DOI 10.36622/1682-7813.2025.28.2.009
УДК 621.372.004

АТАКУЕМЫЕ СРЕДСТВА МАШИННОГО ОБУЧЕНИЯ: АВТОМАТИЗАЦИЯ РИСК-АНАЛИЗА ПРОЦЕССОВ РЕАЛИЗАЦИИ СЦЕНАРИЕВ

Д.А. Нархов, Д.В. Кульшин, К.В. Козина, М.А. Неменущий

В статье рассматривается методология построения информационной системы для оценки рисков атак на средства машинного обучения. Исследованы аналоги данной системы. Для каждого из аналогов выделены их функциональные особенности и их недостатки, относительно применения возможностей данных ресурсов к атакуемым средствам машинного обучения. Автоматизация риск-анализа процессов реализации сценариев представлена в виде прототипа, включающего модули сбора информации об уязвимостях, сопоставления их с техниками MITRE ATLAS, вычисления риска и визуализации результатов. Представлена схема взаимодействия данных модулей. Предложены направления дальнейшего развития системы. Разработанное решение позволяет систематизировать информацию о потенциальных угрозах для автоматизированных систем, сократить время риск-анализа в сравнении с ручными методами, а также обеспечивает гибкость при обновлении базы данных об уязвимостях и сценариях атак.

Ключевые слова: средства машинного обучения, оценка рисков, сценарии атак, уязвимости.

Введение

Современное киберпространство характеризуется высокой динамикой возникновения новых угроз. Поэтому для обеспечения безопасности требуется постоянный мониторинг угроз, их оценка и оперативное реагирование. .

Технологии машинного обучения активно внедряются в различные сферы человеческой деятельности, включая безопасность, медицину, финансы и автоматизацию производственных процессов [1]. Однако их широкое внедрение сопровождается ростом числа атак, направленных на эксплуатацию уязвимостей алгоритмов и их реализации.

Исходя из вышесказанного, одной из важных задач обеспечения информационной безопасности является необходимость систематизации знаний об атаках на средства машинного обучения, а также создание инструмента для построения сценариев и анализа рисков.

Актуальность темы исследования обусловлена:

– стремительным расширением зоны использования средств машинного обучения.

– необходимостью программно-алгоритмической реализации информационной системы, обеспечивающей агрегацию данных об уязвимостях средств машинного обучения и оценку риска.

Исследование аналогов

В рамках исследования был проведен анализ существующих аналогов на предмет выявления их структурных и функциональных особенностей.

Информационный ресурс «Сервис моделирования кибератак по матрице MITRE ATT&CK» [2] разработанный «Код безопасности» содержит следующие особенности:

1) представление на русском языке MITRE ATT&CK,

2) перечень отечественных СрЗИ, которые используются для противодействия техникам,

3) перечень часто реализуемых техник,

4) сегментация техник в соответствии с отраслью деятельности,

5) визуализацию смежных техник для выбранной техники.

Недостатками ресурса «Сервис моделирования кибератак по матрице MITRE ATT&CK» являются:

1) отсутствие техник, направленных на средства машинного обучения,

2) отсутствие возможности построения различных сценариев, помимо существующих цепочек,

3) отсутствие связи с уязвимостями.

Информационный ресурс «Какие техники MITRE ATT&CK выявляют продукты Positive Technologies» [3] созданный «Positive Technologies» обладает следующими особенностями:

1) представление на русском языке MITRE ATT&CK,

2) перечень отечественных СрЗИ, которые используются для противодействия техникам,

3) подборки часто реализуемых техник,

4) способы обнаружения и меры противодействия техникам.

Недостатками данного ресурса являются:

1) отсутствие техник, направленных на средства машинного обучения,

2) отсутствие возможности построения различных сценариев, помимо существующих цепочек,

3) отсутствие связи с уязвимостями.

Таким образом, целью исследования является разработка информационного ресурса, обеспечивающего автоматизированный риск-анализ сценариев атак на средства машинного обучения с учётом уязвимостей.

Для достижения данной цели необходимо решить следующие задачи:

1) разработать алгоритмы функционирования и взаимодействия модулей, обеспечивающих агрегацию данных об уязвимостях и оценку рисков,

2) создать программную реализацию модулей, обеспечивающих агрегацию данных об уязвимостях и оценку рисков.

Алгоритмическая реализация модулей информационной системы

Разработанная информационная система реализована в виде модульной архитектуры. Основными модулями являются:

1) модуль агрегации данных – отвечает за сбор и хранение данных об уязвимостях средств машинного обучения,

2) модуль оценки рисков – отвечает за расчет риска сценариев атак на средства машинного обучения,

3) модуль классификации уязвимостей – предсказывает технику MITRE ATLAS [4] по описанию уязвимости и CWE,

4) модуль визуализации – предназначен для отображения и визуализации данных.

Модуль агрегации данных отвечает за централизованный сбор, нормализацию и хранение сведений об уязвимостях из различных источников. Это позволяет обеспечить обмен данными для последующих процессов классификации и оценки рисков.

Источниками данных являются:

1) NIST NVD – база CVE с подробными описаниями и метриками,

2) БДУ ФСТЭК – отечественный источник сведений об уязвимостях,

3) Huntr – платформа для отправки сообщений об уязвимостях в области машинного обучения.

Задачи, решаемые с помощью данного модуля:

1) сбор данных, с поддержкой запросов к API,

2) обработка отдельных форматов. JSON для NIST, XML для ФСТЭК БДУ, HTML-страницы для Huntr,

3) нормализация и дедупликация. Приведение сведений об уязвимостях в единый формат и объедение дубликатов,

4) загрузка в СУБД для поиска и обработки.

Традиционные системы оценки уязвимостей, такие как CVSS недостаточны при решении проблем, создаваемых в системах, использующих машинное обучение. В связи с этим фонд OWASP разработал систему AIVSS [5], комплексную структуру, адаптированную к рискам безопасности машинного обучения.

Также, необходимым полем является метка техники MITRE ATLAS, к которой относится данная уязвимость.

Данный модуль представляет собой нейросетевую модель, которая на вход получает текстовое описание уязвимости, а выходными данными является метка: к какой технике относится данная уязвимость.

Цель модуля состоит в автоматизации процесса классификации уязвимостей.

Основные компоненты модуля:

1) предобработки текста: очистка от лишних символов, разметка ключевых сущностей,

2) векторизации: преобразование входного текста в числовое представление, с помощью предобученной языковой модели BERT,

3) классификатор техник: нейросетевая архитектура, обученная на размеченном датасете «описание уязвимости – техника ATLAS».

Архитектура данного модуля состоит из следующих основных элементов:

1) входной слой, в котором происходит токенизация с учетом терминов, таких как CVE, CWE, названия средств машинного обучения,

2) контекстный энкодер, который представляет собой предобученную трансформер-модель, дообученную на множестве уязвимостей.

Предложенный нейросетевой модуль решает задачу быстрого и надежного сопоставления описаний уязвимостей с техникой MITRE ATLAS.

На рис. 1 представлен график зависимости точности модели.

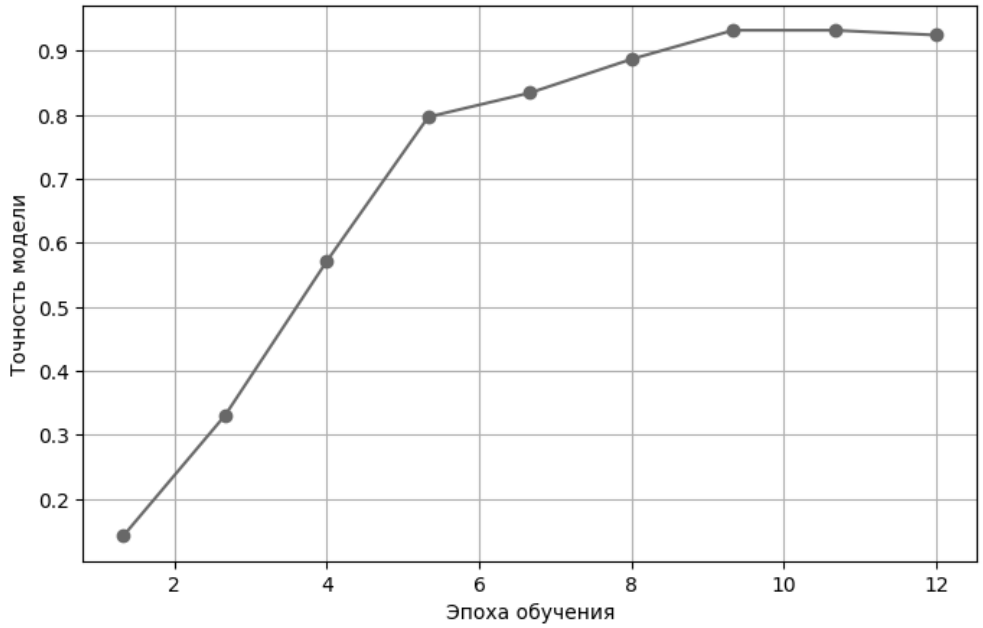

Рис. 1. График точности модели

Ручной метод оценки рисков перестает быть эффективным при обработке сотен уязвимостей. В связи с этим, становится важной задача автоматизации данного процесса, а также создание модуля оценки риска, который получает на вход множество уязвимостей и выдает значения риска.

Целью модуля является автоматический расчет риска, для заданного множества уязвимостей.

Модуль агрегации данных обеспечивает непрерывный сбор и унификацию всей входящей информации. На основе этих данных и строится работа модуля оценки риска.

Для реализации поставленной цели модуль выполняет следующие задачи:

1) подготовка и нормализация данных: приведение метрик CVSS, EPSS, AIVSS к единой шкале,

2) расчет критичности уязвимостей в соответствии с методом, описанном в исследовании [6],

3) вычисление значения риска.

Модуль визуализации предназначен для графического отображения сценариев атак и представления информации о связанных с ними рисках. Данный модуль обеспечивает наглядность и удобство восприятия данных. Основная цель модуля – преобразовать результаты работы модулей агрегации и оценки рисков в понятную и информативную визуальную форму.

Задачи, решаемые модулем:

1) визуализация групп сценариев атак: отображение цепочек техник, в виде графов, где узлы – техники, а рёбра вероятности эксплуатации уязвимостей,

2) отображение рисков всевозможных сценариев,

3) экспорт результатов. Есть возможность экспортировать данные по оценке рисков в формат xlsx,

4) отображение списка уязвимостей с возможностью поиска.

Информационной системы имеет клиент-серверную архитектуру. На сервере хранятся и обрабатываются данные, на клиенте происходит отображение данных и отправление запросов серверу. Схема взаимодействия модулей представлена на рис. 2.

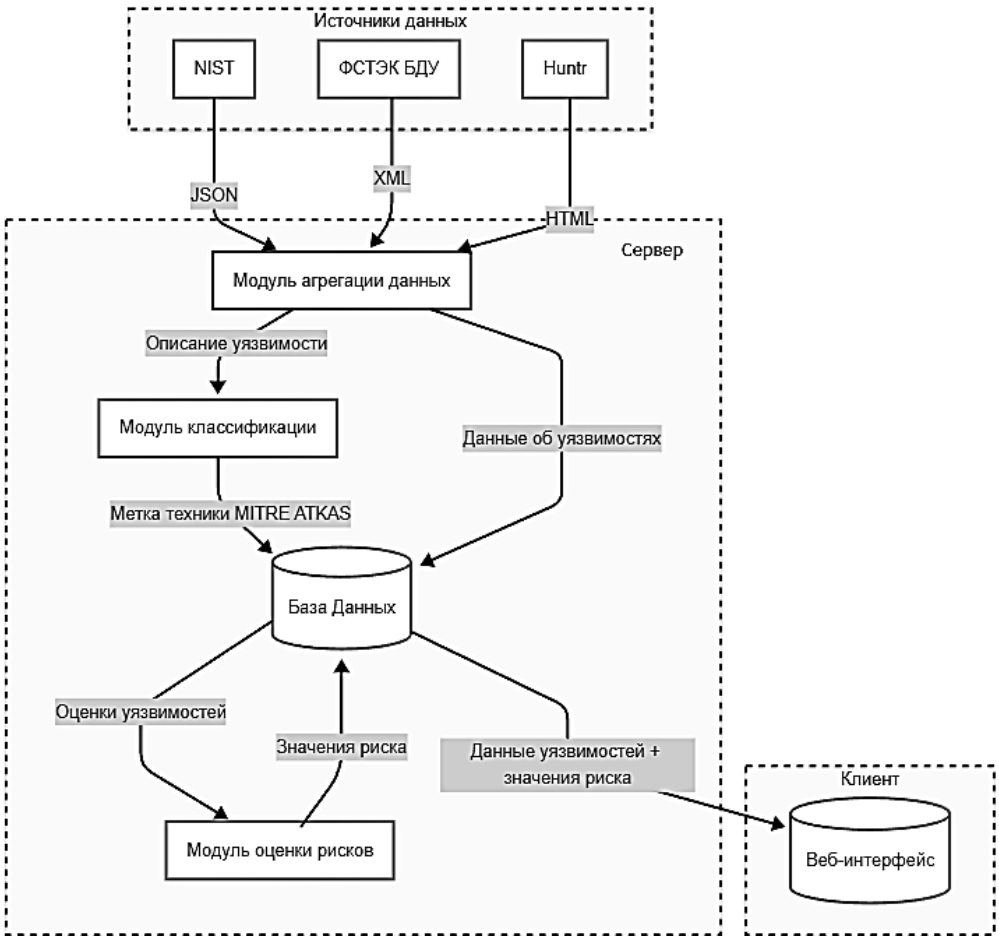

Рис. 2. Схема взаимодействия модулей

Программная реализация

Серверная часть реализована с помощью FastAPI, что обеспечивает высокую производительность. С использованием языка программирования Python. Python является простым и универсальным благодаря множеству различных библиотек, одной из которых является requests. Данная библиотека предоставляет простой интерфейс для работы HTTP-запросами. С её помощью создан модуль агрегации данных. Веб-интерфейс разработан с помощью React.

Запросы делаются с помощью библиотеки axios.

На рис. 3 приведена начальная страница со списком представленных на русском языке техник из MITRE ATLAS, с возможностью загрузить список сценариев содержащихся в json файле в виде последовательностей техник.

На рис. 4 представлена страница графа. Кнопка «разбить на сценарии» выделяет изначальные сценарии с возможностью переключаться между ними.

Кнопка «разбить по длине» позволяет рассмотреть микромодели, которые состоят из разного количества техник.

Рис. 3. Скриншот главной страницы информационного ресурса

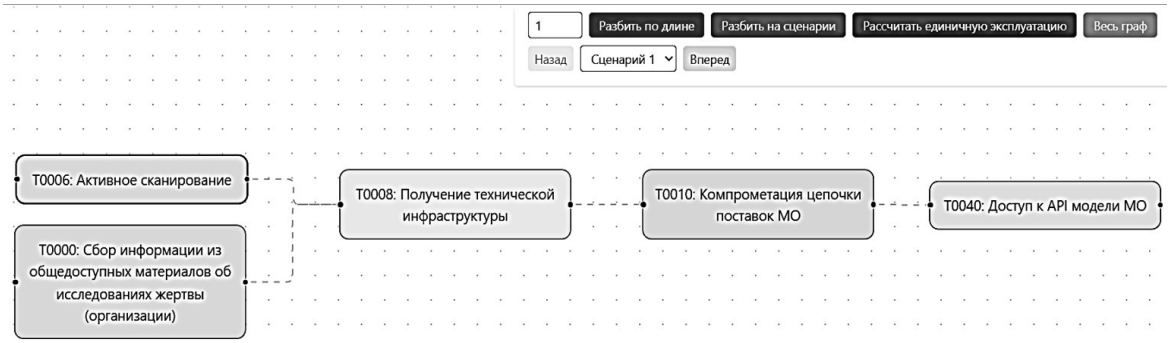

Рис. 4. Скриншот страницы с построенным графом сценария

Идея состоит в том, что при рассмотрении двух изначальных сценариев, а именно: T0006-T0008, T0000-T0010-T0040, появляется возможность рассчитать риск не только этих сценариев, но и сценариев T0006-T0008-T0010-T0040,T0000-T0008 и так далее.

Для полученных множеств уязвимостей производится расчет рисков рисков. В данном случае мы предполагаем, что одна техника эксплуатирует одну уязвимость, поэтому происходит перебор всех конкретных сценариев, в виде «Техника 1 (уязвимость 1) – Техника 2 (уязвимость 2) - …». Для каждого из которого становится возможным найти значение риска данного сценария.

Заключение

Результатом данного исследования является информационная система, автоматизирующая риск-анализ процессов реализации атак на средства машинного обучения.

Описаны цели и задачи модулей данной системы.

Представлена схема взаимодействия модулей системы и интерфейс для работы с данной системой.

Новизна полученных результатов состоит в том, что для оценки уязвимостей помимо CVSS и EPSS используется система AIVSS. В отличии от аналогов используются техники MITRE ATLAS, так как он направлена на системы, связанные с машинным обучением. Реализовано построение сценариев, со связанными уязвимостями.

Для дальнейшего развития системы целесообразно добавить возможность запускать модуль агрегации по расписанию, а также добавить следующие модули:

1) модуль генерации мер противодействия для данных сценариев. Такой модуль может быть построен на основе машинного обучения;

2) модуль многократных эксплуатации, который рассчитывает риск при условии, что одна техника эксплуатирует множество уязвимостей;

3) модуль авторизации и аутентификации.

Описанный выше подход не учитывает наличие эксплоита или зафиксированной атаки, в связи с этим необходимо добавить модуль, который, при наличии эксплоита, нужно дополнительно учитывает временные метки CVSS, и осуществить более глубокий анализ EPSS, так как он показывает вероятность эксплуатации уязвимости.

Список литературы

1. AI for Everything URL: https://www.routledge.com/AI-for-Everything/book-series/ (дата обращения: 20.04.25).

2. Сервис моделирования кибератак по матрице MITRE ATT&CK URL: https://mitre.securitycode.ru/ (дата обращения: 20.04.25).

3. Какие техники MITRE ATT&CK выявляют продукты Positive Technologies URL: https://mitre.ptsecurity.com/ (дата обращения: 20.04.25).

4. MITRE ATLAS URL: https://atlas.mitre.org/ (дата обращения: 20.04.25).

5. OWASP Artificial Intelligence Vulnerability Scoring System URL: https://owasp.org/www-project-artificial-intelligence-vulnerability-scoring-system/ (дата обращения: 20.04.25).

6. Остапенко А.А. Оценка критичности уязвимостей с использованием данных множества риск калькуляторов / А.А. Остапенко // Информация и безопасность. 2024. Т.27.Вып. 4. С. 543-552.

Воронежский государственный технический университет
Voronezh State Technical University

Поступила в редакцию 25.04.25

Информация об авторах

Нархов Дмитрий Андреевич – аспирант, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Кульшин Дмитрий Вячеславович – студент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Козина Ксения Владимировна – студентка Воронежский государственный технический университет e-mail: alexanderostapenkoias@gmail.com

Неменущий Максим Дмитриевич – студент, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

ATTACKABLE MACHINE LEARNING TOOLS: AUTOMATION OF RISK ANALYSIS OF SCENARIO REALIZATION PROCESSES

D.A. Narhov, D.M. Kulshin, K.V. Kozina, M.D. Nemenushchiy

The article deals with the methodology of building an information system for assessing the risks of attacks on machine learning tools. The analogs of this system are studied. For each of the analogs their functional features and their disadvantages are highlighted in relation to the application of the capabilities of these resources to the attacked machine learning tools. Automation of risk-analysis of scenario realization processes is presented in the form of a prototype, including modules for collecting information about vulnerabilities, comparing them with MITRE ATLAS techniques, calculating risk and visualizing results. The scheme of interaction of these modules is presented. Directions for further development of the system are proposed. The developed solution allows to systematize information about potential threats to automated systems, reduce the time of risk analysis in comparison with manual methods, and provides flexibility in updating the database of vulnerabilities and attack scenarios.

Keywords: machine learning tools, risk assessment, attack scenarios, vulnerabilities.

Submitted 25.04.25

Information about the authors

Dmitry A. Narhov – postgraduate student Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com

Dmitry V. Kulshin – student Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com

Ksenia V. Kozina — tudent Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com

Maksim D. Nemenushchiy – student Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com