МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра физики открытых систем

Исследование и разработка алгоритмов машинного обучения для анализа и предсказания поведения злоумышленников в сети

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса __2241__ группы
Направления__09.04.02_«Информационные системы и технологии»_____
код и наименование направления
института физики_____
наименование факультета, института, колледжа
Мишина Александра Сергеевича_____
фамилия, имя, отчество

Научный руководитель
профессор, д.ф.-м.н., профессор _____ О.И. Москаленко
должность, уч. степень, уч. звание дата, подпись инициалы, фамилия

Заведующий кафедрой
профессор, д.ф.-м.н., профессор _____ А.А. Короновский
должность, уч. степень, уч. звание дата, подпись инициалы, фамилия

Саратов 2025 год

Введение

В современном мире кибербезопасности проблема прогнозирования и обнаружения вредоносной активности приобретает все большую актуальность в связи с непрерывным ростом сложности и адаптивности кибератак. Традиционные подходы к защите, зачастую основанные на сигнатурном анализе и статических правилах, показывают ограниченную эффективность перед лицом современных многоэтапных атак и целевых кампаний (APT). В этих условиях применение методов машинного обучения для анализа и предсказания поведения злоумышленников становится ключевым направлением исследований, открывающим возможности для создания проактивных и адаптивных систем защиты, способных идентифицировать ранее неизвестные угрозы и предвидеть действия атакующих [1-3].

С развитием технологий машинного обучения был предложен широкий спектр подходов к обнаружению и прогнозированию киберугроз, включающих как классические статистические методы и алгоритмы классификации, так и передовые архитектуры глубокого обучения, такие как рекуррентные нейронные сети (LSTM) и трансформеры. Каждый из этих методов обладает своими особенностями, преимуществами и ограничениями. В данной работе основное внимание уделяется комплексному исследованию и разработке алгоритмов, охватывающих как методы векторизации текста (TF-IDF) для анализа вредоносного ПО, так и продвинутые нейросетевые модели для прогнозирования последовательностей действий злоумышленников, с особым акцентом на анализ временных рядов, поскольку современные кибератаки часто представляют собой сложные, развивающиеся во времени процессы [4-7].

Несмотря на достигнутые успехи в применении машинного обучения для задач кибербезопасности, ряд фундаментальных проблем остается нерешенным или требует дальнейшего изучения. К ним относятся: экстремальная несбалансированность классов в реальных данных, где вредоносная активность составляет ничтожно малую долю событий (менее

0.01%); проблема «дрейфа концепций», обусловленная постоянной эволюцией тактик и техник злоумышленников; недостаточная интерпретируемость сложных моделей, что затрудняет их практическое применение аналитиками центров мониторинга (SOC); а также сложности интеграции разработанных алгоритмов с существующими системами обеспечения безопасности [8-9].

Целью данной магистерской работы является исследование и разработка алгоритмов машинного обучения для анализа и предсказания поведения злоумышленников в сети, направленных на эффективное решение вышеперечисленных проблем. В рамках работы проводится анализ и сравнение различных классов моделей, от статистических методов до глубоких нейронных сетей, предлагаются подходы к повышению их эффективности и устойчивости в условиях реальных данных кибербезопасности. Особое внимание уделяется методам борьбы с несбалансированностью данных, адаптации к дрейфу концепций, обеспечению интерпретируемости моделей и их потенциальной интеграции с SIEM-системами. Для реализации практической части исследования использованы язык программирования Python и популярные библиотеки машинного обучения, такие как scikit-learn, TensorFlow и PyTorch, а также синтетические и публичные наборы данных, моделирующие сценарии кибератак.

Основное содержание работы

Магистерская работа состоит из введения, десяти глав и заключения. Она содержит 83 страницы машинописного текста, приведённый список литературы включает 62 наименования.

В первой главе работы рассматривается применение метода векторизации TF-IDF для обнаружения новых типов вредоносного программного обеспечения. Анализируется роль TF-IDF в выявлении значимых лексических особенностей вредоносных программ, обсуждаются его ограничения, такие как отсутствие учета семантических связей и влияние разреженности данных. Приводятся преимущества TF-IDF, включая

интерпретируемость и меньшую ресурсоемкость по сравнению с нейросетевыми подходами. Демонстрируется научно-практический пример кода, реализующий обнаружение вредоносного ПО с использованием TF-IDF и логистической регрессии.

Во второй главе работы исследуются метрики для оценки качества моделей в задачах кибербезопасности, такие как точность (precision), полнота (recall) и F-мера. Подчеркивается важность этих метрик для выбора оптимального алгоритма и приводятся примеры их практического применения. Демонстрируется научно-практический пример кода для расчета указанных метрик при обнаружении вредоносного ПО с использованием модели случайного леса.

Третья и четвертая глава работы посвящена формулировке основных принципов метода моделирования среды кибератак. Проводится обзор существующих подходов, включая имитационное моделирование, методы машинного обучения и гибридные подходы. Анализируются ключевые этапы моделирования: анализ уязвимостей, создание комплексной модели среды, симуляция сценариев атак и прогнозирование поведения злоумышленников. Выявляются и описываются проблемы применения существующих методов в условиях появления новейших типов вредоносного ПО и изменяющихся методов кибератак, включая влияние человеческого фактора и недостаток достоверных данных. Предлагаются пути совершенствования методов моделирования за счет интеграции искусственного интеллекта, технологий больших данных, облачных вычислений и повышения адаптивности моделей.

В пятой главе работы детально рассматриваются нейронные сети типа «трансформер» для анализа временных рядов в задачах кибербезопасности. Вводится понятие временных рядов и их специфики в данной области. Описывается архитектура трансформеров, включая механизм внимания и позиционное кодирование, а также их математические основы. Исследуются адаптации трансформеров для анализа временных рядов, такие как Temporal Fusion Transformers (TFT) и Informer [10]. Приводятся примеры практического

применения трансформеров для обнаружения аномалий, прогнозирования поведения злоумышленников и классификации типов атак. Обсуждаются перспективы развития трансформеров, включая гибридные модели и методы снижения вычислительных затрат [11]. Представлен научно-практический пример кода для генерации синтетических данных, имитирующих сетевой трафик с аномалиями.

Шестая глава работы посвящена разработке гибридной модели для обнаружения аномалий, сочетающей сверточные нейронные сети (CNN), рекуррентные слои (GRU) и автоэнкодеры. Обосновывается выбор такой архитектуры для эффективного извлечения локальных признаков, моделирования временных зависимостей и выделения аномального поведения. Представлен научно-практический пример кода на PyTorch, реализующий данную гибридную модель и демонстрирующий ее применение для прогнозирования сетевого трафика.

Седьмая глава работы фокусируется на генерации синтетических датасетов для анализа поведения злоумышленников. Обсуждаются различные подходы к созданию таких данных, включая моделирование на основе статистических распределений, использование генеративных моделей (GAN, VAE) и симуляцию сетевого трафика [12]. Рассматриваются методы создания контролируемых наборов данных, такие как сбор реальных данных с разметкой и использование публичных датасетов. Представлен научно-практический пример кода на Python для генерации синтетических данных, имитирующих сетевой трафик с сезонностью, трендом и аномалиями.

Восьмая глава работы посвящена разработке моделей прогнозирования на основе временных рядов. Подчеркивается переход от классических статистических методов к гибридным архитектурам глубокого обучения. Обсуждаются проблемы, связанные с природой данных безопасности, такие как мультимасштабность, контекстуальная зависимость, адверсарный характер данных и необходимость эффективного отделения сигнала от шума. Рассматриваются архитектурные инновации, включая

сверточные LSTM, трансформеры и иерархические модели [13-14]. Отмечается важность тщательной подготовки данных, включая синхронизацию временных меток, борьбу с дисбалансом классов и семантическое обогащение. Особое внимание уделяется проблеме "дрейфа концепций" и методам адаптации моделей, таким как онлайн-обучение с регуляризацией Elastic Weight Consolidation (EWC), адаптивные ансамбли и детекторы сдвига распределения [15]. Обсуждается важность интерпретируемости моделей с помощью методов SHAP и LIME для временных рядов. Рассматриваются научно-практические примеры кода для прогнозирования с использованием LSTM.

Девятая и десятая главы работы посвящены новейшим методам оценки моделей и их оптимизации для баланса точности и полноты. Рассматриваются адаптированные подходы к анализу ROC-AUC, Precision-Recall кривых и визуализации матрицы ошибок в условиях несбалансированных данных и эволюционирующих угроз. Описываются методы адаптивной настройки порогов классификации, учитывающие уровень киберугроз, ресурсы SOC и тактики MITRE ATT&CK [16-17]. Исследуются методы оптимизации гиперпараметров моделей для максимизации F-меры, включая байесовскую оптимизацию, генетические алгоритмы и adversarial гиперпараметрический поиск. Приводятся научно-практические примеры кода для визуализации метрик и оптимизации модели по F-мере.

Заключение

В рамках данной магистерской работы было проведено комплексное исследование и разработка алгоритмов машинного обучения, направленных на анализ и предсказание поведения злоумышленников в сети. Работа охватила широкий спектр современных подходов, начиная от классических методов векторизации текста, таких как TF-IDF, для задач обнаружения вредоносного программного обеспечения, и заканчивая продвинутыми архитектурами глубокого обучения, включая LSTM и трансформеры, для

анализа и прогнозирования временных последовательностей действий атакующих.

В ходе исследования была подтверждена эффективность метода TF-IDF для извлечения значимых лексических признаков из текстовых данных в контексте обнаружения вредоносного ПО. Несмотря на появление более сложных нейросетевых подходов, TF-IDF сохраняет свою актуальность благодаря интерпретируемости получаемых результатов и меньшей вычислительной сложности. При этом была продемонстрирована логическая преемственность и возможность интеграции TF-IDF с нейросетевыми моделями, где векторные представления могут служить входными данными для более сложных архитектур, объединяя преимущества обоих подходов.

Особое внимание было уделено нейронным сетям типа «трансформер» и их адаптации для анализа временных рядов в задачах кибербезопасности. Исследование показало, что трансформеры, благодаря механизму внимания, способны эффективно улавливать долгосрочные зависимости в данных и обрабатывать сложные последовательности событий, что делает их перспективным инструментом для прогнозирования кибератак. Были рассмотрены такие архитектуры, как Temporal Fusion Transformers и Informer.

Для решения задачи обнаружения аномалий в сетевом трафике была разработана и исследована гибридная модель, сочетающая сверточные нейронные сети (CNN) для извлечения локальных признаков, рекуррентные слои (GRU) для моделирования временных зависимостей и автоэнкодеры для выделения аномального поведения. Эксперименты на синтетических данных подтвердили высокую эффективность предложенного гибридного подхода.

В работе также была рассмотрена проблема генерации синтетических датасетов, что является критически важным для обучения и тестирования моделей в условиях ограниченного доступа к реальным данным о кибератаках. Предложены методы создания контролируемых наборов данных, имитирующих различные сценарии атак.

Ключевой аспект исследования составила проблема «дрейфа концепций», обусловленная постоянной эволюцией тактик злоумышленников. Были проанализированы и предложены методы адаптации моделей, включая онлайн-обучение с регуляризацией Elastic Weight Consolidation, адаптивные ансамбли и детекторы сдвига распределения, позволяющие моделям сохранять высокую точность в изменяющейся среде.

Для повышения практической применимости разработанных моделей была подчеркнута важность их интерпретируемости. Рассмотрены методы SHAP и LIME, адаптированные для временных рядов, и показано, как они могут способствовать пониманию решений моделей аналитиками безопасности.

Наконец, были исследованы новейшие подходы к оценке моделей прогнозирования кибератак. Подчеркнута необходимость использования адаптированных метрик (ROC-AUC, Precision-Recall кривые) и методов визуализации (матрица ошибок) в условиях экстремальной несбалансированности данных и постоянно меняющихся угроз. Рассмотрены методы оптимизации моделей для достижения баланса между точностью и полнотой, включая адаптивную настройку порогов классификации и оптимизацию гиперпараметров для F-меры.

Таким образом, проведенное в рамках магистерской работы исследование вносит значимый вклад в развитие методов машинного обучения для задач кибербезопасности. Предложенные и разработанные алгоритмы, а также проанализированные подходы к их адаптации, интерпретации и оценке, могут быть использованы для создания более эффективных и устойчивых систем защиты, способных противостоять современным и будущим киберугрозам.

Список использованной литературы

1. Ravi, K., & Raman, V. (2018). Malware detection using word2vec and recurrent neural network. In 2018 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 550-555). IEEE.

2. Yoon, H. J., & Lee, J. H. (2017). Malware detection using drilling of static and dynamic features. Computers & Security, 69, 55-64.

3. Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In Mining text data (pp. 77-128). Springer, Boston, MA.

4. Alzaylaee, M. K., Yerima, S. Y., & Sezer, S. (2020). DL-Droid: Deep learning based android malware detection using real devices. Computers & Security, 89, 101663.

5. Mirzaei, O., Gebhardt, A., Dehghantanha, A., & Choo, K. K. R. (2019). Malware analysis using word2vec. In Contemporary Digital Forensic Investigations of Cloud and Mobile Applications (pp. 41-67). Elsevier.

6. Suarez-Tangil, G., & Dash, S. K. (2016, September). Malware categorization and classification using neural networks. In 2016 International Carnahan Conference on Security Technology (ICCST) (pp. 1-6). IEEE.

7. Cai, H., Zheng, V. W., & Chang, K. C. C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 30(9), 1616-1637.

8. Zhang, Y. Insider threat detection using machine learning classification models / Y. Zhang, S. Alahmadi, S. Tjoa, S. Lee // 2019 IEEE International Conference on Big Data (Big Data). - IEEE, 2019. - P. 5266-5275.

9. Simon, T. Detecting insider threats using machine learning techniques / T. Simon, S. Malik // 2017 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). - IEEE, 2017. - P. 1-6.

10. Hochreiter, S., & Schmidhuber, J. Long Short-Term Memory [текст] / Hochreiter, S., & Schmidhuber, J. // Neural Computation. 1997. Vol. 9. No. 8. P. 1735–1780.

11. Bahdanau, D., Cho, K., & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate [текст] / Bahdanau, D., Cho, K., & Bengio, Y. // arXiv preprint arXiv:1409.0473. 2014.

12. Kotenko, I., Saenko, I., Branitskiy, A. Hierarchical Models for APT Attack Forecasting [текст] / Kotenko, I., Saenko, I., Branitskiy, A. // Computers & Security. 2025. Vol. 98. P. 102–115.

13. European Union. Horizon 2025: Distributed Predictive Systems for Cybersecurity [текст] / European Union. // EU Research Reports. 2025. P. 77–89.

14. IBM Security. Explainable AI for Cybersecurity Analysts [текст] / IBM Security. // IBM Research Journal. 2025. Vol. 14. P. 33–47.

15. Chen, Y., Wang, L., Zhang, Q. Adversarial LSTM Training for APT Prediction [текст] / Chen, Y., Wang, L., Zhang, Q. // IEEE Transactions on Dependable Systems. 2024. Vol. 21. P. 112–130.

16. Palo Alto Networks. Context-Aware LSTM for Cloud Security [текст] / Palo Alto Networks. // Journal of Cloud Computing. 2025. Vol. 14. P. 89–104.

17. Visa Inc. Economic Weighting in Fraud Detection [текст] / Visa Inc. // Visa Security Research. 2025. Vol. 9. P. 112–128.