ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА МЕТОДА ADVERSARIAL TRAINING НА ЗАДАЧЕ БИНАРНОЙ КЛАССИФИКАЦИИ

Хаджинова Н. В., Агель А. А., Пашковец М. В.
Кафедра информационных технологий автоматизированных систем,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {agel, matvey.pashkovez01}@gmail.com, khajynova@bsuir.by

Быстрое распространение систем, основанных на машинном обучении (ML), обострило проблему их безопасности и устойчивости к целенаправленным атакам. Это поставило под угрозу достоверность обработки информации. В частности, атаки обхода (evasion attacks) с использованием состязательных примеров (adversarial examples) представляют серьезную угрозу для систем классификации данных, включая их способность корректно интерпретировать входные данные. В настоящей работе рассматривается и оценивается метод Adversarial Training (Состязательное обучение) как ключевая стратегия обеспечения надежности и достоверности классификационных моделей. В качестве механизма генерации состязательных примеров использован итерационный метод Projected Gradient Descent (PGD). Целью работы является разработка и экспериментальная оценка метода повышения устойчивости классификаторов к атакам обхода для обеспечения защиты систем обработки данных.

Введение

Автоматизированные системы, основанные на машинном обучении, стали неотъемлемым компонентом информационных инфраструктур. Однако их широкое внедрение обострило проблему защиты от целевых атак, направленных на компрометацию целостности и достоверности данных [1]. Проблема обеспечения безопасности ML-систем решается различными путями, включая как создание робастных моделей, так и разработку интеллектуальных агентов, способных активно противодействовать атакам [2]. В настоящей работе мы фокусируемся на первом подходе. Особую угрозу представляют состязательные атаки (adversarial attacks), которые позволяют злоумышленнику манипулировать входными данными, чтобы заставить модель принимать заранее определенные неверные решения. Такие атаки, как отравление (poisoning) данных на этапе обучения и обход (evasion) на этапе эксплуатации, используют добавление к данным малозаметных возмущений (в рамках заданной ℓ_p-нормы), что подрывает саму основу доверия к автоматизированным системам [3].

Проблема обеспечения робастности ML-агентов к этим атакам является ключевой задачей в области безопасности искусственного интеллекта. Для агентов, выполняющих задачи мониторинга, классификации или принятия решений, обеспечение устойчивости к состязательным возмущениям является важным условием для сохранения целостности информационных систем и противодействия целевым кибератакам.

I. Adversarial Training как метод защиты

Adversarial Training (AT) является наиболее эффективным и широко используемым методом повышения робастности моделей к состязательным примерам. Суть метода заключается в включении состязательных примеров в обучающий набор данных. Модель обучается не только на чистых данных, минимизируя обычную функцию потерь $L(\theta, x, y)$, но и на данных с максимально возможным возмущением δ в пределах допустимой ℓ_p-нормы ε, которые максимизируют функцию потерь:

$$\min_{\theta} \mathbb{E}_{(\theta,x,y)\sim D} \left[L(\theta, x, y) + \min_{\|\delta\|_p \leq \varepsilon} L(\theta, x + \delta, y) \right],$$

где θ — параметры модели, x – входные данные, y — истинный класс, и D — распределение данных [3-5]. Этот процесс можно интерпретировать как обучение модели, которое минимизирует функцию потерь в наихудшем случае, обеспечивая таким образом своего рода регуляризацию. Компромисс между робастностью и точностью, исследованный в [5], является важным аспектом при применении данного метода.

II. Экспериментальная часть

В данной работе принципы обеспечения робастности и метод состязательного обучения демонстрируются на примере двумерного классификатора, решающего синтетическую задачу бинарной классификации (Класс 0 и Класс 1). Моделирование атаки демонстрирует целенаправленное искажение входных признаков (x_1, x_2) с минимальной величиной возмущения [4]. Эксперимент проводился с использованием языка Java для чистого воспроизведения алгоритмов.

Для исследования выбрана модель линейного классификатора (логистическая регрессия). Модель имеет два веса и смещение, использует функцию активации Сигмоида. Обучение проводится с использованием функции потерь Mean

Squared Error. Выбор MSE обусловлен необходимостью упрощения демонстрации принципов обеспечения достоверности обработки информации.

Для демонстрации использован синтетический 2D-набор данных из 100 точек. Входные признаки (x_1, x_2) генерируются случайным образом в диапазоне [0,1] и являются линейно разделимыми.

Для обучения модели применялся двухэтапный подход. На первом этапе модель обучалась в течение 1000 эпох с использованием стандартного метода градиентного спуска исключительно на чистых данных, формируя базовую модель. На втором этапе проводилось состязательное обучение в течение 1000 эпох. На каждой итерации для каждого исходного образца x с помощью метода FGSM генерировался состязательный аналог x_{adv}, используя текущие веса модели и возмущение с ℓ_∞-нормой ε=0.2, применяемое однократно. Затем модель обучалась на модифицированных образцах, минимизируя функцию потерь.

Для оценки надежности обработки информации использовались два типа входных данных:

1. Исходные данные (x): Немодифицированные входные образцы из набора данных. Точность обработки этих данных (Accclean) характеризует стандартную производительность системы в штатных условиях;
2. Модифицированные данные (xadv): Образцы, целенаправленно искаженные путем добавления малозаметного возмущения (δ). Возмущение генерировалось с помощью одношагового метода FGSM с максимальной допустимой ℓ_∞-нормой ε=0.2.

Совместное использование этих двух типов данных позволяет количественно оценить уязвимость системы обработки информации: высокая точность на исходных данных при значительном снижении на модифицированных образцах свидетельствует о недостаточной надежности обработки и уязвимости к целенаправленным искажениям.

В эксперименте сравнивались две модели, каждая из которых обучалась в течение 1000 эпох: Базовая Модель, обученная только на чистых данных, и Устойчивая Модель, обученная на состязательных примерах (x_{adv}), генерируемых на каждой итерации. Оценка проводилась на полном наборе данных с ε=0.2.

Таблица 1 – Итоговые результаты робастности

Модель	Acc_{clean}(Чистые данные)	Acc_{adv}(Состязательные данные)
Базовая Модель	95,00 %	30 %
Устойчивая Модель	62,00 %	62,00 %

Заключение

В настоящей работе разработан и экспериментально проверен метод повышения устойчивости алгоритмов классификации в автоматизированных системах к целенаправленным искажениям входных данных. Ключевым результатом является создание методики защиты на основе состязательного обучения (Adversarial Training), которая позволяет повысить робастность модели к атакам с обходом защиты. Разработанный протокол проверки целостности данных и механизм генерации состязательных примеров формируют основу для встраивания модуля обнаружения аномалий в конвейер обработки информации. В качестве дальнейшего развития данного исследования представляет интерес интеграция предложенного метода в агентные системы, способные к активной защите, как это показано в [2].

Экспериментальные результаты подтвердили эффективность предложенного метода, продемонстрировав увеличение устойчивости классификатора. Полученные метрики работы системы в нормальных и атакованных условиях позволяют оптимально настроить баланс между производительностью и надежностью автоматизированных систем обработки информации.

1. Goodfellow, I. J. Explaining and Harnessing Adversarial Examples / I. J. Goodfellow, J. Shlens, C. Szegedy // International Conference on Learning Representations (ICLR). – 2015.
2. Khajynava, N. Adaptation of adversarial machine learning for training agents to counter data attacks / N. Khajynava, Z. Mutero, A. Adam // Технические средства защиты информации. – Минск, 2025. – С. 385–387.
3. Surekha, M. A Comprehensive Analysis of Poisoning Attack and Defence Strategies in Machine Learning Techniques / M. Surekha, A. K. Sagar // 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). – 2024.
4. Madry, A. Towards deep learning models resistant to adversarial attacks / A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu // International Conference on Learning Representations (ICLR). – 2017.
5. Zhang, H. Theoretically principled trade-off between robustness and accuracy / H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, M. I. Jordan // International Conference on Machine Learning (ICML). – 2019.