УДК 004.41

ПРИМЕНЕНИЕ АНСАМБЛЕВЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ФИЛЬТРАЦИИ ЛОЖНЫХ СРАБАТЫВАНИЙ В SIEM-СИСТЕМАХ

А.Г. Бокун

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», г. Минск, Республика Беларусь

Аннотация. В статье рассмотрена проблема высокой доли ложных срабатываний в современных системах мониторинга событий информационной безопасности (SIEM). Во введении обоснована актуальность внедрения легковесных механизмов автоматизации для снижения нагрузки на операторов систем защиты. В основной части приведено описание типовой архитектуры ядра SIEM-системы и выделены ключевые этапы обработки данных, на которых возможно внедрение инструментов машинного обучения. Предложено решение на основе алгоритма «Случайный лес», сочетающее в себе устойчивость к переобучению и прозрачность принятия решений за счет оценки важности параметров. В заключении сделан вывод о преимуществе использования ансамблей решающих деревьев перед тяжеловесными нейросетевыми моделями при решении задач классификации инцидентов в локальных сетях.

Ключевые слова: информационная безопасность, SIEM-система, ложные срабатывания, машинное обучение, классификация событий, случайный лес, корреляция событий.

APPLICATION OF ENSEMBLE MACHINE LEARNING METHODS FOR FALSE POSITIVE FILTERING IN SIEM

A.G. Bokun

Educational Institution "Belarusian State University of Informatics and Radioelectronics", Minsk, Republic of Belarus

Abstract. The article addresses the issue of high false positive rates in modern Security Information and Event Management (SIEM) systems. The introduction substantiates the relevance of implementing lightweight automation mechanisms to reduce the workload on security operations center (SOC) analysts. The main body describes the typical architecture of a SIEM core and identifies key data processing stages where machine learning tools can be effectively integrated. The author proposes a solution based on the Random Forest algorithm, which combines resistance to overfitting with decision-making transparency through feature importance evaluation. The conclusion emphasizes the advantages of using decision tree ensembles over heavyweight neural network models for incident classification tasks within local networks.

Keywords: information security, SIEM, false positives, machine learning, event classification, random forest, event correlation.

Введение

Вопрос защиты информации является сегодня одним из приоритетных направлений для всех хозяйствующих субъектов. Разные экономические агенты обладают разным количеством данных, которые необходимо защищать, и разными возможностями эту защиту имплементировать. Если малому бизнесу достаточно просто соблюдать базовые правила информационной безопасности, то большим предприятиям с разветвленной внутренней информационной сетью не обойтись без комплексных решений вроде SIEM систем.

Архитектура таких систем хорошо известна и прошла проверку временем, однако она имеет ряд недостатков, которые можно решить с помощью современных подходов. Одним из таких недостатков являются ложные срабатывания. Цель работы предложить легковесный механизм для решения проблемы обнаружения ложных событий информационной безопасности в SIEM системах на основе алгоритмов машинного обучения.

Основная часть

В начале считаем необходимым кратко рассмотреть общую архитектуру современных SIEM систем (их ядра) для определения наиболее оптимального места, где можно имплементировать решение.

Для сбора данных такие системы используют два основных подхода: агентный и безагентный сбор. По сути эти методы отличаются необходимостью установки дополнительного ПО на целевой узел: агентные системы предполагают установку приложения-агента на узел, который активно собирает данные и отправляет их в аналитические центры системы. Безагентные системы полагаются на пассивный сбор данных через протоколы вроде syslog и чтение обобщенных файлов (если речь идет про Windows).

После непосредственного сбора данных об узле информация пересылается в аналитическое ядро SIEM-системы (далее просто ядро), где проходят несколько этапов обработки. Обобщая эти этапы, их можно обозначить следующим образом.

1. Нормализация – процесс приведения разрозненных данных к общей форме, которая описана правилами нормализации. Правила определяют какие поля должны быть в конечной структуре и как с ними далее работать.

2. Агрегация – процесс объединения похожих событий в одно. Здесь система сверяется с правилами агрегации и приводит группирует однородные данные в единые сущности.

3. Корреляция – процесс выявления инцидентов безопасности на основе некоторого контекста событий. Система наработала окно агрегированных событий и может проверить его на наличие событий, образующих инцидент безопасности (согласно правилам корреляции). После чего оператор оповещается о созданном событии.

Помимо трех основных этапов иногда встречаются дополнительные, в их число могут входить следующие.

1. Обогащение – процесс добавление дополнительных метаданных скоррелированому инциденту, как правило используются заранее прописанные в правилах обогащения источники дополнительной информации.

2. Добавление контекста – система помогает рассматривать инциденты не как отдельные происшествия, а ставить их в общий аналитический контекст.

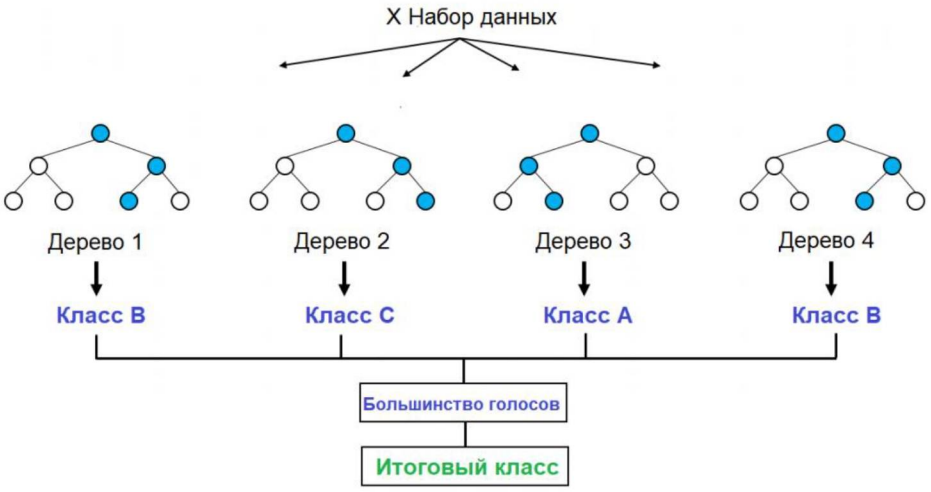
Современные SIEM-системы обладают широким перечнем интеллектуальных инструментов, однако основной расчет все еще делается на корректное написание правил. Обнаружение инцидентов информационной задачи – крайне комплексная и персонифицированная задача, условия которой меняются от одной сети к другой. Невозможно дать одно универсальное решение, поэтому специалисты по информационной безопасности работают отдельно с каждой сетью, учитывают их специфику и с помощью правил затачивают SIEM под конкретную задачу. Такая гибкость является несомненным преимуществом, однако тут же появляется и проблема – правила нельзя написать идеально, поэтому время от времени происходят ложные срабатывания – система реагирует на легальную активность, которая тем не менее попала под ее правила корреляции.

Данная проблема стоит крайне остро в рутине оператора SIEM-систем, так как специалист вынужден тратить свое время на поиск и закрытие ложных срабатываний вместо того чтобы заниматься мониторингом. Полноценной автоматизации решения этой проблемы на рынке все еще нет: некоторые системы (Wazuh) оставляют все на откуп оператору, некоторые (MaxPatrol, KUMA) предлагают довольно тяжеловесные решения на основе полноценных нейронных сетей, которые впрочем больше подходят под решение более глобальных и общих задач. В связи с этим предлагаем свое решение на основе легких моделей машинного обучения.

Определим тип решаемой задачи. Имеется некоторая фиксированная структура данных, данная структура заполняется однородной информацией. Необходимо разделить объекты этой структуры на два класса, для текущей задачи это «ложное срабатывание» и «не ложное срабатывание». То есть это задача классификации с неизменной

структурой данных. Для решения данного типа задач обычно используются модели на основе деревьев принятий решений и градиентный бустинг.

Однако стоит отметить, что важным требованием к модели является прозрачность принятия решений. То есть оператор должен точно понимать, почему модель выбрала тот или иной вариант, чтобы принять конечное информированное решение. Такой набор требований делает модели на основе решающих деревьев более применимыми. Конкретно мы предлагаем модель, которая называется «случайный лес». Случайный лес – это модель машинного обучения, которая использует ансамбль случайных деревьев для создания предсказаний. Решением всего ансамбля будет среднее решение деревьев, входящих в него. Упрощенная схема работы случайного леса представлен на рисунке.



Общая схема работы модели случайного леса
The general scheme of the random forest model

В модели случайного леса есть две политики принятия решений: жесткое и мягкое голосование:

– жесткое голосование – каждый классификатор (решающее дерево) делает предсказание, а предсказанием ансамбля (случайного леса) будет решение, за которое проголосовало большинство;

– мягкое голосование – каждый классификатор предоставляет процентовку с вероятностями принятия того или иного решения, в этом случае суммирующий механизм ансамбля при принятии окончательного решения учитывает не сами голоса моделей, а именно полученные вероятности.

Архитектурно случайный лес крайне устойчив к переобучению, что важно, так как обучающая выборка будет довольно однородной (так как обучение будет происходить на данных конкретной сети). Также важным аспектом является то, что оно может продемонстрировать

коэффициент важности параметров, это делает решения более прозрачными и помогает оператору при аналитике инцидентов.

Заключение

В итоге можно сказать, что предложенный подход решает проблему перегрузки операторов SIEM-систем ложными срабатываниями за счет внедрения легковесного механизма классификации на базе машинного обучения. Выбор модели «Случайный лес» обоснован ее устойчивостью к переобучению и способностью ранжировать важность параметров, что обеспечивает необходимую прозрачность принятия решений. В отличие от тяжеловесных нейросетей, такой алгоритм легче адаптируется к специфике конкретной сети и позволяет специалистам сосредоточиться на анализе реальных инцидентов.

Список использованных источников

1. Ахикян А.И., Данилюк С.С. (2024) Алгоритм машинного обучения адаптивный случайный лес и его применение. *Вестник науки.* (75), 1393-1402.

2. Цымбал Ф.А. (2022) Управление инцидентами безопасности и событиями (SIEM). *Столыпинский вестник.* (4), 2121-2129.

References

1. Akhikyan, A. I., & Danilyuk, S. S. (2024). Adaptive Random Forest Machine Learning Algorithm and Its Application. *Vesnik Nauki*, (75), 1393-1402.

2. Tsymbal, F. A. (2022). Security Information and Event Management (SIEM). *Stolypinskiy Vestnik*, (4), 2121-2129.

Сведения об авторе

Бокун А.Г., магистрант, ассистент кафедры информатики, учреждение обраования «Белорусский государственный университет информатики и радиоэлектроники», a.bokun@bsuir.by.

Information about the author

Bokun A., Master Student, Assistant at the Informatics Department, Educational Institution "Belarusian State University of Informatics and Radioelectronics", a.bokun@bsuir.by.