УДК 004.056.5:378.3
Шифр специальности ВАК: 2.3.6

И.А. Ветров, А.И. Сацута, В.В. Подтопельный

Специфика выявления компрометации обучающих данных систем искусственного интеллекта

Актуальность. В настоящее время большое распространение получили системы искусственного интеллекта, при этом многие из них уязвимы к атакам отравления обучающих данных: атакуя, злоумышленник целенаправленно вводит вредоносные экземпляры записей в датасет, что приводит к снижению точности модели или появлению скрытых бэкдоров. ***Цель исследования.*** Разработка подхода к анализу способов негативного воздействия на обучающие наборы, который позволит не только определять наиболее оптимальные воздействия, но и количественно определять степень их серьёзности (предполагается разработка системы метрик для количественной оценки атак отравления и построение модели оптимальных действий злоумышленника на основе марковских процессов принятия решений). ***Методы.*** Применены: двусторонний тест Колмогорова–Смирнова и тест хи-квадрат для сравнения распределений признаков, расстояние Дженсена–Шеннона для оценки дрейфа данных, метод Adversarial Validation, а также алгоритм итерации по значениям для нахождения оптимальной политики в марковском процессе принятия решений. ***Научная новизна.*** Предложена формальная модель взаимодействия злоумышленника и аналитика как марковского процесса принятия решений с функцией вознаграждения, одновременно учитывающей снижение точности модели, неразличимость отравленных данных и степень их статистического дрейфа относительно контрольного набора. ***Результаты.*** Получены зависимости вероятности успешной атаки от её масштаба для различных состояний доверия аналитика. Выявлены оптимальные и субоптимальные последовательности атакующих воздействий. Показано, что предложенные метрики количественно характеризуют как эффективность, так и скрытность атаки. ***Практическая значимость.*** Предложенный подход позволяет на этапе разработки систем искусственного интеллекта оценить восприимчивость используемых наборов данных к атакам отравления и обосновать превентивные меры защиты: регулярный контроль JS-дивергенции, проверку распределения меток классов и применение нескольких статистических тестов для обнаружения аномалий.

Ключевые слова: сетевая атака, нейросеть, датасет, матрица признаков, функция активации, язык программирования Python, марковский процесс принятия решений.
DOI: 10.21293/1818-0442-2026-29-1-124-134

Введение

Атаки на модели машинного обучения стали актуальной проблемой в свете их широкой интеграции в различные сферы, такие как финансовые услуги, здравоохранение, безопасность и автономные системы. Основные типы атак включают в свой состав имитационные атаки, атаки с целью обмана, утечку конфиденциальной информации через обратное проектирование моделей, что может нарушить устойчивость самих моделей. Под устойчивостью моделей к атакам понимают меру их чувствительности к изменениям во входных и в обучающих данных.

Одной из ключевых проблем является уязвимость моделей к незначительным изменениям во входных данных, что может существенно повлиять на их функционирование и привести к серьезным последствиям, особенно в критически значимых системах. Атаки отравления моделей машинного обучения, также известные как атаки на обучающие данные, представляют собой угрозы, при реализации которых злоумышленник целенаправленно вводит некорректные или вредоносные данные в обучающий набор, что может вывести модель из строя или заставить принимать ошибочные решения [1]. Атаки на модели машинного обучения, известные как «adversarial attacks», оказывают значительное влияние на их производительность и надежность. Эти атаки представляют собой серьезную угрозу, поскольку предполагают использование замаскированных данных при обучении, что затрудняет их обнаружение. В контексте специфики функционирования моделей машинного обучения такие атаки сложно выявлять и предотвращать. Необходимым условием для осуществления данных атак является доступ к хранилищам подготавливаемых данных либо возможность влиять на инструменты их сбора.

Задачей злоумышленника в таком случае является изменение или формирование обучающего набора данных X на этапе сбора информации таким образом, чтобы итоговая модель была обучена на отравленном наборе X_p, который будет гарантировать достижение злоумышленником своих целей. В этом случае специфика модели нарушителя подразумевает следующее:

1. Цель злоумышленника: компрометация целостности и конфиденциальности датасета для снижения качества модели (точность, надёжность), создание условий, при которых модель будет выполнять задачи злоумышленника (например, целевое искажение предсказаний для конкретных входов или классов, создание «слепых зон» для определенных паттернов, активация бэкдора по специфическому триггеру в данных).

2. Мотивация злоумышленников разнообразна. Хотя прямая финансовая выгода таких атак может быть менее очевидной, чем при других видах кибер-

угроз, финансовый аспект исключать нельзя. Мотивация злоумышленников включает как прямую выгоду (например, манипуляция финансовыми моделями), так и косвенную (компрометация конкурента в рамках корпоративного шпионажа или конкурентной борьбы). Другие распространенные мотивации следующие: любопытство и интеллектуальные вызовы, политические/идеологические причины, поиск и эксплуатация уязвимостей. При выборе цели злоумышленник фокусируется на датасетах для критически важных моделей, где последствия успешной атаки максимальны (финансовый ущерб, репутационные потери, безопасность), или на моделях, переобучение которых исключительно ресурсоемко. Доступность конвейера данных для внедрения элементов компрометации является важнейшим фактором при выборе цели.

3. Уровень осведомлённости об атакуемом объекте и доступ к нему должен быть высоким. Это позволяет злоумышленникам эффективно планировать и осуществлять свои действия.

4. Требования, предъявляемые к знаниям и ресурсам злоумышленника при осуществлении атаки, следующие: злоумышленник должен обладать знаниями об архитектуре и поведении целевой модели (например, понимать используемые алгоритмы, знать или уметь оценить порог уверенности предсказаний), и, как правило, иметь доступ к репрезентативным данным (тестовому набору) для оценки эффективности своей атаки. Он также должен контролировать параметры атаки, такие как доля отравляющих данных.

5. Специфика подхода к реализации следующая: злоумышленник должен обладать возможностью осуществлять изменения меток параметров, генерировать экземпляры, смещающие границы решений, искажать непрерывные признаки (для внесения шума). Следует учитывать, что в этих условиях атака может проводиться итеративно (постепенное внедрение ложных данных) и адаптивно (корректировка стратегии на основе наблюдаемых промежуточных результатов обучения, если доступны), чтобы повысить скрытность и эффективность.

Таким образом, последовательность вредоносных действий предполагает поэтапное приближение к требуемому злоумышленнику результату. При этом злоумышленнику необходимо верифицировать результаты атакующих воздействий. Для подтверждения эффективности атаки до эксплуатации целевой модели злоумышленник может использовать методы обучения shadow model или методы сложного анализа статистических свойств отравленного датасета, что требует дополнительных вычислительных ресурсов и глубокого понимания процесса обучения. Таким образом, последовательность атакующих действий должна злоумышленником предусматриваться заранее и включать этапы компрометации и верификации результатов.

Можно предположить, что злоумышленник будет стремиться экстремально быстро приблизиться к требуемым результатам. Однако следует учитывать, что важнейшим фактором успешности атаки является возможность злоумышленника обеспечить скрытность компрометации, которая напрямую связана с моделью взаимодействия с компрометируемым объектом. Отравленные данные должны быть статистически неотличимы от легитимных при стандартном анализе (EDA).

Также злоумышленник может использовать тактику clean-label poisoning (корректные метки, но вредоносные признаки). Злоумышленник стремится минимизировать процент отравленных данных для снижения риска обнаружения. Эффект атаки должен проявляться только при работе обученной модели на специфических входах (триггерах) или в виде скрытого снижения обобщающей способности.

6. Уровень квалификации злоумышленника должен быть высоким (процедура компрометации датасетов требует знания принципов реализации обучения моделей искусственного интеллекта). Требуется качественная математическая подготовка, которая необходима для проведения математического анализа промежуточных результатов и оценки эффективности атак.

7. Тип нарушителя: внешний и/или внутренний. Внутренний нарушитель должен иметь легитимный доступ к конвейеру данных (сбор, очистка, разметка, хранение). Внешний нарушитель должен обладать возможностью эксплуатации уязвимостей в системах сбора данных (веб-скрапинг с подменой, скомпрометированные API, IoT-устройства), конвейерах обработки данных (уязвимости в инструментах ETL/ELT, хранилищах данных), процессах разметки данных (подразумевается внедрение злонамеренных разметчиков или компрометация платформы), иметь возможность компрометации сторонних поставщиков данных или библиотек предобработки.

Примерами атак отравления могут быть атаки смены меток (целевые или нецелевые), добавление шума к чистым данным, атаки с чистой меткой. В задачах классификации или многоклассовой классификации целевые атаки отравления направлены на понижение вероятности верного определения целевых классов. Нецелевые атаки направлены на понижение количественных метрик моделей, например, таких, как точность или полнота. За последнее время проведено множество исследований, посвященных анализу безопасности как датасетов, так и моделей машинного обучения [1–7]. Эти работы охватывают различные аспекты, такие как выявление уязвимостей в данных и оценка устойчивости моделей к различного рода атакам, включая как атаки на наборы данных, так и атаки на сами модели.

Исследователями предлагались различные подходы по анализу эффективности тех или иных атак отравления (так, например, рассматривается и сравнивается эффективность атак отравления данных путём смены меток и отравления параметров датасетов) [2]. В данном исследовании рассматриваются вопросы восприимчивости моделей к таким видам атак, специфика влияния атак на метрики точности моде-

лей, кроме того, анализируется зависимость коэффициента успешности атаки от процентов отравления исходных данных. Аналогичные работы существуют в задачах регрессии, но уже с использованием соответствующих метрик моделей [3, 4]. Подобные исследования проводятся в области состязательных атак на модели машинного обучения, результатами которых является анализ эффективности соответствующих атак и их аналитическое обоснование [5–7].

Вопросы компрометации обучающих данных можно рассматривать в рамках моделирования атакующих воздействий на датасеты систем искусственного интеллекта (в данном случае используется модель «деревья решений»). Соответственно, при моделировании требуется изучить: специфику формирования контрольного набора данных (относительно которого проводится анализ схожести используемых при обучении данных); специфику формирования тестового набора данных, компрометируемых во время атаки.

Следует выделить метапараметры системы моделирования: название столбца меток и наименование класса большинства. В результате можно будет получить множество оптимальных и субоптимальных наборов атакующих воздействий для различных состояний модели системы искусственного интеллекта, которые характеризуются степенью доверия. Это позволит определить эффективность возможных действий злоумышленника и дать оценку качества тестового набора данных.

В этом случае актуальны следующие способы компрометации датасетов: атака смены меток, целевая атака смены меток, атака внесения шума. Все подобные атаки исследуются с учетом наличия чистого тестового набора и доли данных, подвергаемых изменению (отравлению).

Датасеты, рассматриваемые в предлагаемом моделировании, создавались по методологии SMOTE, которая использует алгоритм ближайших соседей для генерации синтетических примеров каждого класса меньшинства в заданном объёме и затем присваивает сгенерированным наблюдениям метку класса большинства. В итоге контрольный набор формируется из исходных (чистых) данных, а тестовый набор – из модифицированных (отравленных) экземпляров.

Метрики атак отравления, описывающие возможность и последствия их реализации

При моделировании порядка компрометации датасетов оценка схожести и различимости контрольного и отравленного набора данных является важнейшей задачей, сутью которой является определение того, какое влияние оказала атака на структуру набора данных, на распределение его признаков и классов. Для получения исходных данных процесса моделирования атаки решается задача оценки схожести контрольного и отравленного наборов данных. Эта задача аналогична задаче анализа дрейфа данных. Таким образом, имеет место сравнение полученных наборов данных посредством проведения статистических тестов и измерение метрик расстояния между моделями.

Поскольку в исходном наборе данных присутствуют как непрерывные, так и категориальные признаки, используются разные статистические тесты.

Для всех тестов за нулевую гипотезу H_0 принимается равенство или идентичность эталонного и эмпирического распределений вероятностей или частот. Другими словами, проверяется, действительно ли данные принадлежат одной генеральной совокупности. За конкурирующую гипотезу H_1 принимается различие распределений анализируемого признака в контрольном и полученном наборах данных (т.е. $H_1 = H_0$).

Результатом проведения тестов с заданным уровнем значимости α (α ∈ {0,1; 0,05; 0,01}) является получение *p*-значения. Уровень значимости α представляет собой вероятность ошибки первого рода (табл. 1, 2), т.е. это вероятность отбрасывания гипотезы H_0, когда она на самом деле справедлива. На основе *p*-значения делается статистический вывод о том, обнаружены ли статистически достоверные различия между распределениями.

Для анализа непрерывных признаков предлагается использовать двусторонний непараметрический тест Колмогорова–Смирнова, важной особенностью которого является чувствительность к различиям в форме распределений и их сдвигу [8]. Статистикой D (1) данного теста является максимальная абсолютная разница между двумя кумулятивными функциями распределения: $F_{1,n}(x)$ для первой выборки объёмом n и $F_{2,m}(x)$ для второй выборки объёмом m. Таким образом, данный критерий количественно определяет расстояние между эмпирическими функциями распределения (рис. 1).

$$D_{n,m} = \sup_{x} \left| F_{1,n}(x) - F_{2,m}(x) \right|. \qquad (1)$$

Таблица 1

Практический смысл уровней значимости

Уровень значимости	Решение	Возможный статистический вывод
$p > 0{,}1$	H_0 не может быть отклонена	Статистически значимые различия не обнаружены
$p \le 0{,}1$	Есть основания сомневаться в истинности H_0	Различия обнаружены на уровне статистической тенденции
$p \le 0{,}05$	Значимые сомнения в истинности H_0	Обнаружены статистически значимые различия
$p \le 0{,}01$	Отклонение H_0	Различия обнаружены на высоком уровне статистической значимости

Таблица 2

Ошибки первого и второго рода

	Решение	
	Не отвергать H_0	Отвергнуть H_0
Справедлива H_0	Правильное с вероятностью 1 – α	Ошибочное с вероятностью α
Справедлива H_1	Ошибочное с вероятностью β	Правильное с вероятностью 1 – β

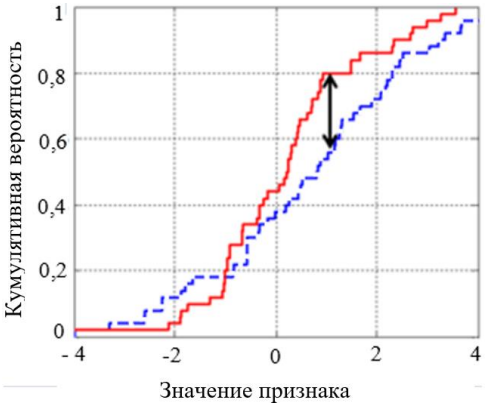

Рис. 1. Статистика теста Колмогорова–Смирнова

Для анализа категориальных признаков используется тест хи-квадрат. Статистика критерия согласия хи-квадрат вычисляется по формуле (2), в которой n_i – число выборочных значений, $P_i(\theta)$ – это вероятность попадания значения в i-й интервал (параметр θ оценивается по контрольному набору данных), а n – это общее число наблюдений. Данный тест имеет ограничение на количество наблюдений в выборке. Кроме того, он предполагает, что количество наблюдений в каждом классе не меньше 5.

$$\chi^2_n = n\sum_{i=1}^{k}\frac{\left(\frac{n_i}{n} - P_i(\theta)\right)^2}{P_i(\theta)}\ . \qquad (2)$$

Рассмотренные статистические тесты отвечают на вопрос, принадлежит ли признак из контрольного набора данных и этот же признак из отравленного набора данных одной генеральной совокупности, но не отвечают на вопрос о том, насколько они схожи или различны.

Для определения степени похожести используется расстояние Дженсена–Шеннона (Jensen–Shannon distance, далее JSD- или JS-дивергенция) (позволяет измерить информационное расстояние между двумя вероятностными распределениями). Данная мера основывается на дивергенции Дженсена–Шеннона и является её квадратным корнем. В отличие от дивергенции Кульбака–Лейблера, расстояние Дженсена–Шеннона является симметричным и ограниченным на отрезке [0, 1]. Оно рассчитывается по формуле (3), где $P = \{p_1, p_2, \ldots, p_n\}$, $Q = \{q_1, q_2, \ldots, q_n\}$ – это входные распределения, KLD – дивергенция Кульбака–Лейблера (4, 5), а M – среднее распределение (рассчитывается как $M = 1/2\,(P + Q) = \{m_1, m_2, \ldots, m_n\}$). Таким образом, для определения степени похожести наборов данных используется кумулятивная JS-дивергенция, которая определяет степень различия двух наборов данных в целом.

$$\mathrm{JSD}(P||Q) = \sqrt{\frac{1}{2}\mathrm{KLD}(P||M) + \frac{1}{2}\mathrm{KLD}(Q||M)}, \qquad (3)$$

$$\mathrm{KLD}(P||M) = \sum_{i=1}^{n} p_i \log\frac{p_i}{m_i}\ , \qquad (4)$$

$$\mathrm{KLD}(Q||M) = \sum_{i=1}^{n} q_i \log\frac{q_i}{m_i}\ . \qquad (5)$$

Кроме того, с целью обеспечения более комплексного подхода используется методика обнаружения дрейфа Adversarial validation (далее – AV) или метод подмены задачи [9]. Целью данного подхода является обучение бинарного классификатора на объединенных наборах данных, где данные контрольного набора помечаются с помощью метки 0, а остальные – с помощью метки 1. Приводимая методика основывается на предположении о том, что если экземпляры отравленного датасета статистически не отличаются от экземпляров контрольного набора, то модель будет с трудом различать данные различных групп, следовательно, классификационная метрика (в данном случае используется метрика $\mathrm{AUC_{ROC}}$ (Area Under the Receiver Operating Characteristic Curve – площадь под ROC-кривой), которая позволяет измерять качество бинарного классификатора) должна быть $\mathrm{AUC_{ROC}} \approx 0{,}5$. В противном случае, если обученная модель способна успешно разделять группы данных, то между данными наблюдаются существенные статистические различия, а значит, метрика $\mathrm{AUC_{ROC}} \gg 0{,}5$ [9].

Далее при моделировании к отравленным моделям применяется конвейер машинного обучения, который задаётся пользователем, иначе реализация данного подхода сводится к решениям модели AutoML [10]. Для обучения многоклассового различения применяется ансамблевый алгоритм градиентного бустинга (в программной реализации используется sklearn), который имеет важные преимущества (высокая точность прогнозирования, устойчивость к переобучению) [11]. Таким образом, при обучении моделей на отравленных и чистых данных допустимо применять алгоритм градиентного бустинга с перекрёстной проверкой и ранней остановкой, размер вариационной выборки составляет 25%.

Следует учесть специфику оценки качества обученной модели (используется метрика $\mathrm{AUC_{ROC}}$ (6)). Оценка качества подразумевает вычисление площади под ROC-кривой, которая строится как параметрическая зависимость точек ($\mathrm{FPR_t}$, $\mathrm{TPR_t}$), полученных при варьировании порога принятия решения $t \in [0,1]$ с использованием формул (7) и (8).

$$\mathrm{AUC_{ROC}} = \int_0^1 \mathrm{TPR}(t)\ \mathrm{dFPR}(t)\ , \qquad (6)$$

$$\mathrm{TPR}_t = \frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FN}_t}, \qquad (7)$$

$$\mathrm{FPR}_t = \frac{\mathrm{FP}_t}{\mathrm{FP}_t + \mathrm{TN}_t}. \qquad (8)$$

В данном случае TP_t, FP_t, TN_t, FN_t – это элементы матрицы ошибок при фиксированном пороге t, а ROC-кривая задаёт отображение порога t в плоскость ($\mathrm{FPR_t}$, $\mathrm{TPR_t}$), что позволяет рассматривать $\mathrm{AUC_{ROC}}$ как интеграл Римана–Стилтьеса по функции $\mathrm{FPR}(t)$.

Для конечной выборки $\mathrm{FPR}(t)$ является ступенчатой функцией, поэтому интеграл Стилтьеса сводится к конечной сумме по точкам скачков, что соответствует стандартной дискретной процедуре вычисления AUC через упорядочивание предсказа-

ний. В случае многоклассовой классификации используется стратегия «один против одного». Для каждой пары классов вычисляется значение AUC_{ROC}, а затем результаты усредняются по всем парам (9):

$$AUC_{ROC_{ovo}} = \\ = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k>j}^{c} \left(AUC_{ROC}(j|k) + AUC_{ROC}(k|j) \right), \quad (9)$$

где c – количество классов, а $AUC(j|k)$ – метрика класса j относительно k, где класс j является положительным, а k – отрицательным.

При реализации и обучении моделей типа AV также используется ансамблевый алгоритм градиентного бустинга с двумя классами.

Полученные решения используются как параметры функции вознаграждений модели марковских процессов принятия решений (МППР). То есть они определяются с помощью меры успешности атаки, которая оценивается через снижение точности модели, а также с помощью меры статистического дрейфа, вычисляемой при сравнении актуального набора данных с контрольным.

Поиск эффективных последовательностей действий злоумышленника при компрометации обучающих данных

Марковские процессы принятия решений являются мощным инструментом для моделирования стохастических систем, где выбор действий агентом влияет на переходы его между состояниями [12]. Состояния МППР в данном контексте соответствуют различным фазам взаимодействия между злоумышленником и аналитиком, который проверяет новые поступающие данные, тогда как действия агента отражают различные стратегии модификации данных.

Формально полностью наблюдаемый Марковский процесс принятия решений представляет собой множество (S, A, P, R, γ), где

– $S = \{s_1, s_2, \ldots, s_n\}$ – множество состояний;

– $A = \{a_1, a_2, \ldots, a_k\}$ – множество действий;

– P: $S \times A \to S$ – функция перехода между состояниями;

– R: $S \times A \times S \to \mathbf{R}$ – функция вознаграждения, дающая агенту мгновенное вознаграждение за применённое им действие;

– $\gamma \in [0,1]$ – коэффициент дисконтирования, суть которого заключается в определении того, насколько награда в настоящем важнее награды в будущем (при равенстве 1 модель будет пытаться максимизировать ожидаемую сумму будущих вознаграждений).

Важным понятием в контексте Марковских процессов принятия решений является политика $\pi \in \Pi$ (14), которая описывает стратегию выбора действия $a \in A$ для состояния $s \in S$, где Π – это множество всех возможных политик.

Целью моделирования при использовании МППР является максимизация суммарной ожидаемой награды (10) на шагах t, которая достигается путём нахождения оптимальной политики π^* (11) из множества всевозможных политик, где $V^{\pi}(s)$ – это ожидаемая полезность стратегии π из начального состояния s_{start} (12).

$$V^{\pi}(s) = \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t), \quad (10)$$

$$\pi^*(s_{start}) = \operatorname{argmax}_{\pi} V^{\pi}(s_{start}), \quad (11)$$

$$V_{k=1}^{\pi}(s) = R(s, \pi(s), s') + \gamma \sum_{s'} P(s' | s, \pi(s)) V^{\pi}(s') \quad (12)$$

Таким образом, модель МППР формулируется относительно входного тестового набора данных. Агент марковского процесса принятия решений может находиться в одном из возможных состояний, которые формируют множество состояний $S = \{S_у, S_б, S_н, S_с, S_в\}$ и представляют собой требования, выдвигаемые аналитиком при проверке набора данных. Данные состояния имеют следующий смысл:

– $S_в$ – состояние, описывающее высокие ожидания по похожести набора данных на контрольный набор; для проведения статистических тестов в данном состоянии используется уровень доверия $\alpha = 0{,}10$, характеризующее высокие требования к сходству, по совместительству данное состояние является начальным;

– $S_с$ – состояние, описывающее средние ожидания по похожести набора данных на контрольный набор; для проведения статистических тестов используется уровень доверия $\alpha = 0{,}05$;

– $S_н$ – состояние, описывающее низкие ожидания по похожести набора данных на контрольный набор, для проведения статистических тестов используется уровень доверия $\alpha = 0{,}01$;

– $S_б$ – состояние блокировки, описывающее невозможность набора данных успешно пройти проверку на сходство;

– $S_у$ – состояние успеха, описывающее успешное прохождение проверки на сходство для одного из состояний доверия.

Кроме того, состояния $S_б$ и $S_у$ являются терминальными состояниями, это означает, что при достижении агентом этих состояний моделирование завершается.

Множество действий задаётся как пересечение множеств, соответствующих разновидностям атак, $\boldsymbol{A} = \boldsymbol{A}_{LF} \cup \boldsymbol{A}_{TLF} \cup \boldsymbol{A}_{Noise} \cup \boldsymbol{A}_{AdvEx}$, где

– $\boldsymbol{A}_{LF}$ – множество действий, соответствующих атаке смены меток (Label Flipping (LF)) с заданной степенью воздействия;

– $\boldsymbol{A}_{TLF}$ – множество действий, соответствующих атаке целевой смены меток (Targeted Label Flipping (TLF)) с заданной степенью воздействия;

– $\boldsymbol{A}_{Noise}$ – множество действий, соответствующих атаке внесения шума (Noise Injection (Noise)) с заданной степенью воздействия;

– $\boldsymbol{A}_{AdvEx}$ – множество действий, соответствующих атаке добавления синтетических наблюдений (Adversarial Examples (AdvEx)) с заданной степенью воздействия.

В соответствии с заданными множествами состояний и действий, модель возможно визуализировать, используя граф (рис. 2).

Функция переходов в МППР представляет собой отображение, которое с заданной вероятностью переводит агента из состояния s в состояние s' посредством действия a.

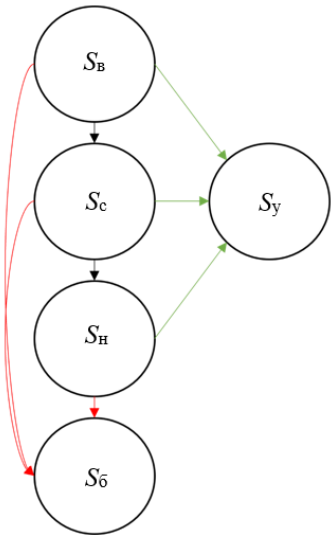

Рис. 2. Граф Марковского процесса принятия решений

Вероятность смены состояния вычисляется с использованием формулы полной вероятности (13):

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i). \qquad (13)$$

Здесь n – это количество столбцов в наборах данных, $P(A)$ – вероятность события A, заключающегося в идентичности двух наборов данных, $P(B_i)$ – вероятность события B_i, означающего, что аналитик выбирает i-е столбцы наборов, а условная вероятность $P(A|B_i)$ означает вероятность того, что наборы данных будут идентичны, основываясь на выбранном i-м признаке.

События B_i образуют полную группу событий, означающую, что $\Sigma P(B_i) = 1$, данные вероятности вычисляются посредством нормализации весов важности признаков и столбца меток. Так как столбец меток задаёт задачу классификации, которая без него нецелесообразна, он имеет вес, равный сумме важностей всех признаков.

С практической точки зрения это означает, что аналитик с наибольшей вероятностью обратит своё внимание на изменения в распределении и балансе классов. Массив важностей столбцов представляет собой следующий список, где w_i – это вес i-го признака:

$$\left[w_1, w_2, \ldots, w_{n-1}, \sum_{i=1}^{n-1} w_i\right]. \qquad (14)$$

Таким образом, $P(B_i)=w_i / \sum_{i=1}^{n} w_i$ для i=1,2,…,n, а вероятность (15) определяется путём проведения для столбца i соответствующего статистического теста, где α – уровень доверия, задаваемый текущим состоянием агента, H_0 – это гипотеза об идентичности распределений столбцов.

$$P(A|B_i)=\begin{cases}1-\alpha, \text{ гипотеза } H_0 \text{ не отвергнута} \\ (\text{различия не значимы}); \\ 0, \text{ гипотеза } H_0 \text{ отвергнута} \\ (\text{различия не значимы}).\end{cases} \qquad (15)$$

Зависимость вероятностей проведения атаки от их масштаба с учетом состояния доверия (низком, среднем, высоком) при условии использовании сформированных (синтетических) датасетов позволяет определить специфику выбора действий злоумышленником (или отсутствие выбора (clean)) в процессе атак (рис. 3).

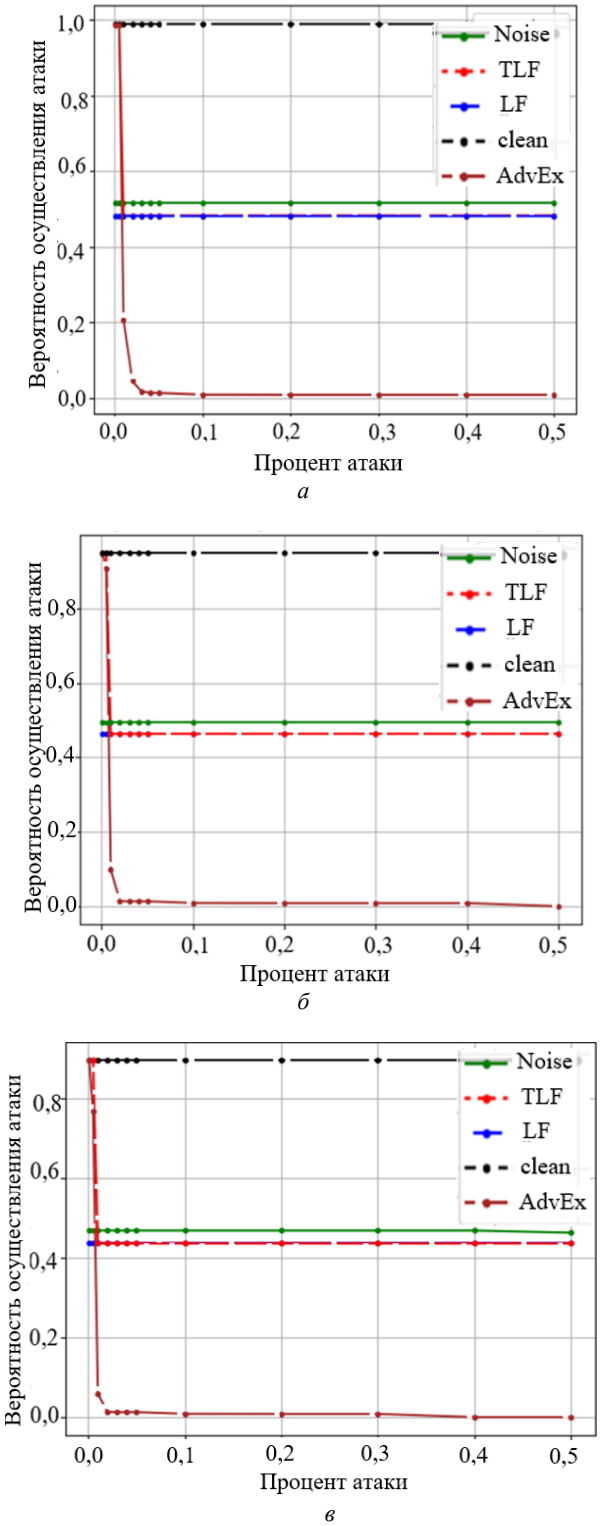

Рис. 3. Зависимость вероятностей проведения атаки от их масштаба для состояния: низкого доверия – *а*; среднего доверия – *б*; высокого доверия – *в*

Функция вознаграждений определяется с помощью меры успешности атаки (оценивается через снижение точности модели) и с помощью меры дрейфа сравниваемого набора данных относительно контрольного. Таким образом, функция наград представляет собой (16), где

– $s \in \{S_в, S_с, S_н\}$ – текущее состояние;

– $\beta \in [0, 1]$ – это коэффициент, обозначающий, насколько сильно следует учитывать сдвиг в данных при создании награды; $\beta = 1$ означает, что результирующий дрейф данных полностью учитывается, $\beta = 0$ означает, что при определении оптимальной атаки значение имеет только итоговая метрика;

– $\text{CONTROL}_{\text{DS}}$ – контрольный набор данных;

– DS_a – отравленный набор данных посредством атаки a;

– $\text{Model}_{\text{DS}_a}$ – модель, обученная на отравленном наборе данных DS_a;

– $\text{AV}_{\text{Model}_{\text{DS}_a}}$ – adversarial validation модель, обученная отличать отравленный набор данных от контрольного;

– $\text{AUC}_{\text{ROCmacro}}\left(\text{Model}_{\text{DS}_a}\right)$ – макроAUC_{ROC} – метрика для модели $\text{Model}_{\text{DS}_a}$;

– $D\left(\text{Model}_{\text{DS}_a}\right)$ – метрика неразличимости, вычисляемая на основе $\text{AV}_{\text{Model}_{\text{DS}_a}}$ (17) [13];

– $\text{JSD}(\text{CONTROL}_{\text{DS}}, \text{DS}_a)$ – кумулятивная JS-дивергенция между контрольным и отравленным набором данных.

$$R\left(s, a, s_y\right) = \left(1 - \text{AUC}_{\text{ROCmacro}}\left(\text{Model}_{\text{DS}_a}\right)\right) + \beta \frac{\left(1 - D\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right)\right) + \left(1 - \text{JSD}\left(\text{CONTROL}_{\text{DS}}, \text{DS}_a\right)\right)}{2}. \quad (16)$$

Значение функции вознаграждения не нормировано и может превышать 1, однако это не влияет на сравнение политик при фиксированном β. Для переходов, не ведущих в состояние S_y, значение функции вознаграждения принимается равным 0, так как только успешное прохождение проверки даёт положительный сигнал агенту. Использование метрики $\text{AUC}_{\text{ROCmacro}}$ обусловлено возможной несбалансированностью в распределении классов. Следовательно, применение метрики Accuracy может привести к неправильной интерпретации результатов (рис. 4).

Кроме того, $\text{AUC}_{\text{ROCmacro}}$ позволяет объективно оценить обученную модель для разных порогов принятия решений, что означает меньшее влияние на оценку гиперпараметров модели (рис. 5).

При этом следует учитывать, что $\text{AUC}_{\text{ROC}}\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right)$ является AUC_{ROC}-метрикой для модели $\text{AV}_{\text{Model}_{\text{DS}_a}}$, пороговым значением которой является 0,5, ввиду того, что в таком случае модель классификации распознаёт наблюдения не лучше случайного классификатора.

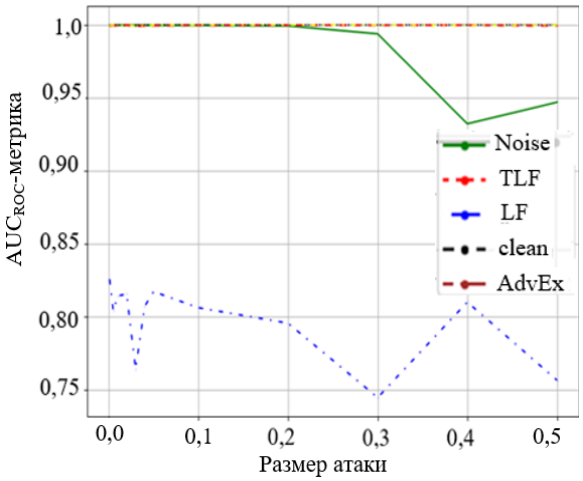

Рис. 4. Зависимость AUC_{ROC}-метрики от масштаба атаки

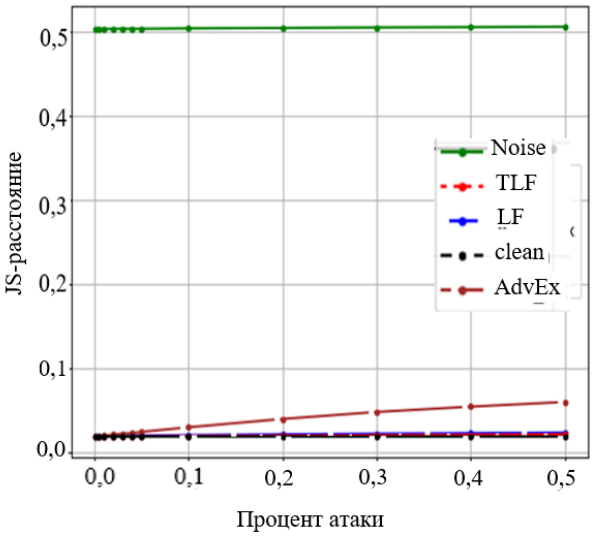

Рис. 5. Зависимость расстояния Дженсена–Шеннона от масштаба атаки

Для определения оптимальной стратегии для каждого состояния используется алгоритм итерации по значениям (Value Iteration) [13]. Данный алгоритм имеет следующие преимущества:

1. Точность результатов.
2. Скорость вычисления оптимальной стратегии.
3. Простота интерпретации работы алгоритма и его результатов.

На вход алгоритму итерации по значению передаются функции переходов и вознаграждения, параметр дисконтирования γ и порог сходимости Δ_0. Целью алгоритма является нахождение такой политики, которая максимизирует функцию ценности состояния (17), путём итеративного обновления функции ценности для каждого состояния (с помощью условия оптимальности Беллмана (18)) [13].

$$D\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right) = \begin{cases} \left(\text{AUC}_{\text{ROC}}\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right) - 0{,}5\right) \times 2, \\ \text{AUC}_{\text{ROC}}\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right) > 0{,}5, \\ 0, \ \text{AUC}_{\text{ROC}}\left(\text{AV}_{\text{Model}_{\text{DS}_a}}\right) \le 0{,}5. \end{cases} \quad (17)$$

Основным фактором при определении оптимальной политики является достижение сходимости значений $V(s)$ (разница между значениями ценности текущей и значениями ценности предыдущей итерации должна быть минимальна, т.е. ниже какого-либо заданного порога Δ_0).

$$V(s) = \max_{a \in A} \sum_{s' \in S} P(s,a,s') \left[R(s,a,s') + \gamma V(s') \right]. \quad (18)$$

Для нахождения оптимальной политики для всех состояний необходимо вычислить значения ценности при условии, что $V(S_у) = V(S_б) = 0$ (так как $S_у$ и $S_б$ являются терминальными состояниями успеха и блокировки соответственно). Требуется определить Марковский процесс принятия решений на основе вышеперечисленных установок при заданных процентах отравления [0,1; 0,2]. Пространство действий A в этом случае представляет собой множество $\boldsymbol{A} = \boldsymbol{A}_{\text{LF}} \cup \boldsymbol{A}_{\text{TLF}} \cup \boldsymbol{A}_{\text{Noise}} \cup \boldsymbol{A}_{\text{AdvEx}}$, где

– $\boldsymbol{A}_{\text{LF}} = \{\text{LF}_{0,1}, \text{LF}_{0,2}\}$ – множество действий, которые производят случайные смены меток у 10% объектов датасета ($\text{LF}_{0,1}$) и у 20% объектов ($\text{LF}_{0,2}$);

– $\boldsymbol{A}_{\text{TLF}} = \{\text{TLF}_{0,1}, \text{TLF}_{0,2}\}$ – множество действий, подразумевает целевые смены меток у 10% объектов датасета ($\text{TLF}_{0,1}$) и у 20% объектов ($\text{TLF}_{0,2}$);

– $\boldsymbol{A}_{\text{Noise}} = \{ \text{Noise}_{0,1}, \text{Noise}_{0,2}\}$ – множество действий, которые реализуют добавление случайного шума к признакам 10% объектов ($\text{Noise}_{0,1}$) и к признакам 20% объектов ($\text{Noise}_{0,2}$);

– $\boldsymbol{A}_{\text{AdvEx}} = \{ \text{AdvEx}_{0,1}, \text{AdvEx}_{0,2}\}$ – в этом случае реализуется генерация и добавление синтетических объектов в объёме 10% от исходного датасета ($\text{AE}_{0,1}$) и в объёме 20% (AE0,2).

Упрощённые формулы для функции ценности состояний $V(S_н)$, $V(S_с)$ и $V(S_в)$ в МППР соответствуют структуре исходной модели. Они точно отражают топологию переходов, учитывают терминальные состояния ($S_у$, $S_б$) с нулевой ценностью ($V(S_у) = V(S_б) = 0$), а также корректно применяют коэффициент дисконтирования только для нетерминальных состояний.

Формула полезности для состояния $S_н$ (низкие ожидания, $\alpha = 0{,}01$) допускает переходы только в терминальные состояния $S_у$ (успех) и $S_б$ (блокировка), что соответствует структуре графа переходов, она максимизирует ожидаемое вознаграждение по всем действиям (19).

$$V(S_н) = \max_{a \in A} \sum_{s' \in (S_у,\, S_б)} P(S_н, a, s') + \left[R(S_н, a, s') + \gamma V(s') \right]. \quad (19)$$

Для состояния $S_с$ (средние ожидания, $\alpha = 0{,}05$) слагаемые для терминальных состояний ($S_у$, $S_б$) равны нулю, поэтому эти состояния исключены из формального описания (20).

$$V(S_с) = \max_{a \in A} \sum_{s' \in (S_у,\, S_б)} P(S_с, a, s') * R(S_с, a, s') + P(S_с, a, S_н) * \left(R(S_с, a, S_н) + \gamma V(S_н) \right). \quad (20)$$

Формула полезности для состояния $S_в$ (высокие ожидания, $\alpha = 0{,}10$) следующая (21):

$$V(S_в) = \max_{a \in A} \sum_{s' \in (S_у,\, S_б)} P(S_в, a, s') \times R(S_в, a, s') + P(S_в, a, S_н) \times \left(R(S_в, a, S_н) + \gamma V(S_н) \right) + P(S_в, a, S_с) \times \left(R(S_в, a, S_с) + \gamma V(S_н) \right). \quad (21)$$

Формулы точно отражают топологию переходов в предложенной модели $S_н \to \{S_у, S_б\}$; $S_с \to \{S_у, S_б, S_н\}$; $S_в \to \{S_у, S_б, S_н, S_с\}$ (см. рис. 2).

Ограничения сформированной модели

Основным ограничением процесса моделирования является необходимость в достаточном объёме входных наборов данных: в случае контрольного датасета это следует из требования к его репрезентативности в той или иной степени относительно генеральной совокупности. С практической точки зрения это означает, что контрольный набор должен быть достаточного размера.

В противном случае если данное требование не выполняется, сравнительный анализ наборов данных с большей долей вероятности даст некорректный результат. Данный результат может возникнуть по причине того, что параметры распределений признаков не будут достаточно близки к параметрам соответствующих генеральных совокупностей, что окажет своё влияние на определение вероятности сходства наборов данных.

Кроме того, требование о достаточном объёме выдвигается и к тестовому набору данных. В данном случае это требование возникает из-за необходимости наличия достаточного количества наблюдений для обучения моделей. В противном случае даже чистая модель будет обладать низкими метриками точности, что окажет своё негативное влияние на процесс определения оптимальных действий агента.

Следует отметить, что представленные результаты получены на синтетических данных. Данный выбор обусловлен общим направлением исследования, которое предполагает разработку формальной модели взаимодействия злоумышленника и аналитика, а также системы метрик для количественной оценки атак отравления [14].

Использование синтетических данных с контролируемыми параметрами распределений и баланса классов позволяет получить воспроизводимые результаты и избежать влияния неконтролируемых артефактов, присущих публичным датасетам. Предложенная модель МППР и система метрик не зависят от природы данных и могут быть применены к любым табличным датасетам.

Вместе с тем валидация на реальных данных позволяет подтвердить работоспособность предложенного подхода и продолжить его развитие. Для демонстрации работоспособности (и частичного решения задач, связанных с валидацией) был проведён эксперимент на публичном наборе NSL–KDD [15].

Выбор датасета обусловлен наличием как непрерывных, так и категориальных признаков, что соответствует условиям применимости используемых статистических тестов.

Задача сведена к бинарной классификации «норма–атака». Эксперимент был ограничен атакой смены меток как наиболее прямым способом нарушения целостности обучающих данных. Выборка разделена на обучающую, контрольную и тестовую в соотношении 70/15/15, обучающая выборка подвергалась отравлению в диапазоне от 0 до 30% с шагом 5%.

Для каждого уровня отравления выполнено три запуска, результаты усреднены. При отсутствии отравления значение AUC_{ROC} составило 0,914. С увеличением доли отравленных данных до 10% метрика снизилась до 0,833 (падение 8,8%), при 20% – до 0,742 (падение 18,8%), при 30% – до 0,650 (падение 28,8%). Характер зависимости монотонный и согласуется с результатами, полученными на синтетических данных.

Дополнительно проведено сравнение с альтернативным подходом, использующим только расстояние Дженсена–Шеннона без AV: при 10% отравления JS-дивергенция составила 0,084, тогда как AV AUC вырос до 0,612, что свидетельствует о наличии скрытых статистических изменений, не фиксируемых одним лишь расстоянием между распределениями. Таким образом, совместное применение метрик повышает устойчивость обнаружения атак, а предложенная модель сохраняет работоспособность и на реальных данных. Полномасштабная валидация на расширенном спектре публичных датасетов и сопоставление с существующими методами защиты составляют предмет дальнейших исследований.

Заключение

Таким образом, был предложен подход к анализу способов негативного воздействия на обучающие наборы, позволяющий не только определять наиболее оптимальные воздействия, но и количественно определять степень их серьёзности.

Метрики атак, вычисляемые на основе статистических тестов с использованием метода Колмогорова–Смирнова (для непрерывных признаков, т.е. сравнения эмпирических функций распределения), метода хи-квадрат (для категориальных признаков), метода для различения распределений (расстояние Дженсена–Шеннона), позволяют выявить уровень компрометируемости обучающих данных для моделей машинного обучения и, соответственно, моделей ИИ в целом.

Модель действий злоумышленника на основе МППР позволяет прогнозировать последовательность действий злоумышленника. Предложенный подход позволяет на раннем этапе разработки системы искусственного интеллекта определить восприимчивость используемых наборов данных к атакам отравления, что дает возможность принять превентивные меры защиты с целью минимизации рисков поражения систем ИИ.

В качестве защитных мер можно предложить регулярную проверку JS-дивергенции и Adversarial Validation, контроль целостности и распределения меток классов, использование нескольких статистических тестов для обнаружения аномалий.

Литература

1. Намиот Д.Е. Схемы атак на модели машинного обучения // International journal of open information technologies. – 2023. – Т. 11, № 5. – С. 68–86.

2. Federated Learning Under Attack: exposing vulnerabilities through Data Poisoning Attacks in Computer Networks / E. Nowroozi, I. Haider, R. Taheri, M. Conti // IEEE Transactions on Network and Service Management. – 2025. – Vol. 22, № 1. – P. 822–831.

3. Backdoor Attacks Against Dataset Distillation / Y. Liu, Z. Li, M. Backes, Y. Shen // Proceedings of the Network and Distributed System Security Symposium (NDSS). – San Diego, California, USA, 2023. – P. 1–18.

4. Jagielski M. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning / M. Jagielski, A. Oprea, B. Biggio // Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP). – San Francisco, CA, USA: IEEE, 2018. – P. 19–35.

5. Adversarial attacks against supervised machine learning based network intrusion detection systems / E. Alshahrani, D. Alghazzawi, R. Alotaibi, O. Rabie // PLoS ONE. – 2022. – Vol. 17 (10). – Art. No. e0275971. [Электронный ресурс]. – URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275971#sec015, свободный (дата обращения: 10.09.2024)

6. Awal M.A. Investigating Adversarial Attacks in Software Analytics via Machine Learning Explainability / M.A. Awal, M. Rochan, K.R. Chanchal // Software Quality Journal. – 2025. – Vol. 33, No. 3. – P. 1–31.

7. Evasion Attacks against Banking Fraud Detection Systems / M. Carminati, L. Santini, M. Polino, S. Zanero // 23rd International Symposium on Research in Attacks, Intrusions and Defenses: USENIX Association. – San Sebastian, Spain: Curran Associates, Inc., 2020. – P. 285–300.

8. Finner H. Two-Sample Kolmogorov–Smirnov-type tests revisited: old and new tests in terms Of local levels / H. Finner, V. Gontscharuk // The Annals of Statistics. – 2018. – Vol. 6A, No. 46. – P. 3014–3037.

9. Pan J. Adversarial Validation Approach to Concept Drift Problem in User Targeting Automation Systems at Uber / J. Pan, V. Pham, M. Dorairaj // Proceedings of the AdKDD '20 Workshop. – ACM, 2020. – С. 1–6.

10. Gang L. Automating Construction of Machine Learning Models with Clinical Big Data: Proposal Rationale and Methods / L. Gang, B.L. Stone, M.D. Johnson // JMIR Research Protocols. – 2017. – Vol. 6, No. 8. – Art. No. e175. DOI: 10.2196/resprot.7757.

11. Бурков А. Машинное обучение без лишних слов / А. Бурков. – СПб.: Питер, 2020. – 192 с.

12. Кохендерфер М. Алгоритмы принятия решений / М. Кохендерфер, Т. Уилер, К. Рэй. – М.: ДМК-Пресс, 2023. – 684 с.

13. Wang Y. On the use of adversarial validation for quantifying dissimilarity in geospatial machine learning prediction / Y. Wang, M. Khodadadzadeh, R. Zurita-Milla // GIScience and Remote Sensing. – 2025. – Vol. 62, No. 1. – P. 1–25.

14. Common Attack Pattern Enumerations and Classifications [Электронный ресурс]. – URL: https://capec.mitre.org/, свободный (дата обращения: 12.05.2024).

15. NSL-KDD dataset – набор данных для оценки систем обнаружения вторжений [Электронный ресурс]. – URL: https://fkie-cad.github.io/COMIDDS/content/datasets/nsl_kdd_dataset/, свободный (дата обращения: 21.10.2025).

Ветров Игорь Анатольевич
Канд. техн. наук, доцент образовательно-научного кластера «Институт высоких технологий»
Балтийского федерального университета им. И. Канта
А. Невского ул., 14, г. Калининград, Россия, 236041
Тел.: +7-906-216-47-19
Эл. почта: vetrov.gosha2009@yandex.ru

Сацута Анатолий Игоревич
Студент Балтийского федерального ун-та им. И. Канта
А. Невского ул., 14, г. Калининград, Россия, 236041
Тел.: 8-911-865-86-81
Эл. почта: toha.satzuta@yandex.ru

Подтопельный Владислав Владимирович
Ст. преп. Института цифровых технологий (ИЦТ)
Калининградского государственного технического университета (КГТУ)
Советский пр-т, 1, г. Калининград, Россия, 236022
ORCID: 0000-0002-7618-3224
Тел.: +7-900-353-98-81
Эл. почта: ionpvv@mail.ru

Поступила в редакцию: 07.02.2026.
Принята к публикации: 13.05.2026.

Vetrov I.A., Satsuta A.I., Podtopelny V.V.

Specifics of Detecting Training Data Compromise in Artificial Intelligence Systems

***Relevance*.** Artificial intelligence systems are currently widely deployed, but many of them are vulnerable to training data poisoning attacks. When attacking, an attacker purposefully introduces malicious instances into the dataset, which leads to a decrease in model accuracy or the emergence of hidden backdoors. ***Purpose of the study*.** Development of an approach to analyzing ways of negative impact on training sets. This approach will not only determine the most optimal impacts, but also to quantify their severity (the study proposes developing a system of metrics for quantitatively assessing poisoning attacks and building a model of optimal actions of an attacker based on Markov decision processes). ***Methods*.** The following tests were applied: The two-sample Kolmogorov–Smirnov test and the chi-square for comparing feature distributions, the Jensen–Shannon distance to estimate data drift, the Adversarial Validation method, and the value iteration algorithm to find the optimal policy in the Markov decision process (MDP). ***Novelty*.** A formal model of the interaction between an attacker and an analyst is proposed as MDP with a reward function that simultaneously accounts for the decrease in model accuracy, the indistinguishability of poisoned data and the degree of their statistical drift relative to a control set. ***Results*.** The dependences of the probability of a successful attack on its scale for various states of analyst trust are obtained. Optimal and suboptimal sequences of attack actions are identified. It is shown that the proposed metrics quantitatively characterize both the effectiveness and stealth of the attack. ***Practical significance*.** The proposed approach allows for assessing the susceptibility of datasets to poisoning attacks during the development of artificial intelligence systems and substantiating preventative measures: regular monitoring of JS divergence, checking the distribution of class labels and applying several statistical tests to detect anomalies.

Keywords: network attack, neural network, dataset, feature matrix, activation function, Python programming language, Markov decision processes.
DOI: 10.21293/1818-0442-2026-29-1-124-134

References

1. Namiot D.E. [Schemes of attacks on machine learning models]. *International journal of open information technologies*, 2023, vol. 11, no. 5, pp. 68–86 (in Russ.).

2. E. Nowroozi, I. Haider, R. M. Conti. Taheri. Federated Learning Under Attack: exposing vulnerabilities through Data Poisoning Attacks in Computer Networks. *IEEE Transactions on Network and Service Management*, 2025, vol. 22, no.1, pp. 822–831

3. Y. Liu, Z. Li, M. Backes, Y. Shen. Backdoor Attacks Against Dataset Distillation. Proceedings of the Network and Distributed System Security Symposium (NDSS), San Diego, California, USA, 2023, pp. 1–18.

4. M. Jagielski, A. Oprea, B. Biggio [et al.] Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *Proceedings of the 2018 IEEE Symposium on Security and Privacy* (*SP*), San Francisco, CA, USA, IEEE, 2018. pp. 19–35.

5. Alshahrani E., Algazzawi D., Alotaibi R., Rabie O. Adversarial attacks against supervised machine learning based network intrusion detection. *PLOS.One*, 2022, vol. 17 (10), art. no. e0275971. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275971#sec015 (accessed: 10 September 2024)

6. Awal, M.A. Investigating Adversarial Attacks in Software Analytics via Machine Learning Explainability // *Software Quality Journal*, 2025, vol. 33, no.3, pp. 1–31.

7. Carminati M., Santini L., M. Polino, Zanero S. Evasion Attacks against Banking Fraud Detection Systems. 23rd International Symposium on Research in Attacks, Intrusions and Defenses: USENIX Association, San Sebastian, Spain, Curran Associates, Inc., 2020, pp. 285–300.

8. Finner, H. Two-Sample Kolmogorov–Smirnov-type tests revisited: old and new tests in terms Of local levels. *The Annals of Statistics*, 2018, vol, 6A, no. 46, pp. 3014–3037.

9. Pan J., Dorairaj M., Chen H., Lee J. Adversarial Validation Approach to Concept Drift Problem in User Targeting Automation Systems at Uber. Proceedings of the AdKDD '20 Workshop, ACM, 2020, no. 1–6.

10. Gang L., Stone B.L., Johnson M.D. Automating Construction of Machine Learning Models with Clinical Big Data: Proposal Rationale and Methods. *JMIR Research Protocols*, 2017, vol. 6 (8), art. no. e175. DOI: 10.2196/resprot.7757.

11. Burkov A. *Mashinnoye obucheniye bez lishnikh slov* [Machine learning without unnecessary words]. St. Petersburg, Peter, 2020, 192 p. (in Russ.).

12. Kohenderfer M., Wheeler T., Ray K. *Algoritmy prinyatiya resheniy* [Algorithms of decision–making]. Moscow, DMK-Press, 2023, 684 p. (in Russ.).

13. Wang Y. On the use of adversarial validation for quantifying dissimilarity in geospatial machine learning prediction. *GIScience and Remote Sensing*, 2025, vol. 62, no 1. pp. 1–25.

14. General descriptions and classifications of attack patterns Available at: https https://capec.mitre.org, free (accessed: 12 May 2024).

15. NSL-KDD – dataset [Electronic resource]. Available at: https://www.unb.ca/cic/datasets/nsl.html, free (accessed: 21 October 2025).

Igor A. Vetrov
Candidate of Engineering Sciences, Associate Professor,
Institute of Physical and Mathematical Sciences
and Information Technologies,
Immanuel Kant Baltic Federal University
14, A. Nevsky St., Kaliningrad, Russia, 236041
Phone: +7-906-216-47-19
Email: vetrov.gosha2009@yandex.ru

Anatolii I. Satcuta
Student, Immanuel Kant Baltic Federal University
14, A. Nevsky St., Kaliningrad, Russia, 236041
Phone: +7-911-865-86-81
Email: toha.satzuta@yandex.ru.

Vladislav V. Podtopelny
Senior Lecturer, Institute of Digital Technologies,
Federal State Budgetary Educational Institution of Higher
Education Kaliningrad State Technical University
1, Sovetsky Ave., Kaliningrad, Russia, 236022
ORCID: 0000-0002-7618-3224
Phone: +7-900-353-98-81
Email: ionpvv@mail.ru

Received: 07.02.2026.
Accepted: 13.05.2026.