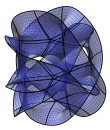

DOI: 10.18287/2541-7525-2024-30-4-147-160 УДК 004.056.5:378.3

Особенности использования марковских процессов принятия решений при моделировании атак на системы искусственного интеллекта

Ветров*[1] *И.А.*, *Подтопельный*[2] *В.В.

[1] *Балтийский федеральный университет имени И. Канта, г. Калининград, Российская Федерация; vetrov.gosha2009@yandex.ru (И.А.);*
[2] *Калининградский государственный технический университет (КГТУ), г. Калининград, Российская Федерация; ionpvv@mail.ru (В.В.);*

Поступила: 14.10.2024
Рассмотрена: 20.11.2024
Принята: 25.11.2024

Научная статья

Аннотация. В данной работе проводится исследование особенностей моделирования атак на системы искусственного интеллекта. При построении модели используются марковские процессы принятия решений. Предлагается многоуровневый подход к интерпретации состояний системы, включающий несколько этапов детализации состояний. Этот подход основан на методологии MITRE ATLAS и Методике оценки угроз ФСТЭК. При формировании вектора учитывается специфика модели нарушителя и рассматриваются два основных режима моделирования: on-time и off-time. Описывается порядок формирования наград на абстрактном уровне (без конкретизации действий злоумышленника) построения модели.

Ключевые слова: сетевая атака; уязвимость; марковские процессы; моделирование; стратегия; политика; метод обучения.

Введение

Современные системы искусственного интеллекта становятся объектом хакерских атак. Специфика таких нападений заключается в возможности манипулирования логикой вычислительной модели и данными датасетов. При атаках часто используется внедрение ложной информации в общий поток данных при доступности интерфейса ввода данных. Для предотвращения подобных нападений необходимо выявить возможные векторы атак на системы искусственного интеллекта [1; 2]. При этом следует учитывать специфику реагирования различных вычислительных моделей ИИ на атакующие воздействия:

1. Нейронные сети и модели машинного обучения (линейная регрессия, решающие деревья) проявляют высокую уязвимость к состязательным атакам.
2. Средняя степень уязвимости наблюдается к атакам типа «отравление данными», которые могут ухудшить производительность всех типов моделей.
3. Атаки типа «кража модели» также имеют среднюю степень уязвимости; они особенно опасны для моделей с программными интерфейсами (API).
4. Атаки типа «определение принадлежности» направлены на поиск конкретных примеров из обучающих данных и уязвимы для нейронных сетей.
5. Атаки типа «смена меток» могут снизить производительность моделей всех типов.
6. Нейросети демонстрируют низкую степень уязвимости к атакам типа «инверсия модели».

При изучении атак следует учитывать специфику действий злоумышленника. В этом случае множество действий целесообразно разделить на несколько групп. Первая группа действий связана с возможностями злоумышленника взаимодействовать с вычислительной моделью ИИ, получать полный доступ к исходным данным и алгоритмам ее работы в течение атаки. Вторая группа связана с ограниченными возможностями злоумышленника. Подразумевается, что у злоумышленника отсутствует доступ к любым данным, раскрывающим параметры и особенности построения вычислительной модели, но при этом нарушителю доступен входящий и исходящий поток данных, используя который можно раскрыть характеристики системы ИИ (модель активно тестируется сконструированными запросами, или воссоздается вычислительная модель на стороне атакующего с последующим ее тестированием). Третья группа особенностей подразумевает наличие у злоумышленника возможностей первой группы с ограничениями следующего типа: отсутствует прямой отклик модели, присутствует неполнота сведений о классификационной модели ИИ. Кроме того, следует выделить те атаки, которые не связаны с активным взаимодействием с моделью ИИ (обычно это атаки на различные датасеты). В этом случае цель злоумышленника — исказить параметры при обучении модели таким образом, чтобы вызвать в процессе классификации данных необходимую злоумышленнику реакцию от системы ИИ. В целом можно отметить, что атакующие воздействия возможны при наличии следующего:

1. Известность и распространенность моделей (типовые или индивидуальные).

2. Доступность обучающих данных, прототипов и интерфейсов — чем больше данных для обучения модели находится в общем доступе, тем выше ее уязвимость.

Если рассмотреть действия злоумышленника более детально, можно выделить основные группы уязвимостей:

1. Общий доступ к базам данных и датасету проблемной области вычислительной модели.

2. Отсутствие верификации источников данных, что позволяет изменять данные или метки данных.

Также стоит отметить сопутствующие уязвимости:

1. Слабые механизмы аутентификации и авторизации.

2. Доступность программных интерфейсов.

3. Отсутствие регистрации и учета событий, связанных с изменением источников данных датасетов.

4. Возможность использования методов социальной инженерии.

5. Слабая сетевая защита и неправильная конфигурация сетевой инфраструктуры.

В условиях активного внедрения искусственного интеллекта в современные информационные системы вопрос доверия к исходным обучающим данным и технологиям, используемым в процессе обучения, становится ключевым фактором, определяющим безопасность вычислительных моделей. Соответственно, необходимо исследовать специфику моделирования атак с учетом доступности данных систем ИИ в процессах аудита, рассмотреть аспекты моделирования атакующих воздействий с учетом современных методик описания сценариев атак. При этом следует учитывать, что процедура моделирования должна быть соотнесена с существующими методиками построения сценариев атак, то есть применима в практике аудита информационной безопасности.

1. Исследование специфики атак на системы ИИ и определение специфики моделирования

Существуют множество методов, которые используются для построения и анализа сценариев атак (векторов атак):

1. Деревья атак.
2. Методы, связанные с использованием машинного обучения.
3. Байесовские сети доверия.
4. Сети Петри — Маркова.
5. Теория нечетких множеств.
6. Теория игр.
7. Теория графов.
8. Теория случайных процессов.

Для определения вектора атаки часто используют модель на основе марковских процессов. Это обусловлено тем, что марковские процессы позволяют учитывать фактор неопределенности (результаты действий злоумышленника не всегда предсказуемы), также присутствует возможность структурирования порядка изучаемых действий и состояний. Применение марковских процессов для исследования компьютерных атак рассматривается в ряде научных работ, что доказывает их применимость в задачах анализа атак в сфере информационной безопасности [3–7].

Отдельно нужно отметить преимущества моделирования на основе марковских процессов принятия решений (МППР): модель позволяет принимать оптимальные решения в условиях неопределенности и изменяющейся среды; учитывать последствия принятых решений и строить долгосрочные и оптимальные стратегии (посредством использования алгоритмов максимизации ожидаемой награды), что особенно полезно в задачах прогнозирования атакующих воздействий; применять методы динамического программирования и обучения с подкреплением.

При этом существуют ограничения у данного способа моделирования: марковский процесс принятия решений предполагает, что состояние системы полностью описывает текущую ситуацию и не учитывает историю действий и состояний. Этого может быть недостаточно для некоторых задач, где порядок прошедших действий и состояний играет важную роль; переходы между состояниями являются стохастическими и марковскими, то есть не зависят от предыдущих состояний и действий. В реальных задачах это может быть неверным предположением; существуют проблемы «проклятия размерности», когда количество состояний и действий становится слишком большим, что затрудняет решение и требует больших вычислительных ресурсов. Несмотря на эти ограничения, марковский процесс принятия решений остается мощным инструментом для моделирования и решения задач в различных областях, таких как робототехника, финансы, управление производством и другие. В целом марковский процесс принятия решений является важным инструментом для принятия оптимальных решений в условиях неопределенности. Марковская модель атаки на системы ИИ описывается кортежем состояний действий, наград (S, A, P, R,γ). При построении модели используются следующие обозначения:

1. $S = \{s_1, s_2, ..., s_n\}$ — множество вершин-состояний системы, в которых находится атакуемая модель ИИ (состояние интерпретируется по методике MITRE) [8; 9].
2. $A = \{(s_i, s_j) | s_i, s_j \in S, a \in \{1, \ldots, 11\}\}$ — множество действий злоумышленника в процессе атаки.

3. $R\left(s, a, s^{'}\right)$ — награда за успешное достижение состояний в результате реализации перехода.

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

5. $\pi(s, a)$, $\pi(a|s) = P[A_i = a|S_i = s]$ — функция, описывающая распределение вероятностей выбора действий злоумышленника в состоянии $s \in S$, которое соответствует достигнутому этапу атаки.

6. $V_\pi(s)$ — ценность состояния — это величина характеризует вознаграждение нападающего. В данной модели она сопоставлена с метрикой уязвимости и набором эксплуатирующих действий в границах стратегии атаки π. Определяется в соответвии с формулой (1)

$$V_\pi(s) = M\left[R_{i+1} + \gamma V_\pi\left(S_{i+1}\right)|S_i = s\right],\ s \in S. \tag{1}$$

7. $V^*_{i+1}(s)$ — функция ценности состояния как достигнутой тактики, которую возможно использовать для следующего перехода (2).

$$V^*_{i+1}(s) = \max_{a \in A} \sum_{s' \in S} P\left(s, a, s^{'}\right)\left[R\left(s, a, s^{'}\right) + \gamma V^*_i\left(s'\right)\right],\ \ \forall s^{'} \in S. \tag{2}$$

8. γ — коэффициент дисконтирования.

9. M — математическое ожидание случайной величины.

10. $G_i = R_{i+1} + \mathrm{yR}_{i+2} + y^2 R_{i+3} + \ldots = \sum_{k=0}^{\infty} y^k R_{i+k+1}$ — функция ценности на i-м шаге.

11. $\pi^*(s)$ — функция оптимальной политики (3):

$$\pi^*(s) = \operatorname*{argmax}_{a \in A} \sum_{s' \in S} P\left(s, a, s^{'}\right)\left[R\left(s, a, s^{'}\right) + \gamma V^*_i\left(s'\right)\right]. \tag{3}$$

Количество состояний и переходов в системе зависит от ее инфраструктуры, логической организации вычислительной модели и ее алгоритмов. Чтобы определить стратегию поведения злоумышленника, можно использовать моделирование на основе марковских процессов принятия решений (МППР). В этом случае политики будут стационарными, то есть не зависящими от времени.

В процессе моделирования необходимо ввести параметры, которые описывают затраты на реализацию стратегии нападения. Эти затраты включают в себя обман механизмов классификации систем идентификации (СИИ) на уровне весов нейросети атак (модификация классифицирования) и на уровне разделения источников данных на доверенные и недоверенные (навязывание ложных обучающих данных).

Анализ параметров, влияющих на принятие оптимальных решений злоумышленником, является важной задачей. Оптимальная стратегия нападения строится с помощью уравнения Беллмана [7].

2. Формирование модели

Выбор типа состояния зависит от уровня абстракции модели и характера действий злоумышленника. При более детальном анализе используются состояния, определяемые как успех действий злоумышленника. Эти действия классифицируются как тактики и техники, соответствующие методологиям MITRE Atlas или ФСТЭК [8; 9].

Однако можно абстрагироваться от конкретных действий злоумышленника, сосредоточившись на общих фазах атаки. В этом случае можно выделить несколько состояний:

1. Контролируемое взаимодействие. В этом состоянии злоумышленник и атакуемая сторона полностью контролируют канал передачи данных и реакцию атакуемой вычислительной модели при осуществлении нелегитимных действий.

2. Легитимное взаимодействие. Злоумышленник может взаимодействовать с вычислительной моделью в обычном режиме, не оказывая на нее компрометирующего воздействия.

3. Доверенное взаимодействие. Злоумышленник взаимодействует с вычислительной моделью, которая не различает легитимные и компрометирующие воздействия на ее алгоритмы и данные.

Ключевым фактором в этой модели является вычисление наград. Изначально все награды равны нулю. Затем при переходах вычисляются новые награды, и для предотвращения неправильной оценки (максимизации ценности множества неэффективных действий) используется коэффициент дисконтирования, равный 0,9. Это позволяет найти наиболее эффективный путь, то есть набор действий злоумышленника, который приводит к наибольшим наградам. Логика получения наград определяется в зависимости от условий моделирования и типа атаки. В случае успешного перехода злоумышленник получает более высокую награду, что служит стимулом для продолжения его действий. Если же действия не приносят успеха, награды либо отсутствуют, либо становятся отрицательными.

Таким образом, можно выделить четыре основных состояния, которые наблюдаются в режиме онлайн, то есть когда происходит атака на вычислительную модель:

- контролируемое взаимодействие (C);
- легальное взаимодействие (L);
- доверенное взаимодействие (T);
- блокировка (B).

Модель взаимодействия приведена на рисунке 2.1. При моделировании атак на системы искусственного интеллекта используются абстрактные состояния действия, которые приводят к изменениям и появлению новых состояний системы. В результате выстраивается определенный порядок действий и состояний.

Важно отметить, что злоумышленник должен быть осведомлен об изменениях системы. В противном случае он будет использовать шаблонные действия или алгоритмы вредоносных программ, не имея возможности контролировать или отслеживать состояние атакуемой системы. Такой подход к моделированию соответствует типологии атак на системы ИИ. Если злоумышленник способен получать отклик от системы в реальном времени, эти атаки можно отнести к атакам «белого» или «серого ящика». В случае отсутствия отклика нападение может рассматриваться как атака типа «черный ящик». При моделировании важно учитывать логику злоумышленника. Если при моделировании определяется наилучший порядок действий без учета особенностей модели нарушителя, то считается, что он является высококвалифицированным и имеет максимальный доступ к атакующим инструментам. Такой режим можно отнести к режиму «вне контекста времени атаки», который ограничивает возможности нарушителя (режим off-time). Полностью учесть модель нарушителя сложно, так как она неизвестна в момент атаки. Однако, если учитывать вариативность модели нарушителя, можно вводить специальные действия, которые будут характеризовать нарушение логики наилучшего пути реализации атакующих воздействий. Это покажет, что злоумышленник не всегда действует наилучшим образом (режим on-time). В этой логике злоумышленник может возвращаться в нейтральное состояние при поиске наилучшего альтернативного пути. Если он продолжит атакующие действия в контексте контролируемого взаимодействия, система переведет систему в состояние блокировки. Таким образом, злоумышленник будет извещен о негативных последствиях для него, если у него нет доступа к интерфейсу вычислительной модели.

Приведенные состояния (рис. 1) описывают модель атаки типа «отравление данными» в контексте взаимодействия двух сторон в реальном режиме времени (режим on-time). При этом следует учитывать, что модель, описывающая сценарий предполагаемой атаки, определяемой при аудите информационных систем, создается с учетом поиска наилучшей последовательности действий злоумышленника вне контекста противодействия в реальном режиме реального времени защищающейся стороны (режим off-time).

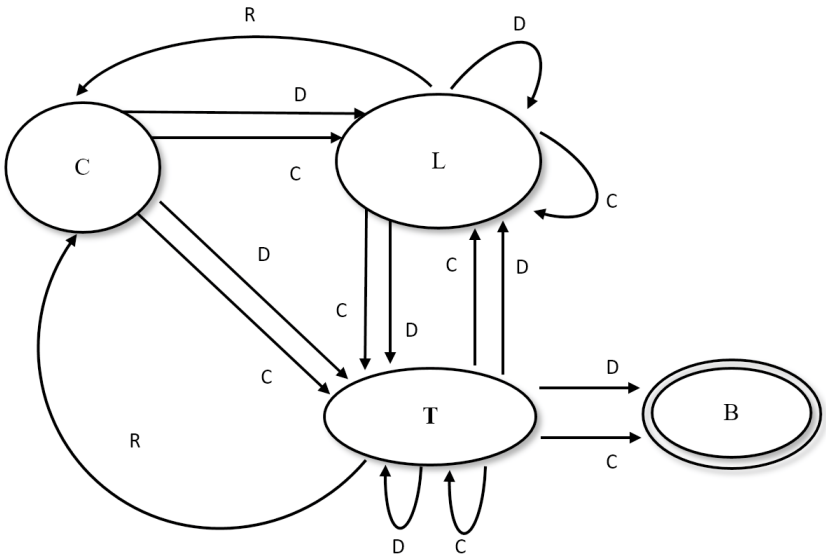

Рис. 1. Модель МППР для атак в режиме on-time
Fig. 1. MDP-model for on-time data poisoning attacks

Для построения сценария атак подобного рода часто используют методологию MITRE ATLAS (далее — MITRE) [4; 10]. На рисунке 2 приведена упрощенная (абстрагированная) классификация тактик MITRE, в которой тактики объединены на основе сходства функционального назначения техник в несколько классов, а именно:

1. Разведка предполагает выявление слабостей (уязвимостей к воздействию на системы ИИ), проводится на разных этапах атаки: на начальном и промежуточных этапах при продвижении в последовательности компрометирующих действий.

2. Сопровождение включает в свой состав подготовку средств эксплуатации уязвимостей и условий для успешного нападения. Также может осуществляться на начальном и промежуточных этапах.

3. Компрометация содержит действия, направленные на извинение порядка функционирования СИИ, в том числе ее вычислительной модели.

4. Компрометация — это заключительная фаза, наступление которой означает — достигнут успех атаки и злоумышленник нанес ущерб системе.

Изначально злоумышленнику доступно нейтральное состояние СИИ. Модель МППР для сценария атаки в режиме off-time на модели и системы ИИ предполагает также четыре состояния, с которыми злоумышленник может столкнуться при взаимодействии с СИИ [9]:

- разведка (T1);
- подготовка (T2);
- компрометация (T7);
- достижение целевого состояния (*T12*).

Рис. 2. Упрощение методологии MITRE ATLAS
Fig. 2. Simplification of the MITER ATLAS methodology

Модель взаимодействия приведена на рисунке 3. Отсутствие обратных дуг, обозначающих действие R, в данной модели обусловлено характером построения сценария при пассивном аудите (определяется наилучший сценарий атаки злоумышленника при учете того, что злоумышленнику известно все об атакуемой системе, средства защиты предустановлены, а система при этом изменяется только под действиями злоумышленника). В этом случае откат в предыдущее состояние означает отсутствие успеха в выбранной стратегии поведения злоумышленника, и, следовательно, возникает потребность в смене вектора атаки.

При моделировании атакующих воздействий доступны следующие стратегии (действия):

1. Обход системы (обман) (D).
2. Легитимное взаимодействие с системой (C).
3. Откат до предыдущего состояния (R).

Злоумышленник начинает с предоставления легитимных данных, чтобы завоевать доверие системы искусственного интеллекта (ИИ). Затем он вводит возмущения, которые приводят к неверной классификации данных и получению вознаграждения R. Начальные значения этого вознаграждения отражают как выгоды, так и потери, которые он понес до достижения легального взаимодействия с системой.

Размер вознаграждения R за действия a определяется на основе следующего принципа: величина ресурсов, затраченных на обход модели (обман), должна превышать вознаграждение

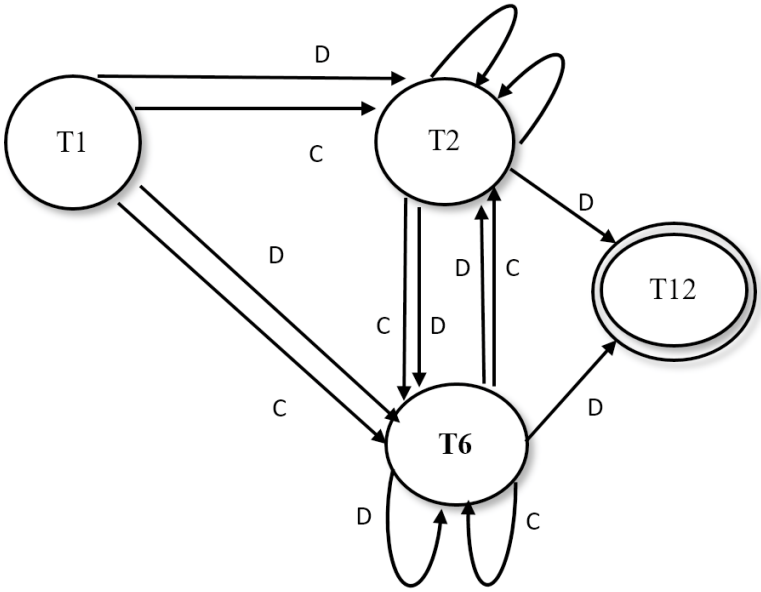

Рис. 3. MDP-модель для off-time, связанных с отравлением данных
Fig. 3. MDP-model for off-time, connected with Data Poisoning attacks

за легальное взаимодействие. В случае отката (блокировки действий злоумышленника) вознаграждение отсутствует, а при переходе в предыдущее состояние оно заменяется штрафом. Таким образом, порядок вознаграждений R за каждое новое состояние будет соответствовать следующему набору правил:

1. Для моделирования в режиме on-time: $R(CT) > R(LT) > R(LC) > R(CC)\ R(C-$ $-L)\ R(TT) > R(TC)\ R(TL)R(CB)$.

2. Для моделирования в режиме off-time: $R(T1T7) > R(T2T7) > R(T2T1) >$ $> R(T1CT1)\ R(T1-T2)\ R(T7T7) > R(T7T1)\ R(T7T2) > R(T1T12)$.

При моделировании начальные значения для всех состояний устанавливаются равными нулю. Далее для каждого состояния вычисляются новые значения (3; 4). Этот процесс повторяется до тех пор, пока значения не достигнут равновесного состояния и не изменятся. Кроме того, учитывается максимальное количество повторений (например, 1000), чтобы избежать попадания в бесконечный цикл, когда значения меняются очень незначительно.

Итерация значений — это простой итерационный алгоритм для определения оптимальной функции значений V*, которая сходится к правильным значениям.

При создании модели важно учитывать ряд ограничений, которые соответствуют логике осуществления атак на модель искусственного интеллекта. Во-первых, все переходы (множество A) из одного состояния в другое должны в сумме составлять 1. Это условие важно для направленных графов модели. Во-вторых, если одновременно применяются как легитимные, так и нелигитимные действия, то их суммарная вероятность должна превышать вероятности только легитимных или только нелигитимных переходов. Если реализуются нелигитимные действия, то вероятность попадания в состояние, при котором повышается вероятность последующего перехода в заблокированное состояние, должна быть выше, чем вероятность того, что новое состояние будет полностью доверенным (действия злоумышленника признаются легитимными).

Ключевым фактором в построении модели является определение величин ценности состояний, что позволяет выявить наилучшую последовательность фаз (состояний) атаки с точки зрения злоумышленника. Определение значений для каждого состояния с использованием функции полезности модели на основе МППР для общей атаки в режиме on-time на СИИ приведено в выражениях ниже. Нужно отметить, что подобный подход уже использовался при рассмотрении аспектов компрометации с применением социальной инженерии, но при этом не

использовалась интерпретация состояний с учетом специфики MITRE ATLAS, Методики ФСТ-ЭК, не применялись режимы построениям модели (on-line, on-line), не учитывалось наличие намеренных действий, связанных со сбросом достигнутых состояний [11].

$$(V_{i+1}^{*}(s=T)=\max_{a\in A}\sum_{s'\in S}P\left(s,a,s'\right)\ \left[R\ \left(s,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right]. \tag{4}$$

$$V_{i+1}^{*}\ (s=T)=\max\begin{cases}P(T,D,L)\,[R(T,D,L)+\gamma V_{i}^{*}(L)]+\\ P(T,D,T)\,[R(T,D,T)+\gamma V_{i}^{*}(T)]<(D)>\\ ---------\\ P(T,C,L)\,[R(T,C,L)+\gamma V_{i}^{*}(L)]+\\ P(T,D,T)\,[R(T,C,T)+\gamma V_{i}^{*}(T)]<\ (S)>\\ ---------\\ P(T,R,C)\,[R(T,R,C)+\gamma V_{i}^{*}(C)]<(R>\end{cases}. \tag{5}$$

$$V_{i+1}^{*}(s=L)=\max_{a\in A}\sum_{s'\in S}P\left(L,a,s'\right)\ \left[R\ \left(L,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right]. \tag{6}$$

$$V_{i+1}^{*}\ (s=L)=\max\begin{cases}P(L,D,L)\,[R(L,D,L)+\gamma V_{i}^{*}(L)]+\\ P(L,D,B)\,[R(L,D,B)+\gamma V_{i}^{*}(B)]+\\ P(L,D,T)\,[R(L,D,T)+\gamma V_{i}^{*}(T)]<(D)>\\ ---------\\ P(L,C,L)\,[R(L,C,L)+\gamma V_{i}^{*}(L)]+\\ P(L,C,B)\,[R(L,C,B)+\gamma V_{i}^{*}(B)]+\\ P(L,C,T)\,[R(L,C,T)+\gamma V_{i}^{*}(T)]<\ (S)>\\ ---------\\ P(L,R,C)\,[R(L,R,C)+\gamma V_{i}^{*}(C)]<(R>\end{cases}. \tag{7}$$

$$V_{i+1}^{*}(s=C)=\max_{a\in A}\sum_{s'\in S}P\left(C,a,s'\right)\ \left[R\ \left(C,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right]. \tag{8}$$

$$V_{i+1}^{*}(s=C)=max\begin{cases}P(C,D,L)\,[R(C,D,L)+\gamma V_{i}^{*}(L)]+\\ P(C,D,B)\,[R(C,D,B)+\gamma V_{i}^{*}(C)]<(D)>\\ ---------\\ P(C,C,L)\,[R(C,C,L)+\gamma V_{i}^{*}(L)]+\\ P(C,C,C)\,[R(C,C,C)+\gamma V_{i}^{*}(C)]<(S)>\end{cases}. \tag{9}$$

При определении значений для каждого состояния с использованием функции полезности модели на основе МППР для общей атаки в режиме off-time на СИИ не предполагается возвратных состояний.

$$V_{i+1}^{*}(s=T6)=\max_{a\in A}\sum_{s'\in S}P\left(s,a,s'\right)\ \left[R\ \left(s,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right]. \tag{10}$$

$$V_{i+1}^{*}\ (s=T6)=\max\begin{cases}P(T6,D,T2)\,[R(T6,D,T2)+\gamma V_{i}^{*}(T2)]+\\ P(T6,D,12)\,[R(T6,D,B)+\gamma V_{i}^{*}(T12)]+\\ P(T6,D,T6)\,[R(T6,D,T6)+\gamma V_{i}^{*}(T6)]<(D)>\\ ---------\\ P(T6,C,T2)\,[R(T6,C,T2)+\gamma V_{i}^{*}(T2)]+\\ P(T6,D,T6)\,[R(T6,C,T6)+\gamma V_{i}^{*}(T6)]<(S)>\\ ---------\\ P(T6,R,T1)\,[R(T6,R,T1)+\gamma V_{i}^{*}(T1)]<(R)>\end{cases}. \tag{11}$$

$$V_{i+1}^{*}\ (s=T2)=\max\begin{cases} P\left(T2,D,T2\right)\left[R\left(T2,D,T2\right)+\gamma V_{i}^{*}\left(T2\right)\right]+ \\ P\left(T2,D,T12\right)\left[R\left(T2,D,T12\right)+\gamma V_{i}^{*}\left(T12\right)\right]+ \\ P\left(T2,D,T6\right)\left[R\left(T2,D,T6\right)+\gamma V_{i}^{*}\left(T6\right)\right]<(D)> \\ ---------- \\ P\left(T2,C,T2\right)\left[R\left(T2,C,T2\right)+\gamma V_{i}^{*}\left(T2\right)\right]+ \\ P\left(T2,C,T6\right)\left[R\left(T2,C,T6\right)+\gamma V_{i}^{*}\left(T6\right)\right]<(S)> \\ ---------- \\ P\left(T2,R,T1\right)\left[R\left(T2,R,T1\right)+\gamma V_{i}^{*}\left(T1\right)\right]<(R)> \end{cases}. \tag{12}$$

$$V_{i+1}^{*}\left(s=T2\right)=\max_{a\in A}\sum_{s'\in S}P\left(T2,a,s'\right)\ \left[R\ \left(T2,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right]. \tag{13}$$

$$V_{i+1}^{*}\left(s=T1\right)=\max_{a\in A}\sum_{s'\in S}P\left(T1,a,s'\right)\ \left[R\ \left(T1,a,s'\right)+\gamma V_{i}^{*}\left(s'\right)\right] \tag{14}$$

$$V_{i+1}^{*}(s=T1)=\max\begin{cases} P\left(T1,D,T2\right)\left[R\left(T1,D,T2\right)+\gamma V_{i}^{*}\left(T2\right)\right]+ \\ P\left(T1,D,T6\right)\left[R\left(T1,D,T6\right)+\gamma V_{i}^{*}\left(T6\right)\right]<(D)> \\ ---------- \\ P\left(T1,C,T2\right)\left[R\left(T1,C,T2\right)+\gamma V_{i}^{*}\left(T2\right)\right]+ \\ P\left(T1,C,C\right)\left[R\left(T1,C,T6\right)+\gamma V_{i}^{*}\left(T6\right)\right]<S)> \end{cases}. \tag{15}$$

Приведенные значения вознаграждений должны быть масштабированы в заранее определенном диапазоне, чтобы избежать поверхностного эффекта больших вознаграждений. При атаке, когда цели передаются компрометирующие данные, злоумышленник может быть заблокирован, то есть величина выгоды в этом случае будет стремиться к нулевым значениям. Моделирование в соответствии с приведенным порядком вычисления $V_{i+1}^{*}(s)$ должно показывать компромисс между затратами, действиями и последствиями различных стратегий, доступных злоумышленникам, что в итоге позволяет определить наилучшую стратегию нападения. Демонстрация работы модели для режима on-line показывает наилучшие пути реализации атак в ситуации «серый ящик» при заполненных матрицах вознаграждений и матрицах вероятностей переходных состояний (рис. 4). Атака производится на системы ИИ, которые используют нейросеть GAN. Состояния классифицируются в соответствии с фазами атак (упрощение методологии MITRE ATLAS, рис. 3).

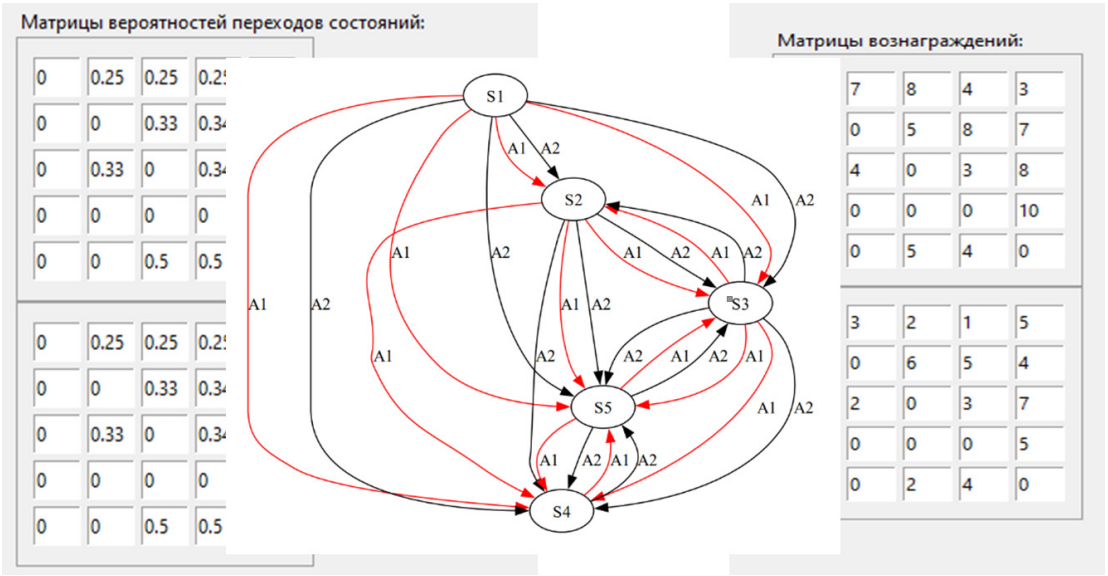

Рис. 4. MDP-модель для off-time, связанных с отравлением данных
Fig. 4. MDP-model for off-time, connected with Data Poisoning attacks

В этом случае награды за переход A1 (успех) определяются как мера уязвимости в текущем состоянии. Награды за переход A2 (не успех) можно определять как величину возможного отката (результат осуществления контрмер, закрывающих уязвимость для действия A1). Контрмеры рассматриваются как противодействие влиянию нелегитимных данных методами машинного обучения [11; 12]. Таким образом, если $A1_{ij} < A2_{\text{ij}}$, то контрмеры перекрывают текущие уязвимости в данном элементе матрицы. Тогда все оптимальные действия для злоумышленника успешны. Злоумышленник может пойти по любому из предоставленных путей в случае равновероятных событий.

Выводы

Таким образом, в процессе исследования проблемной области атак на ИИ были сформированы модели определения наилучших политик (действий) нападения злоумышленника для режимов on-time и off-time с учетом методов MITRE ATLAS и методологии ФСТЭК. Порядок моделирования атак на системы ИИ с использованием МППР, учитывая требования оптимальной политики нападения, позволяет более точно формировать векторы атак, которые используют методы навязывания ложных данных, а также модифицирования логики вычислительной модели. Важной особенностью процесса моделирования является учет специфики формирования наград за переходы между состояниями развертываемой атаки.

Информация о конфликте интересов: авторы заявляют об отсутствии конфликта интересов.

Цитирование. Ветров И.А., Подтопельный В.В. Особенности использования марковских процессов принятия решений при моделировании атак на системы искусственного интеллекта // Вестник Самарского университета. Естественнонаучная серия / Vestnik of Samara State University. Natural Science Series. 2024. Т. 30, № 4. С. 147–160. DOI: 10.18287/2541-7525-2024-30-4-147-160.

Игорь Анатольевич Ветров (vetrov.gosha2009@yandex.ru) – кандидат технических наук, доцент, методический руководитель по УГСНП «Информационная безопасность», Образовательно-научный кластер «Институт высоких технологий», Балтийский федеральный университет имени И. Канта, 236041, Российская Федерация, г. Калининград, ул. Александра Невского, 14.
Владислав Владимирович Подтопельный (ionpvv@mail.ru) – старший преподаватель, Институт цифровых технологий, Калининградский государственный технический университет, 236022, Российская Федерация, г. Калининград, Советский пр., 1.

Литература

[1] Котенко И.В., Саенко И.Б., Лаута О.С., Васильев Н.А., Садовников В.Е. Атаки и методы защиты в системах машинного обучения: анализ современных исследований // Вопросы кибербезопасности. 2024. № 1 (59). С. 24–37. DOI: https://doi.org/10.21681/2311-2024-1-24-37. EDN: https://elibrary.ru/izqdls.

[2] Намиот Д.Е. Схемы атак на модели машинного обучения // International Journal of Open Information Technologies. 2023. Т. 11, № 5. С. 68–86. URL: https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya?ysclid=m5dmh9jnct363700583.

[3] Xiaofan Zhou, Simon Yusuf Enoch, Dan Dong Seong Kim. Markov Decision Process For Automatic Cyber Defense. You I., Youn T.Y. (eds.) Information Security Applications. WISA

2022. Lecture Notes in Computer Science, vol 13720. Springer, Cham, 2023. P. 313–329. DOI: https://doi.org/10.1007/978-3-031-25659-2_23.

[4] Booker L.B., Musman S.A. A Model-Based, Decision-Theoretic Perspective on Automated Cyber Response. DOI: https://doi.org/10.48550/arXiv.2002.08957.

[5] Zheng J., Namin A.S. Defending SDN-based IoT Networks Against DDoS Attacks Using Markov Decision Process. 2018 IEEE International Conference on Big Data (Big Data). DOI: https://doi.org/10.1109/BigData.2018.8622064.

[6] Щеглов А.Ю. Защита компьютерной информации от несанкционированного доступа. Санкт-Петербург: Наука и Техника, 2004. 384 с. URL: https://reallib.org/reader?file=523140.

[7] Кохендерфер М., Уилер Т., Рэй К. Алгоритмы принятия решений / пер. с англ. В.С. Яценкова. Москва: ДМК Пресс, 2023. 684 с. URL: https://znanium.ru/catalog/document?id=445338.

[8] Методический документ «Методика оценки угроз безопасности информации» (утв. Федеральной службой по техническому и экспортному контролю 5 февраля 2021 г.). Москва, 2021. 83 с. URL: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-5-fevralya-2021-g.

[9] MITRE ATLAS // MITRE ATT&CK [Электронный ресурс]. URL: https://atlas.mitre.org, свободный (дата обращения: 02.05.2024)

[10] Горбачев И.Е., Глухов А.П. Моделирование процессов нарушения информационной безопасности критической инфраструктуры // Труды СПИИРАН. 2015. Вып. 1 (38). С. 112–135. URL: https://www.elibrary.ru/item.asp?id=23342077. EDN: https://elibrary.ru/tquqzx.

[11] Faranak Abri, Jianjun Zheng, Akbar Siami Namin, Keith S. Jones. Markov Decision Process for Modeling Social Engineering Attacks and Finding Optimal Attack Strategies // IEEE Access. 2022. Vol. 10. P. 109949–109968. DOI: http://doi.org/10.1109/ACCESS.2022.3213711.

[12] Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Синтез модели машинного обучения для обнаружения компьютерных атак на основе набора данных CICIDS2017. // Труды института системного программирования РАН. 2020. Т. 32, вып. 5. С. 81–94. DOI: https://doi.org/10.15514/ISPRAS-2020-32(5)-6.

[13] Горюнов М.Н., Рыболовлев А.А., Рыболовлев Д.А. Оценка применимости методов машинного обучения для обнаружения компьютерных атак // Информационные системы и технологии. 2020. № 6 (122). С. 103–111. URL: https://elibrary.ru/item.asp?id=44141046. EDN: https://elibrary.ru/bhyjls.

DOI: 10.18287/2541-7525-2024-30-4-147-160

Features of using Markov decision-making processes when modeling attacks on artificial intelligence systems

Vetrov*[1] *I.A.*, *Podtopelny*[2] *V.V.

[1] *Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation; vetrov.gosha2009@yandex.ru (I.A.);*
[2] *Kaliningrad State Technical University (KSTU), Kaliningrad, Russian Federation; ionpvv@mail.ru (V.V.);*

Received: *14.10.2024*
Revised: *20.11.2024*
Accepted: *25.11.2024*

Scientific article

Abstract. In this paper, we study the features of modeling attacks on artificial intelligence systems. Markov decision-making processes are used in the construction of the model. A multilevel approach to the interpretation of system states is proposed, which includes several stages of detailing the states. This approach is based on the MITRE ATLAS methodology and the FSTEC Threat Assessment Methodology. When forming the vector, the specifics of the intruder model are taken into account, and two main modeling modes are considered: on-time and off-time. The procedure for the formation of awards at the abstract level (without specifying the actions of the attacker) of building a model is described.

Key words: network attack; vulnerability; Markov process; modeling; strategy; policy; teaching method.

Information about the conflict of interests: the authors declared no conflicts of interest.

Citation. Vetrov I.A., Podtopelny V.V. Features of using Markov decision-making processes when modeling attacks on artificial intelligence systems. *Vestnik Samarskogo universiteta. Estestvennonauchnaya seriya / Vestnik of Samara State University. Natural Science Series*, 2024, vol. 30, no. 4, pp. 147–160. DOI: 10.18287/2541-7525-2024-30-4-147-160. (In Russ.)

Igor A. Vetrov (vetrov.gosha2009@yandex.ru) – Candidate of Technical Sciences, associate professor, methodological supervisor for the UGSNP "Information Security", Educational and Scientific Cluster "Institute of High Technologies", Immanuel Kant Baltic Federal University, 14, Alexander Nevsky Street, Kaliningrad, 236041, Russian Federation.
Vladislav V. Podtopelny (ionpvv@mail.ru) – senior lecturer, Institute of Digital Technologies, Kaliningrad State Technical University, 1, Sovetsky Avenue, Kaliningrad, 236022, Russian Federation.

References

[1] Kotenko I.V., Saenko I.B., Lauta O.S., Vasiliev N.A., Sadovnikov V. Attacks and defense methods in machine learning systems: analysis of modern research. *Voprosy kiberbezopasnosti*, 2024, no. 1 (59), pp. 24–37. DOI: https://doi.org/10.21681/2311-2024-1-24-37. EDN: https://elibrary.ru/izqdls. (In Russ.)

[2] Namiot D.E. Schemes of attacks on machine learning models. *International Journal of Open Information Technologies*, 2023, vol. 11, no. 5, pp. 68–86. Available at: https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya?ysclid==m5dmh9jnct363700583. (In Russ.)

[3] Xiaofan Zhou, Simon Yusuf Enoch, Dan Dong Seong Kim. Markov Decision Process For Automatic Cyber Defense. In: *You I., Youn T.Y. (eds) Information Security Applications. WISA 2022. Lecture Notes in Computer Science, vol. 13720.* Springer, Cham, 2023, pp. 313–329. DOI: https://doi.org/10.1007/978-3-031-25659-2_23.

[4] Booker L.B., Musman S.A. A model-based, decision-theoretic perspective on automated cyber response. DOI: https://doi.org/10.48550/arXiv.2002.08957.

[5] Zheng J., Namin A.S. Defending SDN-based IoT Networks Against DDoS Attacks Using Markov Decision Process. In: *2018 IEEE International Conference on Big Data (Big Data)*. DOI: https://doi.org/10.1109/BigData.2018.8622064.

[6] Shcheglov A.Yu. Protection of computer information from unauthorized access. Saint Petersburg: Nauka i Tekhnika, 2004, 384 p. Available at: https://reallib.org/reader?file=523140. (In Russ.)

[7] Kochenderfer M., Wheeler T., Wray K. Algorithms for Decision Making. Translated from English by V.S. Yatsenkova. Moscow: DMK Press, 2023, 684 p. Available at: https://znanium.ru/catalog/document?id=445338. (In Russ.)

[8] Methodological document "Methodology for assessing information security threats" (approved by the Federal Service for Technical and Export Control on February 5, 2021). Moscow, 2021, 83 p. Available at: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-5-fevralya-2021-g. (In Russ.)

[9] MITRE ATLAS. Retrieved from the official website of MITRE ATT&CK. Available at: https://atlas.mitre.org, free (accessed 02.05.2024)

[10] Gorbachev I.E., Gluhov A.P. Modeling of processes of information security violations of critical infrastructure. *SPIIRAS Proceedings*, 2015, no. 1 (38), pp. 112–135. Available at: https://www.elibrary.ru/item.asp?id=23342077. EDN: https://elibrary.ru/tquqzx. (In Russ.)

[11] Faranak Abri, Jianjun Zheng, Akbar Siami Namin, Keith S. Jones. Markov Decision Process for Modeling Social Engineering Attacks and Finding Optimal Attack Strategies. *IEEE Access*, 2022, vol. 10, pp. 109949–109968. DOI: http://doi.org/10.1109/ACCESS.2022.3213711.

[12] Goryunov M.N., Matskevich A.G., Rybolovlev D.V. Synthesis of a Machine Learning Model for Detecting Computer Attacks Based on the CICIDS2017 Dataset. *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)*, 2020, vol. 32, no. 5, pp. 81–94. DOI: https://doi.org/10.15514/ISPRAS-2020-32(5)-6. (In Russ.)

[13] Goryunov M.N., Ry'bolovlev A.A., Ry'bolovlev D.A. Evaluating the applicability of machine learning methods to detect computer attacks. *Information Systems and Technologies*, 2020, no. 6 (122), pp. 103–111. Available at: https://elibrary.ru/item.asp?id=44141046. EDN: https://elibrary.ru/bhyjls. (In Russ.)