощью данных, полученных на основе статистического анализа и выявления аномальных источников.

Список использованных источников

[1] Лесько С. А. *Модели и методы защиты веб-ресурсов: систематический обзор* // Cloud of Science.– 2020.– Т. **7**.– № 3.– С. 577–610. ↑268

[2] Duan L., Zhou J., Wu Y., Xu W. *A novel and highly efficient botnet detection algorithm based on network traffic analysis of smart systems* // International Journal of Distributed Sensor Networks.– 2022.– Vol. **18**.– No. 3.– 17 pp. ↑268

[3] Papadogiannaki E., Ioannidis S. *A survey on encrypted network traffic analysis applications, techniques, and countermeasures* // ACM Computing Surveys.– 2021.– Vol. **54**.– No. 6.– id. 123.– 35 pp. ↑268

[4] Singh N. J., Hoque N., Singh K. R., Bhattacharyya D. K. *Botnet-based IoT network traffic analysis using deep learning* // Security and Privacy.– 2024.– Vol. **7**.– No. 2.– id. e355.– 40 pp. ↑268

[5] Zhang P., Ma W., Qian S. *Cluster analysis of day-to-day traffic data in networks* // Transportation Research Part C: Emerging Technologies.– 2022.– Vol. **144**.– id. 103882.– 24 pp. ↑269

[6] Zhongsheng W., Jianguo W., Sen Y., Jiaqiong G. *Retracted: Traffic identification and traffic analysis based on support vector machine* // Concurrency and Computation: Practice and Experience.– 2020.– Vol. **32**.– No. 2.– id. e5292. ↑269

[7] Nie L., Jiang D., Lv Z. *Modeling network traffic for traffic matrix estimation and anomaly detection based on Bayesian network in cloud computing networks* // Annals of Telecommunications.– 2017.– Vol. **72**.– Pp. 297–305. ↑269

[8] Abbasi M., Shahraki A., Taherkordi A. *Deep learning for network traffic monitoring and analysis (NTMA): A survey* // Computer Communications.– 2021.– Vol. **170**.– Pp. 19–41. ↑269

[9] Kotenko I., Saenko I., Kribel A., Lauta O. *A technique for early detection of cyberattacks using the traffic self-similarity property and a statistical approach* // *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (Valladolid, Spain, 10–12 March 2021).– IEEE.– 2021.– ISBN 9781665447645.– Pp. 281–284. ↑269

[10] Kim A., Park M., Lee D. H. *AI-IDS: Application of deep learning to real-time web intrusion detection* // IEEE Access.– 2020.– Vol. **8**.– Pp. 70245–70261. ↑269

[11] Lohrasbinasab I., Shahraki A., Taherkordi A., Jurcut A. D. *From statistical — to machine learning-based network traffic prediction* // Trans. Emerging Tel. Tech..– 2022.– Vol. **33**.– No. 4.– id. e4394.– 20 pp. ↑269

[12] Canola Garcia J. F., Blandon G. E. T. *A deep learning-based intrusion detection and preventation system for detecting and preventing denial-of-service attacks* // IEEE Access.– 2022.– Vol. **10**.– Pp. 83043–83060. ↑269

[13] Aslam M., Ye D., Tariq A., Asad M., Hanif M., Ndzi D., Chelloug S. A., Abd Elaziz M., Al-Qaness M. A. A., Jilani S. F. *Adaptive machine learning based distributed denial-of-services attacks detection and mitigation system for SDN-enabled IoT* // Sensors.– 2022.– Vol. **22**.– No. 7.– id. 2697.– 28 pp. ↑269

[14] Azman M. A., Marhusin M. F., Sulaiman R. *Machine learning-based technique to detect SQL injection attack* // Journal of Computer Science.– 2021.– Vol. **17**.– No. 3.– Pp. 296–303. ↑269

[15] Горюнов М. Н., Мацкевич А. Г., Рыболовлев Д. А. *Синтез модели машинного обучения для обнаружения компьютерных атак на основе набора данных CICIDS2017* // Труды Института системного программирования РАН.– 2020.– Т. **32**.– № 5.– С. 81–94. ↑270

[16] Ahmad I., Imran M., Qayyum A., Ramzan M. S., Alassafi M.Ȯ. *An optimized hybrid deep intrusion detection model (HD-IDM) for enhancing network security* // Mathematics.– 2023.– Vol. **11**.– No. 21.– id. 4501.– 24 pp. ↑270

[17] Moh M., Pininti S., Doddapaneni S., Moh T.-S. *Detecting web attacks using multi-stage log analysis* // *2016 IEEE 6th International Conference on Advanced Computing (IACC)* (Bhimavaram, India, 27–28 February 2016).– IEEE.– 2016.– ISBN 9781467382878.– Pp. 733–738. ↑271

[18] Ross K., Moh M., Moh T.-Sh., Yao J. *Multi-source data analysis and evaluation of machine learning techniques for SQL injection detection* // *ACMSE'18: Proceedings of the 2018 ACM Southeast Conference*, New York: ACM.– ISBN 978-1-4503-5696-1.– id. 1.– 8 pp. ↑271

[19] Lin Y.-D., Wang Z.-Y., Lin P.-Ch., Nguyen V.-L., Hwang R.-H., Lai Y.-Ch. *Multi-datasource machine learning in intrusion detection: Packet flows, system logs and host statistic* // Journal of Information Security and Applications.– 2022.– Vol. **68**.– id. 103248. ↑271

[20] Ju A., Guo Y., Ye Z., Li T., Ma J. *HeteMSD: A big data analytics framework for targeted cyber-attacks detection using heterogeneous multisource data* // Security and Communication Networks.– 2019.– Vol. **2019**.– id. 5483918.– 9 pp. ↑272

[21] Li R., Sun L. *Network security threat detection model based on large-scale multi-source data analysis and perception fusion* // Neural Computing and Applications.– 2025. ↑272

[22] Исаев С. В., Кононов Д. Д. *Исследование динамики и классификация атак на веб-сервисы корпоративной сети* // Сибирский аэрокосмический журнал.– 2022.– Т. **23**.– № 4.– С. 593–600. ↑274

[23] Кононов Д. Д. *Автоматизированная система подготовки обучающих наборов данных активности веб-сервисов на основе разнородных источников*, Свидетельство о регистрации программ для ЭВМ № 2024686983 от 13.11.2024.– 2024. ↑278

Поступила в редакцию 10.07.2025;
одобрена после рецензирования 16.07.2025;
принята к публикации 03.10.2025;
опубликована онлайн 27.11.2025.

Рекомендовал к публикации *Анд. В. Климов*

Информация об авторах:

Дмитрий Дмитриевич Кононов

научн. сотр. отдела Информационно-телекоммуникационных технологий Института вычислительного моделирования СО РАН. Научные интересы: кибербезопасность, анализ сетевого трафика, разработка защищенных веб-сервисов, машинное обучение

0000-0002-8757-5274

e-mail: ddk@icm.krasn.ru

Сергей Владиславович Исаев

к.т.н., доцент, зав. отделом Информационно-телекоммуникационных технологий Института вычислительного моделирования СО РАН. Научные интересы: телекоммуникационные системы и сети, защита информации, интернет-технологии, кибербезопасность, интеллектуальные системы и машинное обучение

0000-0002-6678-0084

e-mail: si@icm.krasn.ru

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*

PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article HARDWARE AND SOFTWARE FOR DISTRIBUTED AND SUPERCOMPUTER SYSTEMS

UDC 004.89+004.056
10.25209/2079-3316-2025-16-4-267-285

Using multilevel data sources to prepare training sets for cyberattack detection

Dmitry Dmitrievich **Kononov**[1], Sergey Vladislavovich **Isaev**[2]

[1, 2] Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, Krasnoyarsk, Russia
[1]✉*ddk@icm.krasn.ru*

Abstract. Network traffic analysis is an integral part of ensuring security in information and telecommunication systems. The use of machine learning provides modern approaches with higher detection rates for cyber threats.

A new approach for generating training datasets is proposed, which introduces a new aggregation unit "session", utilizes signature analysis and multi-level data sources, including heterogeneous ones. A list of requirements for the datasets is generated, which includes preserving the first packets of the connection, preserving hidden areas of the packets, extended information about traffic sources (country, autonomous system number ASN). The additional information will allow to detect attacks of the "hidden communication channel" type. Using the proposed approach, a software package for creating training datasets from multilevel sources at the L7, L4, L3 levels of the OSI model has been developed. In contrast to existing works, real data of network activity as well as long time intervals are used. The proposed approach allows to use the obtained training sets to create more effective methods of intrusion detection and prevention using machine learning techniques. (*In Russian*).

Key words and phrases: Internet, network security, cyber threats, network traffic analysis, datasets, machine learning

2020 *Mathematics Subject Classification*: 68M25; 68-11, 62N86

For citation: Dmitry D. Kononov, Sergey V. Isaev. *Using multilevel data sources to prepare training sets for cyberattack detection*. Program Systems: Theory and Applications, 2025, **16**:4(67), pp. 267–285. (*In Russ.*).
https://psta.psiras.ru/read/psta2025_4_267-285.pdf

© Kononov D. D., Isaev S. V. 2025

References

[1] S. A. Les'ko. "Models and methods of protecting web resources: a systematic review", *Cloud of Science*, **7**:3 (2020), pp. 577–610 (in Russian).

[2] L. Duan, J. Zhou, Y. Wu, W. Xu. "A novel and highly efficient botnet detection algorithm based on network traffic analysis of smart systems", *International Journal of Distributed Sensor Networks*, **18**:3 (2022), 17 pp.

[3] E. Papadogiannaki, S. Ioannidis. "A survey on encrypted network traffic analysis applications, techniques, and countermeasures", *ACM Computing Surveys*, **54**:6 (2021), id. 123, 35 pp.

[4] N. J. Singh, N. Hoque, K. R. Singh, D. K. Bhattacharyya. "Botnet-based IoT network traffic analysis using deep learning", *Security and Privacy*, **7**:2 (2024), id. e355, 40 pp.

[5] P. Zhang, W. Ma, S. Qian. "Cluster analysis of day-to-day traffic data in networks", *Transportation Research Part C: Emerging Technologies*, **144** (2022), id. 103882, 24 pp.

[6] W. Zhongsheng, W. Jianguo, Y. Sen, G. Jiaqiong. "Retracted: Traffic identification and traffic analysis based on support vector machine", *Concurrency and Computation: Practice and Experience*, **32**:2 (2020), id. e5292.

[7] L. Nie, D. Jiang, Z. Lv. "Modeling network traffic for traffic matrix estimation and anomaly detection based on Bayesian network in cloud computing networks", *Annals of Telecommunications*, **72** (2017), pp. 297–305.

[8] M. Abbasi, A. Shahraki, A. Taherkordi. "Deep learning for network traffic monitoring and analysis (NTMA): A survey", *Computer Communications*, **170** (2021), pp. 19–41.

[9] I. Kotenko, I. Saenko, A. Kribel, O. Lauta. "A technique for early detection of cyberattacks using the traffic self-similarity property and a statistical approach", *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (Valladolid, Spain, 10–12 March 2021), IEEE, 2021, ISBN 9781665447645, pp. 281–284.

[10] A. Kim, M. Park, D. H. Lee. "AI-IDS: Application of deep learning to real-time web intrusion detection", *IEEE Access*, **8** (2020), pp. 70245–70261.

[11] I. Lohrasbinasab, A. Shahraki, A. Taherkordi, A. D. Jurcut. "From statistical—to machine learning-based network traffic prediction", *Trans. Emerging Tel. Tech.*, **33**:4 (2022), id. e4394, 20 pp.

[12] J. F. Canola Garcia, G. E. T. Blandon. "A deep learning-based intrusion detection and preventation system for detecting and preventing denial-of-service attacks", *IEEE Access*, **10** (2022), pp. 83043–83060.

[13] M. Aslam, D. Ye, A. Tariq, M. Asad, M. Hanif, D. Ndzi, S. A. Chelloug, M. Abd Elaziz, M. A. A. Al-Qaness, S. F. Jilani. "Adaptive machine learning based distributed denial-of-services attacks detection and mitigation system for SDN-enabled IoT", *Sensors*, **22**:7 (2022), id. 2697, 28 pp.

[14] M. A. Azman, M. F. Marhusin, R. Sulaiman. "Machine learning-based technique to detect SQL injection attack", *Journal of Computer Science*, **17**:3 (2021), pp. 296–303.

[15] M. N. Goryunov, A. G. Mackevich, D. A. Rybolovlev. "Synthesis of a machine learning model for detecting computer attacks based on the CICIDS2017 dataset", *Trudy Instituta sistemnogo programmirovaniya RAN*, **32**:5 (2020), pp. 81–94 (in Russian).

[16] I. Ahmad, M. Imran, A. Qayyum, M. S. Ramzan, M.Ȯ. Alassafi. "An optimized hybrid deep intrusion detection model (HD-IDM) for enhancing network security", *Mathematics*, **11**:21 (2023), id. 4501, 24 pp.

[17] M. Moh, S. Pininti, S. Doddapaneni, T.-S. Moh. "Detecting web attacks using multi-stage log analysis", *2016 IEEE 6th International Conference on Advanced Computing (IACC)* (Bhimavaram, India, 27–28 February 2016), IEEE, 2016, ISBN 9781467382878, pp. 733–738.

[18] K. Ross, M. Moh, T.-Sh. Moh, J. Yao. "Multi-source data analysis and evaluation of machine learning techniques for SQL injection detection", *ACMSE'18: Proceedings of the 2018 ACM Southeast Conference*, ACM, New York, ISBN 978-1-4503-5696-1, id. 1, 8 pp.

[19] Y.-D. Lin, Z.-Y. Wang, P.-Ch. Lin, V.-L. Nguyen, R.-H. Hwang, Y.-Ch. Lai. "Multi-datasource machine learning in intrusion detection: Packet flows, system logs and host statistic", *Journal of Information Security and Applications*, **68** (2022), id. 103248.

[20] A. Ju, Y. Guo, Z. Ye, T. Li, J. Ma. "HeteMSD: A big data analytics framework for targeted cyber-attacks detection using heterogeneous multisource data", *Security and Communication Networks*, **2019** (2019), id. 5483918, 9 pp.

[21] R. Li, L. Sun. "Network security threat detection model based on large-scale multi-source data analysis and perception fusion", *Neural Computing and Applications*, 2025.

[22] S. V. Isaev, D. D. Kononov. "A study of dynamics and classification of attacks on corporate network web services", *Sibirskij aerokosmicheskij zhurnal*, **23**:4 (2022), pp. 593–600 (in Russian).

[23] D. D. Kononov. *An automated system for preparing training datasets of web service activity based on heterogeneous sources*, Svidetel'stvo o registracii programm dlya EVM No 2024686983 ot 13.11.2024, 2024 (in Russian).