УДК 004.056.5:378.3

В.В. Подтопельный

Особенности моделирования атак на модели машинного обучения с использованием марковских процессов принятия решений

Рассматриваются проблемы, возникающие при решении задач моделирования атак на модели искусственного интеллекта, которые интегрированы в современные информационные системы. Приведены и охарактеризованы особенности моделирования с использованием методологии MITRE, применяемой при построении вектора сетевой атаки. Проанализированы специфические особенности использования марковских процессов принятия решений при моделировании атакующих воздействий. Рассмотрена их пригодность для различных процедур определения параметров вектора атаки. При определении специфики моделирования атак на модели искусственного интеллекта рассматриваются уязвимости систем искусственного интеллекта. Изучается формирование вектора атаки в контексте эксплуатации уязвимостей.

Ключевые слова: сетевая атака, уязвимость, марковские процессы принятия решений, стратегия, оптимальная политика, метод обучения, искусственный интеллект.
DOI: 10.21293/1818-0442-2024-27-2-21-30

В современных условиях развития технологий в настоящий момент особенно интенсивно применяются системы искусственного интеллекта (ИИ) для решения различных задач, в том числе в области обнаружения компьютерных атак. При этом следует отметить, что системы и алгоритмы ИИ также уязвимы и также являются целевыми объектами при развертывании последовательности атакующих воздействий. Поэтому требуется проводить анализ защищенности систем ИИ в том числе с применением моделирования компьютерных атак, учитывая специфику функционирования систем машинного обучения.

Не все из существующих методов моделирования одинаково применимы к задачам анализа атак подобного рода, поскольку уязвимости вычислительных моделей ИИ, а также подсистем, осуществляющих сбор и обработку данных ИИ, достаточно специфичны: атаки могут реализовываться на основе эксплуатации заданной неточности работы моделей ИИ на основе манипуляции с исходными обучающими выборками и т.п.

Таким образом, при аудите информационной безопасности возникает необходимость поиска возможных последовательностей атакующих воздействий (векторов в терминологии ФСТЭК), которые включают в свой состав эксплуатацию специфических особенностей (уязвимостей) новых интеллектуальных технологий, встраиваемых в современные информационные системы (ИС) [1].

Описание проблемной области

Можно выделить следующие проблемы при использовании машинного обучения, которые могут повлиять на безопасность обработки данных ИС:

1. Необходимость обеспечения чистоты и достоверности данных. Требуется обеспечить качество данных.

2. Требуется осуществлять тщательный выбор признаков для обучения модели, чтобы обеспечить её эффективность, избегая избыточности данных. Необходимо обеспечить корректный отбор признаков.

3. Обучение и оценка модели требуют значительных вычислительных ресурсов и времени, особенно при работе с большими объемами данных.

4. Существует необходимость использования опыта специалистов для корректной настройки и улучшения моделей.

5. Анализ данных в реальном времени требует устойчивой и производительной системы, способной обрабатывать большие объемы информации и обеспечивать своевременное обнаружение угроз. Требуется обеспечить баланс скорости и качества обработки данных.

Приведенные проблемы позволяют реализовать несколько основных типовых наборов атак на системы ИИ:

1. Атаки «белого ящика» подразумевают полный доступ к модели машинного обучения, включая ее архитектуру, параметры и данные.

2. Атаки типа «черный ящик» подразумевают известность только входных и выходных данных модели. Однако с помощью различных методов, таких как внедрение шума в данные или анализ выходных данных, атакующий может искажать выводы модели или даже извлекать некоторую информацию о ее внутреннем устройстве.

3. Атаки типа «серый ящик» основываются на частичной известности злоумышленнику используемой модели ИИ.

4. Атаки, отравляющие данные. Этот тип атаки заключается во внесении изменений в обучающие данные модели.

5. Атаки, использующие уязвимости программной и аппаратной среды.

Следует отметить: различные вычислительные модели в разной степени уязвимы к указанным атакам. В целом уровень уязвимости зависит от двух факторов: известности и распространённости моделей (они могут быть типовыми и индивидуальны-

ми), распространенности и доступности обучающих данных. Кроме того, следует отметить то, что разные вычислительные модели ИИ (алгоритмы) поразному реагируют на атаки: одни демонстрируют хорошую устойчивость к воздействию на них через эксплуатацию уязвимостей, другие, наоборот, показывают чрезмерную восприимчивость к атакам (табл. 1).

В табл. 2 можно увидеть основные уязвимости, связанные с отравлением данных, используемых при обучении. Наличие этих уязвимостей подчеркивает необходимость строгих мер безопасности и контроля в процессе работы с данными, так как их несоблюдение ведет к нарушению функциональности систем ИИ.

Таблица 1

Уязвимости моделей ИИ

Атаки	Алгоритмы	Вероятность успешного воздействия	Источник
Состязательные примеры (целенаправленные возмущения входных данных, которые приводят к неправильной классификации выводимого)	Нейронные сети, глубокие нейронные сети, SVM, k-NN, логистическая регрессия, линейная регрессия, решающие деревья, ансамбли моделей	Высокая	SecurityNet, NIST, OWASP
Data Poisoning (внесение вредоносных изменений в обучающие данные, чтобы ухудшить производительность модели)	Все алгоритмы	Средняя	SecurityNet, NIST
Кража модели (восстановление модели машинного обучения)	Любые модели через API	Средняя	SecurityNet, NIST
Инверсия модели (восстановление данных обучающих примеров на основе выводов модели)	Нейронные сети, глубокие нейронные сети	Низкая	NIST, IEEE
Определение принадлежности (поиск конкретного примера, использованного для обучения модели\.)	Нейронные сети, глубокие нейронные сети	Средняя	SecurityNet, NIST
Смена меток (модификация меток обучающих данных, чтобы ухудшить производительность модели)	Все алгоритмы	Средняя	SecurityNet, NIST

Таблица 2

Используемые при отравлении данных уязвимости

Уязвимость	Описание	Действия злоумышленника	Вероятность
Незащищенные базы данных	Недостаточная защита базы данных, отсутствие шифрования, слабые пароли	Получение доступа к базе данных через уязвимость и изменение данных или меток	Высокая (30–40%)
Недостатки процедур аутентификации и авторизации	Недостаточные меры контроля доступа, недостаточная сегментация прав доступа	Получение доступа к системе через компрометированные учетные записи или недостаточно защищенные интерфейсы	Высокая (25–35%)
Незащищенные файловые системы	Недостаточная защита файловой системы, отсутствие контроля целостности файлов	Изменение или замена файлов данных после получения доступа к файловой системе	Средняя (20–30%)
Отсутствие проверки данных из внешних источников	Отсутствие валидации и проверки поступающих данных	Внедрение и отправка вредоносных данных через скомпрометированные внешние источники	Средняя (15–25%)
Отсутствие мониторинга и аудита	Недостаточный мониторинг и аудит данных и процессов	Внесение изменений в данные без обнаружения	Низкая (10–20%)
Социальная инженерия	Использование методов социальной инженерии для обмана сотрудников и получения доступа	Обман сотрудников для внесения изменений в данные или получения доступа к системе	Средняя (20–30%)
Недостаточная защита сетевого периметра	Недостаточная защита сетевого периметра, отсутствие сегментации сети	Получение удаленного доступа к системе для внесения изменений в данные	Средняя (15–25%)

При моделировании атак на системы и модели ИИ необходимо учитывать принятые методологии построения вектора атакующих воздействий. Методология MITRE ATLAS уже включает в свой состав действия, обозначающие не только перемещение между узлами, но и между состояниями, которые сопоставлены этапам компрометации вычислительной модели ИИ или системы ИИ в целом [2]. Таким образом, действия злоумышленника при атаке соответствуют тактикам методологии MITRE ATLAS (далее – MITRE), а их осуществление соответствует наступлению состояний, фиксирующих успех действия на некотором этапе атаки. Однако моделирование атак на алгоритмы ИИ (математический аппарат), заслуживает отдельного внимания, поскольку целевым объектом злоумышленника становится

специфика вычислений, некоторые допущения в области точности, трактуемые как уязвимости. Соответственно, необходимо изучить специфику моделирования атак на модели ИИ. И поскольку нейронные сети демонстрирует высокую степень уязвимости к состязательным атакам, целесообразно сосредоточить внимание на исследовании аспектов моделирования состязательных атак, таких как атака FGSM (в данном случае рассматриваемый вариант атаки относится к классу атак «белый ящик») [3]. При проведении подобных атак для формирования входных данных в модель часто используются градиенты потерь, чтобы создать новые данные, которые максимизируют потери, приводящие к неточности классифицирования.

Для определения вектора атаки может применятся моделирование на базе марковских процессов принятия решений (МППР), поскольку данный метод позволяет учесть фактор неопределенности. Это особенно важно при анализе компьютерных атак, уязвимости которых носят вероятностный характер, где атакующие могут изменять свои тактики и стратегии.

Вопрос использования марковских процессов для исследования компьютерных атак рассматривается в ряде научных работ [4–6], что доказывает их применимость для решения задач информационной безопасности.

Определение ограничений при моделировании атак

В банке данных угроз безопасности информации ФСТЭК РФ присутствуют пять возможных угроз на технологии искусственного интеллекта. К ним относятся:

1. УБИ. 218. Угроза раскрытия информации о модели машинного обучения.
2. УБИ. 219. Угроза хищения обучающих данных.
3. УБИ. 220. Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта.
4. УБИ. 221. Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных.
5. УБИ. 222. Угроза подмены модели машинного обучения.

В случае рассмотрения атак FGSM на модели ИИ важность представляют следующие угрозы: УБИ. 218, УБИ. 220, УБИ. 221. Сценарии реализации угроз безопасности информации должны быть определены для соответствующих способов реализации угроз безопасности информации. Определение сценария реализации угроз предусматривает установление последовательности возможных тактик и техник.

Описание возможных тактик и техник рассмотрено в соответствии с глобальной базой знаний о тактиках и способах, основанных на реальных наблюдениях метрической базы MITRE ATLAS (является базой знаний о произошедших инцидентах, тактиках и техниках, применяемых нарушителем для атак на системы машинного обучения и нейронные сети) [2].

В соответствии с матрицей MITRE ATLAS и Методикой оценки угроз безопасности информации ФСТЭК РФ [1, 2], можно привести обобщенный перечень тактик и техник для реализации угроз на системы машинного обучения (табл. 3).

Таблица 3

Перечень тактик атак на модели

Тактики	Техники
Сбор информации о системах и сетях	Сбор информации из публичных источников
	Сбор общедоступной информации о результатах анализа уязвимостей известных моделей
	Сбор информации в открытых репозиториях приложений
	Обход модели при помощи создания состязательных данных, которые нарушают корректную классификацию моделью
	Эксплуатация уязвимостей компонентов систем и сетей
	Доступ к API модели
	Доступ к продукту или сервису, использующему нейронные сети
	Полный доступ к модели
Внедрение и исполнение вредоносного программного обеспечения в системах и сетях	Выполнение вредоносного кода с участием пользователя
	Использование интерпретатора командной строки и скриптов
Закрепление (сохранение доступа) в системе или сети	Отравление обучающих данных
	Создание бэкдора в модели
Сокрытие действий и применяемых при этом средств от обнаружения	Обход модели при помощи создания состязательных данных, которые нарушают корректную классификацию моделью
Несанкционированный доступ и (или) воздействие на информационные ресурсы или компоненты систем и сетей, приводящие к негативным последствиям	Обход модели
	Отказ в обслуживании модели
	Засорение системы машинного обучения мусорными данными. Нарушение целостности модели МО

Следует отметить, что в произвольный момент времени атакуемая система ИИ может находиться в любом из десяти состояний, которые образуют пространство состояний [2].

Подобный подход удобен при рассмотрении посредством марковских процессов принятия решений (МППР) последовательности атакующих воздействий как череды смены состояний и, одновременно,

позволяет отследить их сопряжённость (для этого достаточно применения системы линейных уравнений). Необходимо подчеркнуть сложную природу состояний, поскольку они учитывают с одной стороны, тип тактики (который приведён в методологии ФСТЭК), с другой стороны, поскольку данные состояния достигаются при успешной эксплуатации уязвимости (узлов сети, их программного обеспечения), в пространстве состояний также должна учитываться специфика процесса достижения успеха, т.е. действия, которое позволяют эксплуатировать уязвимости, а также сами уязвимости.

Последовательность состязательной атаки показана на рис. 1. К правильно распознаваемым данным X добавляется шум δX, вычисленный на основании функции потерь модели нейронной сети.

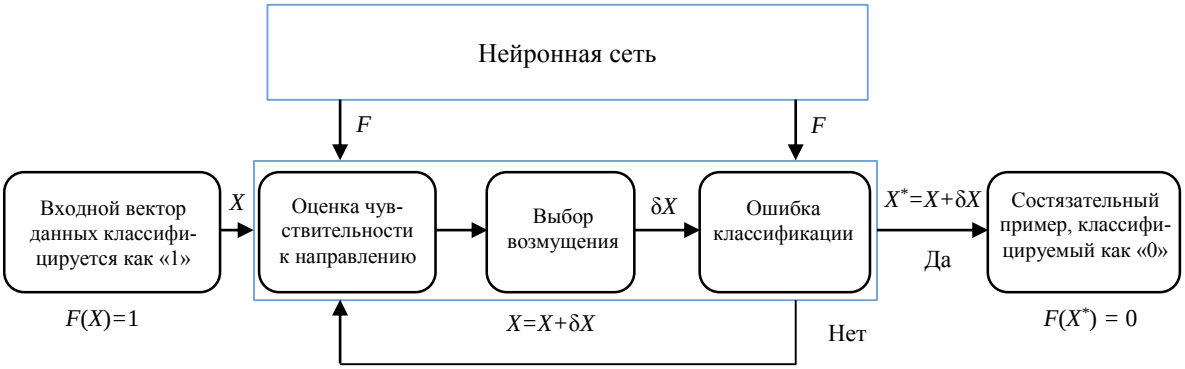


Рис. 1. Изменение данных для обмана классификатора

Модификации входных данных сильно связаны с проблемой устойчивости модели. Под устойчивостью модели понимают меру его чувствительности к возмущениям в исходных данных. Модель считается устойчивой, если при обучении погрешность в изначальных данных поэтапно не снижает точность классификации. При этом достичь неустойчивости работы модели можно другим способом: данные могут быть модифицированы на этапе тренировки, когда в обучающий набор добавляются записи, которые снижают качество классификации. Соответственно, возникает проблема доверия к обучающим датасетам. Эту проблему можно решить путем введения процедур обязательной верификации обучающих выборок. Однако в этом случае атака на уже обученную модель не исключается. Для изучения аспектов моделирования подобных атаки могут использоваться различные нейросети.

В данном случае использовалась простая нейронная сеть (семь слоев, из которых пять скрытых полносвязанных слоев) и сборка генеративных состязательных сетей (GAN), которая включает в свой состав нейросеть-генератор (шесть слоев, из которых четыре скрытых полносвязанных слоев) и нейросеть-дискриминатор (семь слоев, из которых пять скрытых полносвязанных слоев). Сеть GAN переобучена с учетом наличия состязательных атак

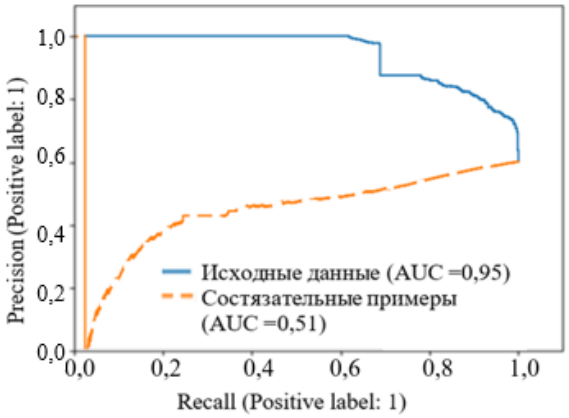
В ходе моделирования атаки была определена случайная выборка 1000 векторов данных из тестового набора. Тестирование атаки проводилось как на исходных данных, так и на сгенерированных состязательных выборках с различной величиной множителя возмущения. Результаты эффективности работы метода оценивались при помощи кривой ROC-AUC показателя и кривой Precision-Recall.

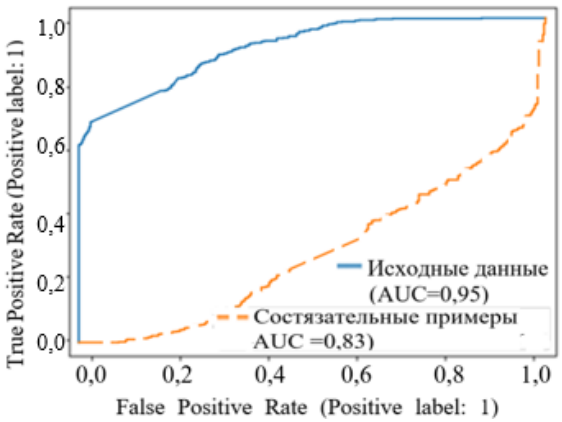
Кривые ROC показывают частоту истинно положительных результатов (TPR) по оси Y и частоту ложных положительных результатов (FPR) по оси X. Показатель полноты является площадью под кривой ROC, и для идеальных моделей он равен 1. Реальные модели должны стремиться к показателю, близкому к 1. Кривая Precision–Recall определяется на основе метрик точности (Precision) и полноты (Recall).

На основе показателей полноты и точности определяется метрика средней точности AP как средневзвешенное значение точности (Precision), достигнутой на каждой итерации с увеличением полноты (Recall), по сравнению с предыдущей итерацией. На рис. 2, *а* показаны результаты атаки на обычную модель нейронной сети. Показатели AP и AUC резко уменьшаются при проведении состязательных атак.

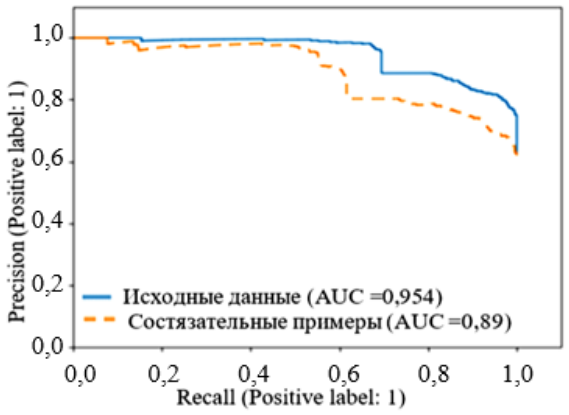
При проведении аналогичной атаки на модифицированную нейросеть (модель GAN), обученную с учетом наличия состязательных атак, показатели эффективности нейронной сети остаются на высоком уровне, и при этом эффективность атаки падает (рис. 2, *б*). Следует учитывать, что диапазон допустимых отклонений уменьшается в процессе классификации данных (при переобучении обычных нейросетей подобное может приводить к появлению множества ошибок, связанных с ложноположительными отказами вычислительной модели).

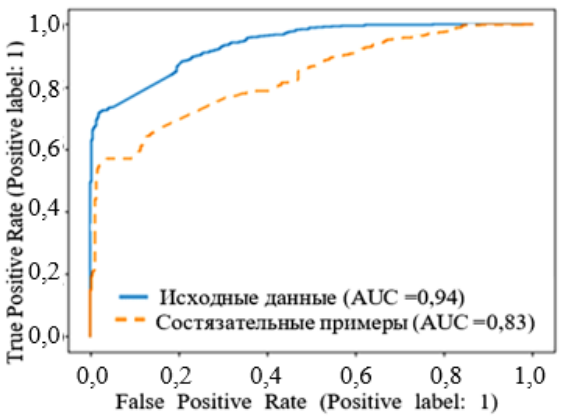
Несмотря на то, что нейронная сеть (модель GAN) показала высокую устойчивость к состязательным атакам (по модели белого ящика), а при проведении атак черного ящика увеличивается время генерации состязательных примеров, а также уменьшается эффективность состязательных выборок из-за отсутствия данных о модели нейронной сети, диапазон доступных возмущений для манипуляций злоумышленника все равно присутствует (что означает увеличение времени его поиска).





а





б

Рис. 2. Графики AP и AUC при моделировании состязательных атак: *а* – на обычную модель нейронной сети; *б* – на обученную модель GAN

Кроме того, стоит отметить, что модифицированная сеть устойчива к уровню возмущений до 20% от исходного вектора данных. Таким образом, задача построения защищенных нейронных сетей сводится к минимизации величины уязвимости и возможности ее эксплуатации при манипуляции с вычислительной моделью ИИ.

Описание модели и ее применение

При использовании марковских процессов принятия решений (МППР) сценарий атаки (вектор атаки) описывается направленным графом, включающим вершины, интерпретируемые как этапы вредоносных действий по перечню MITRE ATLAS. Предполагается, что злоумышленник получает вознаграждение *r*, которое зависит от действия *a* и состояния *s*.

В общем случае применяемый метод включает следующие ключевые этапы:

1. Определение пространства состояний системы и возможных действий по отношению к ним.

2. Назначение вероятностей переходов состояний (постоянных во времени).

3. Определение характеристик выхода (политика действий злоумышленника).

4. Разработка параметров математической модели (формирование матриц вероятностей переходов, вознаграждений) и решение задачи МППР.

5. Анализ результатов.

Марковская модель атаки описывается кортежем (S, A, P, R, γ) где [6]:

1. $S = \{s1, s2, ..., s10\}$ – множество вершин-состояний системы, которые соответствуют успешному осуществлению тактик MITRE ATLAS. Количество состояний и переходов между ними определяется на основе технического исследования целевой инфраструктуры.

2. $A = \{(s_i, s_j) \mid s_i, \quad s_j \in S, \quad a \in \{1, ..., 11\}\}$ – множество переходов, представляемых как набор действий, направленных на продвижение атаки до достижения цели злоумышленника.

3. $R = \{r_{ij} \mid (s_i, s_j) \in S\}$ – это награды за переход в определенное состояние, характеризующие эффективность или сложность действий при эксплуатации уязвимостей.

4. $P(s, a, s')$ – вероятности перехода из состояния $s \in S$ при действии $a \in A$ в состояние $s' \in S$.

5. γ – коэффициент дисконтирования (оценивает ценность будущих наград) $\gamma \in [0;1]$.

6. Политика $\pi(s, a)$ – функция (1), описывающая распределение вероятностей выбора действий злоумышленника в состоянии *s*, которое соответствует достигнутому этапу атаки. Достижение цели моделирования означает нахождение оптимальной политики π^*, следование которой позволяет максимизировать получаемую награду злоумышленника (в данном случае награда зависит от степени найденных уязвимостей). Оптимальная политика целиком описывает поведение атакующего и факти-

чески представляет собой стратегию поведения злоумышленника, т.е. его конкретные действия.

$$\pi(a \mid s) = P\left[A_i = a \mid S_i = s\right]. \quad (1)$$

Функция ценности определяется следующим образом вознаграждений на каждом *i*-м шаге (2).

$$G_i = R_{i+1} + yR_{i+2} + y^2 R_{i+3} = \sum_{k=0}^{\infty} y^k R_{i+k+1}. \quad (2)$$

Ценность состояния – это ожидаемое дисконтированное вознаграждение *R* злоумышленника, начинающего эксплуатировать уязвимость из состояния $s_i = S$ при соблюдении стратегии атаки π. Функция ценности состояния *s* при стратегии π определяется как математическое ожидание дисконтированной суммы будущих вознаграждений *R* (описано уравнением оптимальности Беллмана (3). Для поглощающего состояния графа атаки ценность равна нулю [7, 8].

$$V_\pi(s) = M[R_{i+1} + \gamma V_\pi(S_{i+1}) \mid S_i = S], \; s \in S, \quad (3)$$

где $V_\pi(s)$ – функция ценности состояний (важность достижения состояния успеха эксплуатации уязвимостей атакующим в векторе атаки) при стратегии атаки π; *M* – математическое ожидание случайной величины; γ – коэффициент дисконтирования.

Марковский процесс принятия решений позволяет определить наилучшую стратегию π_* поведения нарушителя в заданной системе. Наилучшая стратегия – это такая стратегия, при соблюдении которой достигается максимальная ожидаемая совокупная награда, которая в дальнейшем будет определятся как параметр компрометации (в том числе и степени общей уязвимости) модели ИИ [7]. При выявлении актуальной политики для злоумышленника может использоваться метод итерации, применимый в МППР [7, 8].

Начальные значения награды (*R*) переходов по вершинам графа атаки для всех состояний устанавливаются равными нулю. Далее для каждого состояния вычисляются новые значения. Общая функция ценности выглядит следующим образом:

$$V_{i+1}^* = \max_{a \in A} \sum_{s' \in S} P(s,a,s') \left[R(s,a,s') + \gamma V_i^*(s') \right], \quad (4)$$
$$\forall s' \in S,$$

где $V_{i+1}^*(s)$ – функция ценности в состоянии *s* из множества возможных состояний *S*, если взять оптимальное действие (осуществить эксплуатацию уязвимостей); $P(s,a,s')$ – вероятность перехода, начиная от состояния s и заканчивая состоянием *s'* после выполнения действия *a*; $R(s,a,s')$ – ожидаемые награды (степень повышения уязвимостей), полученные после состояния перехода от *s* к *s'* после выполнения действия *a* с учетом дисконтирования $\gamma V_i^*(s')$.

Этот процесс повторяется до тех пор, пока значения награды не достигнут равновесного состояния, перестав изменяться. Когда атакующему отправляется отказ в приеме данных, вознаграждение сводится к минимальным значениям при расчете вероятностей наступления тех состояний, которые необходимы злоумышленнику (в соответствии с графом атаки), поскольку наблюдается отсутствие успешного влияния на систему. Начальные значения вознаграждения отражают выгоды и потери, свойственные состоянию, при котором нападающий не достиг легального взаимодействия с атакуемым объектом. Оптимальная политика в этом случае описывается следующим образом:

$$P\pi^*(s) = \arg\max \sum_{s' \in S} P(s,a,s') \left[R(s,a,s') + \gamma V_i^*(s') \right]. \quad (5)$$

Приведенный метод моделирования атак на системы ИИ с использованием МППР позволяет более точно построить вектор атак FGSM. Модели подобного рода могут использоваться также в целях прогнозирования нападений на системы ИИ [9].

Оценка потенциала эксплуатации уязвимостей (в диапазоне от 0 до 10) осуществляется в соответствии со стандартом CVSS 3.0 (Common Vulnerability Scoring System) [9, 10]. Учитывая оценки каждой из уязвимостей в графе атак, возможно оценить вероятности перехода путем нормализации оценок уязвимостей по всем ребрам графа атаки, начиная с исходного состояния системы. Соответственно, можно формально определить вероятность перехода, используя формулу (6):

$$p_{ij} = 1 - \frac{\sum rr_j}{\sum_{k=1}^{n} r_k}, \quad (6)$$

где n – количество уязвимостей, доступных из состояния i; r_j – сумма опасности уязвимостей в состоянии *j*; r_k – сумма оценок опасности всех уязвимостей всех состояний, доступных из состояния *i*.

Исходя из описанных выше условий, формируется матрица переходных вероятностей при выполнении действия a_i в отношении набора состояний.

Важным параметром модели является вознаграждение, которое отражает успех злоумышленника в случае реализации атакуемого воздействия или затраты в случае неудачи.

Ключевым фактором для данного параметра является базовая метрика уязвимости. Чем выше значение данной метрики, тем большее вознаграждение должен получать атакующий при переходе в соответствующее состояние вектора атаки. Коэффициент дисконтирования задает горизонт задачи для нарушителя и определяется в диапазоне от 0 до 1.

Состояния модели атаки описываются в соответствии с десятью тактическими наборами MITRE. Например, состояние T1 понимается как успех действий злоумышленника, который вел разведку, и, собрав необходимую информацию о целевой системе, может перейти к следующим тактикам (этапам) атаки.

Из тактик MITRE непосредственно к модели нападения по вектору состязательной атаки относятся следующие наборы:

1. T1: поиск и агрегация информации о системах, сетях и моделях ИИ, используемых в инфраструктуре.

2. T2: первичный обход модели при помощи создания состязательных данных, которые нарушают корректную классификацию моделью.

3. T3: внедрение и исполнение вредоносного программного обеспечения в системах и сетях.

4. T4.1: закрепление (сохранение доступа) в системе или сети (отравление данных модели).

5. T4.2: создание бэкдора в модели нейронной сети.

6. T7: сокрытие действий и применяемых при этом средств обнаружения.

7. T10: несанкционированный доступ и (или) воздействие на информационные ресурсы или компоненты систем и сетей, приводящие к негативным последствиям.

При визуализации графа учтены все возможные состояния (все наборы тактик MITRE) (рис. 3). Множество действий модели содержит следующее:

1. Разрешить (Allow, «a»): успешное выполнение тактики.

2. Не разрешить (Not-Allow, «n»): отсутствие успеха.

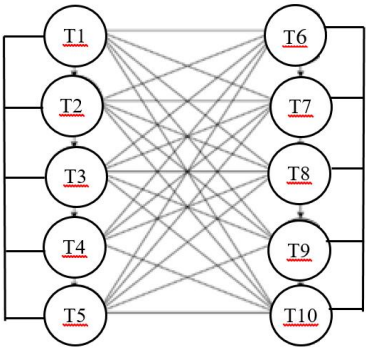


Рис. 3. Граф атаки на модель

Возможные результаты действий:

1. Отсутствие действия (not-compromised).

2. Простаивание – наблюдение не дает четко установить нынешнее состояние системы (unknown.

3. Взаимодействие – процесс использования уязвимости для достижения состояния компрометации по какому-либо этапу (compromised).

Динамика изменений заданных значений матрицы переходов между состояниями и матрицы вознаграждений при обучении приведена (фрагментарно) на рис. 4 и 5.

Можно заметить, что при нулевых и близких к нулю значениях вероятностей переходов также наблюдаются нулевые или отрицательные значения вознаграждений. Это указывает на ослабление связей между состояниями атаки (достигаемыми тактиками), которые незначительно влияют на выбор последовательности действий злоумышленника.

Результаты определения оптимальной политики показывают, что тактики T1, T2, T8 были включены в стратегию атаки (рис. 6). Таким образом, оптимальная последовательность атакующих воздействий включает всего два перехода при условии изначальной успешности достижения состояния T1.

Матрица переходов между состояниями:

```
[[[0.05789954 0.15856926 0.08844104 0.09470629 0.04214539 0.
  0.14922619 0.14593516 0.1190518  0.14402534]
 [0.         0.39669316 0.         0.47196687 0.13133997 0.
  0.         0.         0.         0.        ]
 [0.52697934 0.         0.         0.         0.         0.
  0.47302066 0.         0.         0.        ]
 [0.         0.         0.         0.45164895 0.         0.
  0.         0.54835105 0.         0.        ]
 [0.14879651 0.         0.         0.32873833 0.         0.
  0.52246516 0.         0.         0.        ]
 [0.09888238 0.         0.13605139 0.01627506 0.15588886 0.13419906
  0.22020479 0.0862781  0.07356396 0.07865638]
 [0.00189184 0.00461932 0.13226219 0.12009744 0.0384432  0.07186007
  0.13709279 0.15937011 0.14704516 0.18731788]
 [0.         0.         0.         0.         0.06642054 0.
  0.         0.         0.93357946 0.         ]
```

Рис. 4. Матрица переходов между состояниями

Матрица вознаграждений:

```
[[[-0.09357598 -0.99970764 -0.85979102  0.18980623 -0.82901452
   0.80575298 -0.83338879 -0.1045887   0.31732356 -0.95410244]
 [ 0.          0.         -0.         -0.33638672  0.
  -0.         -0.10847402 -0.         -0.91662604  0.        ]
 [ 0.          0.          0.         -0.93743775  0.97918305
  -0.57017224 -0.88202841  0.20733075 -0.93269793  0.4481216 ]
 [-0.64026389  0.74630463 -0.40566717 -0.4388521   0.29221433
  -0.69314171  0.32204563 -0.85754741 -0.35528272 -0.12167991]
 [-0.63654614  0.         -0.          0.         -0.
  -0.         -0.         -0.          0.06751705  0.55410047]
 [-0.          0.         -0.00786791 -0.         -0.
   0.         -0.         0.          0.          0.        ]
```

Рис. 5. Матрица вознаграждений

```
Iteration                Number of different  actions
   1                 1
   2                 1
   3                 0
Iterating stopped? unchanging policy found
(1, 1, 0, 0, 0, 0, 0, 1, 0, 0)
```

Рис. 6. Результат моделирования – определение оптимальной политики атаки

Классические модели машинного обучения (МО) также подвергаются атакам. Однако в силу специфики их вычислительной модели состязательные атаки мало эффективны (интерпретируемость и меньшая чувствительность к шуму делают их более устойчивыми к состязательным атакам по сравнению с нейронными сетями). В этом отношении переобучение модели не даст того же эффекта, что и при модификации нейронной сети. Однако уязвимость, представленная как допустимый диапазон отклонений входных данных, хотя и в меньшей степени, но всё же влиятт на безопасность вычислительной модели МО. Рассмотрим специфику моделирования атаки на модель машинного обучения, в которой используется метод опорных векторов.

При рассмотрении перечня техник и тактик MITRE были определены наиболее подходящие состояния для модели:

1. T1: Сбор информации в открытых репозиториях приложений.

2. T2.1: Отравление обучающих данных/

3. T2.2: Публикация вредоносных наборов данных.

4. T3: Обход модели машинного обучения.

5. T4.1: Доступ к продукту или сервису, использующему машинное обучение.

6. T4.2: Полный доступ к модели МО.

7. T4.3: Закрепление (сохранение доступа) в системе машинного обучения.

8. T5: Эксфильтрация через программные средства.

9. T7: Несанкционированный доступ и (или) воздействие на информационные ресурсы или компоненты систем и сетей, приводящие к негативным последствиям.

Действия МППР-модели следующие:

1. Разрешить (Allow, «a»): успешное выполнение тактики.

2. Не разрешить (Not-Allow, «n»): отсутствие успеха.

Далее необходимо сформировать значения вознаграждений при переходе от одной тактики к другой (табл. 4).

Таблица 4

Вознаграждения при переходе тактик

Переход	Вознаграждение
1–2	–3
1–3	–15
2–4	–7
3–4	–7
4–5	–4
4–6	–10
4–7	–6
5–8	–3
6–8	–9
7–8	–6
8–9	30

Далее необходимо указать начальные вероятности, которые определяются исходя из количества запросов от клиентского приложения к системе машинного обучения. Вычисления учитывают отсутствие взаимодействия при запросе, как и, соответственно, обращений. Тогда вероятности обращений будут выглядеть, как показанов в табл. 5.

Таблица 5

Вероятности начальных вероятностей

Состояние	Отсутствие обращений	Взаимодействие
1	1	0
2	0	1
3	0,5	0,5
4	0,5	0,5
5	0	1
6	0	1
7	0	1
8	0	1
9	0	1

Используя данные из табл. 5, можно найти вероятности реализации тактик и построить модель графа оптимальной политики.

Модель графа оптимальной политики представлена на рис. 7. Значение S представляет вероятность использования злоумышленником выбранной тактики, а значение a указывает на успех злоумышленника в достижении данного состояния. Нужно отметить, что состояние T1 соответствует успеху действия $a1$ (оно достигается из состояния, которое обычно не рассматривается как элемент графа атаки, подобное состояние можно сопоставить с началом развертывания атаки как целого процесса: от принятия решения о нападении до достижения успеха).

Последовательность проявления состояний с учетом вызвавших их действий (эксплуатаций уязвимостей) приведена на рис. 7 (номером указаны действия, обозначения $S1$–$S10$ указывают состояния, проявляемые при совершении действий).

В табл. 6 приведены итоговые вероятности сопоставленных действий и состояний вектора атаки.

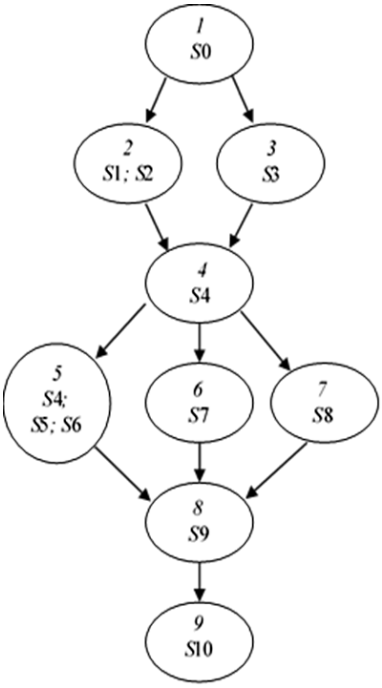


Рис. 7. Модель графа оптимальной политики злоумышленника при атаке на модель машинного обучения

Таблица 6

Итоговые вероятности сопоставленных действий и состояний вектора

Состояние	Действия								
	$a1$	$a2$	$a3$	$a4$	$a5$	$a6$	$a7$	$a8$	$a9$
S0	1,0	0	0	0	0	0	0	0	0
S1	0	0,8	1,0	0	0	0	0	0	0
S2	0	0,2	0	0	0	0	0	0	0
S3	0	0	1,0	0	0	0	0	0	0
S4	0	0	0	1,0	0,65	0	0	0	0
S5	0	0	0	0	0,25	0	0	0	0
S6	0	0	0	0	0,15	0	0	0	0
S7	0	0	0	0	0	1,0	0	0	0
S8	0	0	0	0	0	0	1,0	0	0
S9	0	0	0	0	0	0	0	1,0	0
S10	0	0	0	0	0	0	0	0	1,0

Итоги определения оптимальной политики показывают, что существует несколько вариантов атаки на модель выбранного типа машинного обучения

тактики. При этом последовательность состояний включает больше переходов, что увеличивает количество действий для осуществления атаки на модель МО, что указывает на большие трудозатраты злоумышленника при подборе средств компрометации и осуществлении атакующих воздействий, в сравнении с атакой на модель нейронной сети.

Заключение

Таким образом, при определении особенностей поиска наилучших последовательностей действий атакующего были рассмотрены специфические особенности МППР как метода моделирования атак применительно к нейронной сети и модели машинного обучения.

При построении моделей в качестве входных данных использовались метрики уязвимостей, которые были классифицированы в соответствии с методами эксплуатации уязвимостей и, следовательно, сопоставлены тактикам MITRE (следует учесть, что оценка уязвимости может носить экспертный характер, если четкого соотнесения с тактиками не наблюдается). Вектор атаки на нейросеть включает в свой состав три актуальных состояния, необходимых для достижения целей атаки, в то время как вектор атаки на модель машинного обучения предполагает множество переходов и, следовательно, достижимых состояний, что свидетельствует о более затратном для злоумышленника формировании вектора атаки.

Литература

1. Методический документ «Методика оценки угроз безопасности информации» (утв. Федеральной службой по техническому и экспортному контролю 5 февраля 2021 г.). – 83 с. [Электронный ресурс]. – Режим доступа: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-5-fevralya-2021-g, свободный (дата обращения: 25.07.2024).

2. MITRE ATLAS // MITRE ATT&CK [Электронный ресурс]. – Режим доступа: https://atlas.mitre.org, свободный (дата обращения: 02.05.2024).

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

4. Markov decision process for automatic cyber defense (WISA–2023) / Xiaofan Zhou, Simon Yusuf Enoch, Dan Dong Seong Kim [Электронный ресурс]. – Режим доступа: https:// https://arxiv.org/abs/2207.05436, свободный (дата обращения: 13.05.2024).

5. Booker L.B. A model-based, decision-theoretic perspective on automated cyber response / L.B. Booker, S.A. Musman [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2002.08957, свободный (дата обращения: 13.05.2024).

6. Zheng J. Defending SDN-based IoT networks against ddos attacks using markov decision process / J. Zheng, A.S. Namin // Proceedings – 2018. IEEE International Conference on Big Data, Big Data–2018. – 2018. – P. 4589–4592.

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

8. Саттон Р.С. Обучение с подкреплением: введение / Р.С. Саттон, Э.Дж. Барто. – 2-е изд. – М.: ДМК-Пресс, 2020. – 552 с.

9. Mazengia D.H. Forecasting spot electricity market prices using time series models. Thesis for the degree of master of science in electric power engineering. – Gothenburg: Chalmers University of Technology, 2008. – 89 p.

10. Common Attack pattern enumerations and classifications [Электронный ресурс]. – Режим доступа: https://capec.mitre.org/, свободный (дата обращения: 12.05.2024).

11. Common vulnerability scoring system calculator [Электронный ресурс]. – Режим доступа: https://nvd.nist.gov/vuln-metrics/cvss/v2-calculator, свободный (дата обращения: 13.05.2024).

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

Podtopelny V.V.

Features of modeling the attacks on the machine learning model using Markov decision-making processes

The issues that arise when determining the problem of modeling the impact on artificial intelligence models that are integrated into the network infrastructure are considered. Various research methods are presented and characterized, including the MITER methodology used in constructing the network impact vector. The features of using Markov decision-making processes in modeling attack influences are considered. Their significance for various procedures for determining vector parameters is considered. When constructing the modeling specifics, the authors take into account the features of determining the vulnerabilities of artificial intelligence systems. The vector impact of major exploits of vulnerabilities is being studied.

Keywords: network attack, vulnerability, Markov process modeling, strategy, policy, teaching method, artificial intelligence.
DOI: 10.21293/1818-0442-2024-27-2-21-30

References

1. Metodicheskij dokument «Metodika ocenki ugroz bezopasnosti informacii» (utv. Federalnoj sluzhboj po tehnicheskomu i eksportnomu kontrolyu 5 fevralya 2021 g.). 83 p. Available at: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-5-fevralya-2021-g, free (Accessed: July 25, 2024).

2. MITRA ATLAS // MITRA ATLANT Available at: https://atlas.mitre.org, free (Accessed: May 02, 2024).

3. Namiotsky D.E. [Shemy atak na modeli mashinnogo obucheniya]. *International Journal of Open Information Technologies*, 2023, vol. 11, no. 5, pp. 68–86 (in Russ.).

4. Xiaofan Zhou, Simon Yusuf Enoch, Dan Dong Song Kim. Markov decision-making process for automatic cyberbullying (WISA 2023). Available at: https://arxiv.org/abs/2207.05436, free (Accessed: May 13, 2024).

5. Booker L.B., Musman S.A. A model-based view of automated implementation in cyberspace based on the theory

of decision-making. Available at: https://arxiv.org/abs/2002.08957, free (Accessed: May 13, 2024).

6. Zheng J., Namin A.S. [Protection of the Internet of Things based on SDN from DDoS attacks using the Markov decision-making process]. *Proceedings of the IEEE International Conference on Big Data 2018, Big Data–2018*. 2018, pp. 4589–4592.

7. Kohenderfer M., Wheeler T., Ray K. *At the decision-making algorithm*. M.: DMK-Press, 2023, 684 p. (in Russ.).

8. Sutton R.S., Barto E.J. Algoritmy prinyatiya reshenij. *Engagement with reinforcement: An introduction.* 2nd ed. Moscow, DMK-Press, 2020, 552 p.

9. Mazengia D.H. *Forecasting prices in the electricity market using time series models: dissertation for the degree of Master of Science in the field of electric power engineering*. Gothenburg, Chalmers University of Technology, 2008, 89 p.

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

11. Calculator of the general capability assessment system Available at: https://nvd.nist.gov/vuln-metrics/cvss/v2-calculator, free (Accessed: May 13, 2024).

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