УДК 004.85:004.056.53

ИССЛЕДОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДЕТЕКТИРОВАНИЯ СЕТЕВЫХ АТАК

К. Н. БОРОДИН [1], А. Р. ШЕДОВА[1], М. В. ГАЛИЦКИЙ[1]

[1]*Московский технический университет связи и информатики (МТУСИ)*
(г. Москва, Россия)

E-mail: kirbor2014@yandex.ru

Аннотация. В данной статье представлен сравнительный анализ различных алгоритмов машинного обучения для задачи обнаружения сетевых атак. В работе демонстрируется эффективность использования глубоких нейронных сетей, а также впервые применяются сети Колмогорова-Арнольда для обнаружения сетевых атак. В ходе исследования применялись датасеты, характеризующие различные типы атак, что позволило нам оценить производительность моделей в разных сетевых доменах. Полученные результаты подтверждают актуальность использования современных подходов к детектированию угроз в условиях растущего объема данных.

Abstract. This paper presents a comparative analysis of different machine learning algorithms for the task of network attack detection. The paper demonstrates the effectiveness of using deep neural networks and for the first time applies Kolmogorov-Arnold networks to detect network attacks. In the course of the study, datasets characterizing different types of attacks were used, which allowed us to evaluate the performance of the models in different network domains. The obtained results confirm the relevance of using modern approaches to threat detection in conditions of growing data volume.

Введение

На сегодняшний день одним из самых ценных ресурсов является информация в различных её проявлениях. Благодаря интернету экспоненциально растёт обмен данными между пользователями [1]. Сетевые атаки — это действия, направленные на нарушение работы сетей передачи данных, кражу или порчу информации. В условиях роста обмена данными также повышается количество угроз и модификаций методов атак [2], и задача их обнаружения становится сложнее.

Создание и внедрение эффективных систем для автоматического детектирования уязвимостей, обеспечение безопасности и предотвращение утечки данных внутри инфраструктуры, постоянное обновление и адаптация защитных механизмов имеют первоочередное значение.

Целью нашего исследования является проведение сравнительного анализа различных алгоритмов машинного обучения для задачи обнаружения сетевых атак. В качестве выбранных алгоритмов мы использовали: K-ближайших соседей (KNN) [3], метод опорных векторов (SVM) [4], градиентный бустинг [5], полносвязный персептрон (MLP) [6], сети Калмогорова-Арнольда (KAN) [7], двунаправленные сети с долгосрочной краткосрочной памятью (BiLSTM) [8], трансформер [9].

Для проведения экспериментов мы использовали следующие датасеты: UNSW-NB15 [10], CIC-IDS-2018 [11] и NSL-KDD [12]. Для оценки моделей мы применяли следующие метрики: precision, recall, accuracy, F1-score.

Основной вклад нашего исследования заключается в унифицированном сравнении производительности различных моделей машинного обучения для детектирования сетевых атак в разных доменах. Кроме того, впервые в данной области применяются сети Калмогорова-Арнольда, что делает наш подход инновационным.

Описание данных и методов исследования

Мы выбрали для сравнения перфоманса моделей 3 различных датасета, чтобы иметь более полноценное представление о генерализации, робастности и валидности наших моделей в разных средах при разных видах атак. Далее проведём краткое описание датасетов.

UNSW-NB15 – популярный датасет, созданный для оценки моделей обнаружения сетевых атак. Он был разработан австралийским центром Информационной безопасности (ACCS) для устранения более старого датасета KDDCup99[13]. Трафик в базе данных синтезирован таким образом, чтобы симулировать настоящие сценарии, включая нормальные и злокачественные. Датасет включает в себя такие атаки, как фаззеры, сканеры, backdoor, DoS, exploits, универсальные атаки, shellcode, worms и другие.

NSL-KDD–улучшенная версия датасета KDDCup99, которая состоит из 41 признака для каждой записи о подключении и 1 лейбла, отражающего нормальный ли это трафик или какая-то атака. Датасет является одним из важнейших бенчмарков в сфере обнаружения сетевых угроз. Датасет включает в себя следующие типы атак: DoS, R2l, U2R, Probe.

CIC-IDS-2018 – датасет, созданный канадским институтом информационной безопасности (CIC) и широко используемый для исследования обнаружения сетевых атак. Датасеты предоставляют гигантское количество реалистичных данных в сетевом трафике для обучения моделей машинного обучения. Датасет включает в себя такие типы атак, как BruteForce, SSH, DoS, Heartbleed, Botnet, SQLinjection, XSS, infiltration и другие.

Перед подачей в каждую из моделей числовые признаки нормализовывались – выборочные среднее и среднеквадратичное отклонение приводилось к 1 и 0 соответственно. Для формирования целевой переменой вредоносные строки помечались «1», нормальные «0». Категориальные признаки кодировались с помощью one-hot -кодирования перед подачей в следующие модели: SVM, KNN, MLP, KAN, BiLSTM. Алгоритмы CatBoost и признаковый трансформер используют иные способы обработки таких признаков.

При наличии пропущенных значений, или чрезмерно больших, не входящих в диапазон числового типа, мы заменяли их на среднее значение по признаку. Для обучения мы использовали все возможные признаки, удаление каких-либо строк или столбцов не осуществлялось.

Несмотря на то что в датасетах представлены разные виды атак и даже есть соответствующая разметка, мы не обучали наши модели на задачу классификации типов атак, поскольку нашей целью является разработка автономного детектора.

Каждый из выбранных нами датасетов имеет ярко-выраженный дисбаланс классов. Для корректной оценки, а также возможности иметь сравнение с другими работами мы решили использовать метрики, позволяющие более тонко оценивать модели. Будем считать положительным классом наличие вредоносного трафика, а негативным классом его отсутствие. После обучения модели мы можем рассчитать количество истинно положительных (TP), ложно отрицательных (FN), ложно положительных (FP), истинно отрицательных (TN) результатов. Тогда метрики *accuracy, precision, recall, f1* могут быть определены следующим образом (1-4):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$f1 = 2\frac{precision * recall}{precision + recall} \tag{4}$$

Эксперименты и обсуждение

В таблице 1 показаны результаты моделей на датасете UNSWNB-15. Метрики precisionи recall достаточно высокие, что говорит о способности модели эффективно обнаруживать атаки в сетевом трафике, минимизировать ложно положительные и ложно негативные ошибки.

BiLSTM модель имеет хорошие метрики, что говорит об эффективности моделей, имеющих в своей основе рекуррентную природу. Стоит также отметить, что сети Калмогорова-Арнольда продемонстрировали лучший перфоманс, чем многослойный персептрон, что позволяет их считать интересной альтернативой для дальнейших исследований.

Таблица 1. Результаты на датасете UNSWNB-15

	Accuracy	precision	recall	f1
SVM	0.7808	0.7553	0.7944	0.7742
KNN	0.7708	0.7085	0.8376	0.7672

MLP	0.8170	0.8151	0.8095	0.8127
KAN	0.8353	0.8272	0.8481	0.8375
CatBoost	0.7944	0.7596	0.8215	0.7894
BiLSTM	0.8885	0.8561	0.9217	0.8877
Pheature transformer	**0.9235**	**0.9170**	**0.9432**	**0.9303**
RF	0.7346	0.8084	0.7364	0.7707
AlexNet	0.7389	0.7961	0.7273	0.7601
LeNet-5	0.7111	0.7887	0.7113	0.7480
CNN	0.7461	0.8101	0.7565	0.7824
BiLSTM	0.7224	0.7952	0.7243	0.7581
CNN-BiLSTM	0.7716	0.8263	0.7991	0.8125
RNN	0.883	0.876	0.965	0.918
LSTM	0.899	0.889	0.973	0.929
GRU	0.897	0.886	0.973	0.928

В таблице 2 представлены результаты для датасета NSL-KDD. На этот раз результаты не настолько однозначны. Во-первых, лидера сразу по всем метрикам нет. Во-вторых, нельзя сказать о явном преимуществе рекуррентных моделей над обычными.

Прежде всего, хотелось бы отметить, что применение сверточных нейронных сетей[14] не подходит для данной задачи. Мы полагаем, что это связано с тем, что такие сети ориентированы на обработку локальных зависимостей, тогда как для обнаружения атак требуется анализировать глобальные.

Следующий фактор, на который стоит обратить внимание, — это успешный опыт использования BiLSTM: наша модель показала наилучшие результаты. Кроме того, KAN снова продемонстрировала более высокую производительность по сравнению с MLP. Наконец, наш обучающий пайплайн с CatBoost не только достигает наивысшей accuracy, но и обеспечивает наилучший баланс между precision и recall среди всех рассмотренных моделей.

Почему в данном датасете алгоритм для работы с табличными данными оказался лучше рекуррентных моделей? Мы считаем, что такая ситуация обусловлена природой датасета – он создавался в первую очередь для обработки проблем устаревшего датасета KDD99, а не для изучения новых атак. Строго говоря, этот набор данных можно считать не актуальным для исследований, и результаты CatBoost являются одним из аргументов в эту сторону.

Таблица 2. Результаты на датасете NSL-KDD

	accuracy	precision	recall	f1
SVM	0.8168	0.7406	0.9295	0.8215
KNN	0.8589	0.7861	0.9586	0.8638
MLP	0.8522	0.8011	0.9296	0.8605
KAN	0.8397	0.7459	0.9205	0.8241
CatBoost	**0.8634**	0.7871	0.9693	**0.8688**

BiLSTM	0.8594	0.7851	0.9608	0.8641
Pheature transformer	0.8390	0.7741	0.9315	0.8456
RF	0.7471	0.8133	0.7549	0.783
AlexNet	0.7702	0.7854	0.7724	0.7788
LeNet-5	0.7991	0.8295	0.8001	0.8045
CNN	0.8175	0.8243	0.8271	0.8257
BiLSTM	0.7943	0.8114	0.7965	0.8039
CNN-BiLSTM	0.8358	**0.8582**	0.8449	0.8514
RNN-IDS	0.8328	0.7295	**0.9692**	0.8324

В таблице 3 представлены результаты для датасета CIC-IDS 2018. Хочется отметить: разброс ложноположительных результатов очень большой, об этом говорят результаты по метрике precision. В действительности, данный датасет является самым не сбалансированным среди всех рассмотренных. Так, например, KNN почти свёлся к наивному классификатору, о чём говорит низкий precision и довольно высокий accuracy.

Несмотря на несбалансированный характер данных, MLP и BiLSTM продемонстрировали отличный результат по ложноположительным срабатываниям — модели научились на данном датасете не помечать обычный трафик вредоносным. Однако, для этих моделей вредоносный трафик всё ещё может оставаться замаскированным, что является для нас проблемой. С точки зрения баланса между ложноположительными и ложнонегативными результатами лучший результат в наших экспериментах показала модель табличного трансформера с f1-метрикой 97.89%, уступая по accuracy всего 1%.

Таблица 3. Результаты на датасете CIC-IDS-2018

	accuracy	precision	recall	f1
SVM	0.769	0.1574	0.1146	0.1327
KNN	0.8881	0.0038	0.8387	0.0075
MLP	0.8878	**1**	0.8878	0.9406
KAN	0.9135	0.4877	0.7975	0.6053
CatBoost	0.8886	0.0102	0.7778	0.0201
BiLSTM	0.9205	**1**	0.9178	0.9571
Pheature transformer	0.9624	**0.9916**	0.9666	**0.9789**
LG	-	0.781	0.801	0.791
XGB	-	0.845	0.834	0.839
DT	-	0.8733	0.885	0.879
HCRNN	**0.9725**	0.9633	**0.9712**	0.976

Заключение

Прежде всего, в заключении хотелось бы обсудить ограничения. Обучение проводилось на отдельных, не связанных друг с другом датасетах, каждый из которых был собран в лабораторных условиях, что затрудняет оценку способности моделей к генерализации в реальных условиях. Ещё одна проблема, выявленная для всех рассмотренных моделей, — это компромисс между precision и recall: в большинстве случаев высокая доля ложных срабатываний приводит к чрезмерному количеству ошибочных предсказаний, что в реальных системах может стать критичным. Наконец, большинство лучших моделей являются ресурсозатратными. Это вызывает сложности, так как не всегда возможно развернуть такие модели в режиме реального времени, что приводит к новому компромиссу между качеством и скоростью отклика.

Далее хочется выделить новизну нашего исследования. Результаты на основе всех трёх датасетов показывают сравнение между различными моделями, включающими машинное обучение и глубокое обучение к задаче обнаружения аномалий в сетевом трафике. Модели традиционного ML показывают себя хуже в рамках этого домена в отличие от рекуррентных моделей, которые демонстрируют лучший перфоманс. Относительно трёх рассмотренных датасетов трансформерная архитектура продемонстрировала хорошее качество, что говорит о робастности к различным сложным сценариям обнаружения сетевых атак. Архитектура Калмогорова-Арнольда, менее изученная из-за её недавнего создания, добавляет новизны в наше исследование, поскольку никто не исследовал её применение к данной задаче до этого. Наши выводы релевантны не только в академических, но и в промышленных условиях, где сетевая безопасность занимает критически важное место, благодаря чему открывает путь к внедрению более умных моделей в системы информационной безопасности.

В заключение хотелось бы отметить, что данное исследование подчеркивает значимость применения алгоритмов глубокого обучения для обнаружения атак на сетевой трафик. Однако будущие исследования должны уделять внимание ограничениям моделей в плане генерализации, работе с несбалансированными данными и возможности практического развёртывания.

Список использованных источников

1. ИСИЭЗ: [Электронный ресурс]. Доступ: *https://issek.hse.ru/news/810217750.html* (дата обращения: 10.09.2024).
2. Positive Technologies: [Электронный ресурс]. Доступ: *https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2024-q1/* (дата обращения: 12.09.2024).
3. *Consistency Properties of Nonparametric Discrimination.* In: Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties, pp. 201-210, 1985.
4. Support-vector networks: [Электронный ресурс]. Доступ: *https://link.springer.com/article/10.1007/BF00994018* (дата обращения: 17.09.2024).
5. ArXiv CatBoost: Unbiased Boosting with Categorical Features: [Электронный ресурс] Доступ: *https://arxiv.org/abs/1706.09516* (дата обращения 18.09.24)
6. Carnegie Mellon University: [Электронный ресурс]. Доступ: *https://web.archive.org/web/20151010204407/http://deeplearning.cs.cmu.edu/pdfs/Cybenko.pdf* (дата обращения: 24.09.2024).
7. ArXiv KAN: [Электронный ресурс]. Доступ: *https://arxiv.org/abs/2404.19756* (дата обращения: 25.10.2024).
8. BiLSTM: [Электронный ресурс]. Доступ: *https://www.sciencedirect.com/science/article/abs/pii/S0893608005001206* (дата обращения: 17.09.2024).
9. Transformer: [Электронный ресурс]. Доступ: *https://arxiv.org/abs/2106.11959v2* (дата обращения: 08.10.2024).
10. UNSW-NB15: [Электронный ресурс]. Доступ: *https://ieeexplore.ieee.org/document/7348942* (дата обращения: 18.09.2024).
11. IDS-2018: [Электронный ресурс]. Доступ: *https://ieeexplore.ieee.org/abstract/document/9947235* (дата обращения: 22.09.2024).
12. NSL-KDD: [Электронный ресурс]. Доступ: *https://e-tarjome.com/storage/btn_uploaded/2019-07-13/1563006133_9702-etarjome-English.pdf* (дата обращения: 30.09.2024).
13. KDDCup99: [Электронный ресурс]. Доступ*: https://arxiv.org/pdf/1706.03762* (дата обращения: 09.10.2024).
14. IEEE Xplore: [Электронный ресурс]. Доступ: *https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8998253* (дата обращения: 10.10.2024).