АНАЛИЗ УГРОЗ ЗЛОУМЫШЛЕННОЙ МОДИФИКАЦИИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СИСТЕМ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

Костогрызов А.И.[1], Нистратов А.А.[2]

Цель: *предложить методический аппарат для вероятностного анализа корректности обучаемых программных средств (ПС) в системах с искусственным интеллектом (СИИ) при их разработке и эксплуатации в условиях потенциальных угроз злоумышленной модификации модели машинного обучения (ММО).*

Методы исследования включают: *методы теории вероятностей, методы системного анализа. Подход основан на адаптации разработанных ранее авторских вероятностных моделей, которые доведены до уровня реализации в ГОСТ Р 59341-2021 «Системная инженерия. Защита информации в процессе управления информацией системы».*

Результат: *в условиях принятых предположений и допущений разработаны вероятностные модели для оценки частных рисков невыявления некорректностей в машинном обучении (дообучении) при разработке и эксплуатации ПС, а также метод оценки интегрального риска нарушения корректности машинного обучения в течение задаваемого периода прогноза. Проанализированы актуальные угрозы подмены ММО и модификации ММО путем искажения («отравления») обучающих данных. Разработаны предложения по формированию исходных данных для прогнозирования рисков с использованием предложенных моделей. Подход проиллюстрирован расчетными примерами с количественными оценками, зависимостями рисков от исходных данных и обоснованием рекомендаций.*

Научная новизна: *впервые для условий потенциальных угроз злоумышленной модификации ММО предложены модели и методы количественной оценки частных рисков невыявления некорректностей в машинном обучении при разработке и эксплуатации ПС и интегрального риска нарушения корректности машинного обучения для СИИ в течение задаваемого периода прогноза.*

Ключевые слова: *вероятность, искажение обучающих данных, модель, риск, система, угрозы.*

DOI:10.21681/2311-3456-2023-5-9-24

1. Введение

Системы с искусственным интеллектом (СИИ) все глубже проникают в повседневную жизнь человека. И это далеко не только голосовые помощники в наших персональных телефонах, навигаторы, онлайн карты и иные удобные сервисы. СИИ все чаще используется в системах обеспечения безопасности на основе интеллектуальной обработки огромных потоков разнородной информации, поступающей от различных камер, сенсоров, устройств телеметрии. Программные средства (ПС) СИИ, обновляемые с помощью моделей машинного обучения, помогают соответствующим службам в распознавании лиц и документов, строений и сооружений и их местоположений, в идентификации предпосылок к нарушению информационной, промышленной, транспортной, экологической безопасности, в геологоразведке, медицине, фармацевтике и биологии, в мониторинге соблюдения правил дорожного движения, распознавая условия нарушения и государственные номера транспортных средств нарушителей и др. Эти примеры далеко не исчерпывают практических возможностей СИИ – см., например, [1-3].

В основе эффектов от применения СИИ лежат обучаемые нейронные сети. Искусственные нейронные сети основаны на наборе персептронов, называемых нейронами. Каждый нейрон сопоставляет набор входных данных с выходными, используя функцию активации. Машинное обучение управляет весами

1 Костогрызов Андрей Иванович, заслуженный деятель науки РФ, доктор технических наук, профессор, главный научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Москва, Россия. E-mail: Akostogr@gmail.com

2 Нистратов Андрей Андреевич, кандидат технических наук, старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Москва, Россия. E mail: andrey.nistratov.job@yandex.ru

и функцией активации таким образом, чтобы иметь возможность правильно определять выходные данные. В то время, как однослойная нейронная сеть (или персептрон) - это подход к разработке объектов, глубокая нейронная сеть позволяет изучать объекты, используя необработанные данные в качестве входных данных. За счет этого достигается существенное увеличение производительности СИИ по сравнению с обычным человеческим интеллектом при решении многих практических задач. При этом обеспечение безопасности информации СИИ должно предусматривать возможность противодействия злоумышленным угрозам подмены и модификации ММО. Однако сегодня системная зависимость нарушения нормального функционирования СИИ от этих угроз является не только далеко не прозрачной, но и на количественном уровне не анализируется. Не представляя всей «внутренней кухни» машинного обучения, заказчик и пользователи системы могут вполне воспринимать нарушения ее нормального функционирования за обычное техническое несовершенство, не устанавливая прямой связи со злоумышленными действиями «умного» нарушителя по модификации ММО. Опасность в том, что нарушитель пытается целенаправленно подменить ММО или исказить обучающие данные, вводя тщательно разработанные ложные образцы так, чтобы в конечном итоге скомпрометировать весь процесс машинного обучения.

В рамках настоящей работы для систем, использующих СИИ, из множества различных угроз выделены следующие актуальные угрозы[3]: угроза подмены ММО (УБИ.222) и угроза модификации ММО путем искажения («отравления») обучающих данных (УБИ.221). Это обусловлено следующими соображениями. В наше время нередко разработчики ПС, осуществляющие машинное обучение (дообучение), принадлежат сторонним организациям относительно разработчика систем, использующих СИИ. Они являются основными владельцами ММО, не хотят раскрывать и передавать заказчику и головному разработчику системы исходные тексты, находятся на субконтракте, сами разрабатывают ПС, в которых содержатся результаты машинного обучения, и контролируют его корректность. Обученные и дообученные ПС передаются заказчику и головному разработчику систем, использующих СИИ, для функционального тестирования, после чего оттестированные ПС принимаются в эксплуатацию в системе. Сертификация дообучаемых ПС по требованиям безопасности может оказаться нецелесообразной из-за длительности и дороговизны ее проведения для заказчика, а также из-за возможного нежелания владельцев ММО раскрывать все исходные тексты программ и методы обучения. В этом случае угрозы, связанные со злоумышленной модификацией ММО, становятся остро актуальными и требуют системного анализа.

Применение предлагаемого подхода к решению различных прямых и обратных задач для обеспечения эффективного целевого применения СИИ позволит прогнозировать риски и количественно обосновывать принимаемые решения о стратегии и мерах противодействия рассматриваемым угрозам. При этом под риском понимается сочетание вероятности нанесения ущерба и тяжести этого ущерба (по ГОСТ Р 51898). В работе основное внимание сосредоточено на анализе вероятностного выражения риска, полагается, что возможный ущерб (чаще - репутационный) противопоставляется расчетным значениям рисков и соответствующим условиям моделирования.

Подход учитывает последние взгляды Национального института стандартизации США на таксономию внедрения в СИИ вредоносного машинного обучения, а также основы управления рисками для СИИ[4],[5] и не противоречит им.

2. Характеристика возможных угроз и сценариев их реализации

Краткая характеристика угроз УБИ.222 и УБИ.221, а также возможных злоумышленных действий нарушителей, именуемых атаками (Attacks), приведена со ссылками на обобщенные взгляды в России и международном сообществе, анализирующем риски в СИИ[6] – см. также [1-4] и сноски 3–5.

Угроза УБИ.222 заключается в возможности подмены ММО внутренним нарушителем (с высоким потенциалом). Угроза обусловлена слабостями разграничения доступа в СИИ, реализация угрозы возможна при наличии у нарушителя непосредственного доступа к ММО.

Угроза УБИ.221 заключается в возможности модификации ММО внешним нарушителем (с высоким по-

3 см. сайт ФСТЭК России https://bdu.fstec.ru/ - Банк данных угроз безопасности информации. ФАУ «ГНИИИ ПТЗИ ФСТЭК России». Дата обращения 25.07.2023

4 Adversarial Machine Learning. A Taxonomy and Terminology of Attaks and Mitigations (Вредоносное машинное обучение. Таксономия и терминология атак, и способов снижения их отрицательных последствий). NIST AI 100-2e2023 ipd, 2023. nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

5 Artificial Intelligence Risk Management Framework. NIST AI 100-1, 2023. nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

6 Biggio B., Fumera G. and Roli F. Security evaluation of pattern classifiers under attack. IEEE transactions on knowledge and data engineering 26, 4. 2014. 984-996.

тенциалом) или внутренним нарушителем (со средним или высоким потенциалом) путем искажения («отравления») обучающих данных. Угроза обусловлена недостатками алгоритмов машинного обучения и осуществления процесса машинного обучения. Реализация угрозы возможна при наличии у нарушителя возможности воздействовать на процесс машинного обучения. Атаки с искажением («отравлением») по сути представляют собой целенаправленное злоумышленное изменение обучающих данных во время машинного обучения для компрометации всего процесса машинного обучения («отравление» - это буквальный перевод на русский язык англоязычного термина Poisoning Attack).

Анализ возможностей нарушителя на этапе обучения состоят в следующем. Нарушитель пытается напрямую повлиять на ММО или повредить ее, изменяя набор данных, используемый для обучения. Самая распространенная атака - это простой доступ к частичным или полным данным обучения.

На сегодня выделяются три применимые стратегии атаки для модификации ММО, основанные на возможностях нарушителя – это стратегии ввода данных, модификации данных и искажения логики ММО.

Стратегия ввода данных используется, когда нарушитель не имеет никакого доступа к обучающим данным, а также к алгоритму обучения, но имеет возможность добавить новые данные в обучающий набор. Он может исказить целевую ММО, вставив ложные выборки в обучающий набор данных. Это влечет за собой некорректность машинного обучения при разработке соответствующих ПС.

Стратегия модификации данных используется, когда нарушитель не имеет доступа к алгоритму обучения, но имеет полный доступ к обучающим данным. Нарушитель напрямую искажает обучающие данные (например, путем прямого изменения меток обучающих данных), изменяя их до того, как они будут использованы целевой ММО. Это также влечет за собой некорректность машинного обучения при разработке соответствующих ПС.

Стратегия искажения логики ММО используется, когда нарушитель имеет возможность вмешиваться в алгоритм обучения (например, путем манипулирования входными характеристиками в зависимости от своих возможностей). Это наиболее опасные атаки, поскольку очень трудно разработать стратегию упреждающего противодействия злоумышленным действиям нарушителя, способного законным образом изменить логику обучения (подобного рода нарушения легко могут быть замаскированы под неумышленную «ошибку»). Искажение логики целенаправленно влечет за собой некорректность машинного обучения при разработке соответствующих ПС, поскольку нарушителем контролируется и модифицируется сама целевая ММО.

Злоумышленные возможности нарушителя на этапе тестирования ПС состоят в следующем. Нарушитель пытается напрямую повлиять на ММО или повредить ее, изменяя набор данных. Атаки во время тестирования не влияют на целевую ММО, но приводят к неверным выходным результатам при использовании соответствующих ПС. Эффективность таких атак определяется главным образом объемом доступной нарушителю информации о целевой ММО.

Таким образом, реализация угроз злоумышленных действий по модификации ММО для СИИ рассчитана на «умного» нарушителя, понимающего свои возможности, представляющего и способного поставить достижимые задачи нарушения целостности системы. Вышеизложенные пояснения даны для понимания излагаемой далее формализации и системного анализа рассматриваемых угроз, мер противодействия этим угрозам и соответствующих рисков от реализации этих угроз.

3. Принятые предположения и допущения

Предполагается, что анализ рассматриваемых угроз может быть формализован с использованием понятия моделируемой системы. Получаемые результаты вероятностного моделирования используются в приложении к исходной СИИ, в интересах которой проводятся соответствующие исследования.

Под моделируемой системой понимается система, для которой решение задач системного анализа осуществляется с использованием ее формализованной модели и, при необходимости, формализованных моделей учитываемых сущностей в условиях их применения. В свою очередь под целостностью моделируемой системы понимается такое ее состояние, которое отвечает целевому назначению модели системы в течение задаваемого периода прогноза.

Примечание. В качестве модели системы могут выступать формализованные сущности, объединенные целевым назначением. Например, при проведении системного анализа в принимаемых допущениях, ограничениях и предположениях модель может формально описывать процесс, функциональные действия, множество активов или множество этих или иных сущностей в их целенаправленном применении в задаваемых условиях (по ГОСТ Р 59341).

С учетом неопределенностей расчет вероятностных показателей делается при условии или в предпо-

ложении реальной или гипотетической повторяемости возможных событий и их независимости. Для математической формализации приняты следующие предположения:

- в общем случае защищенность моделируемой системы от рассматриваемых угроз зависит от корректности машинного обучения при разработке и эксплуатации ПС;
- к началу периода прогноза целостность моделируемой системы полагается обеспеченной, в условиях неопределенностей возникновение и разрастание различных угроз целостности моделируемой системы описывается в терминах случайных событий;
- для различных вариантов развития угроз существуют технологии и меры для выявления признаков возникновения источников угроз и воспрепятствования реализации угрозам, а также следов реализации угроз.

Кроме того, делается предположение о наличии возможностей по определению предпосылок к реализации угроз, а также возможностей по приемлемому восстановлению нарушаемых условий функционирования моделируемой системы (с точки зрения противодействия угрозам). Обоснованное использование выбранных мер противодействия угрозам является предупреждающими контрмерами.

С учетом различных неопределенностей относительно возможных угроз принято допущение о пуассоновских потоках моментов возникновения событий на временной оси и об экспоненциальном распределении времени развития угроз. Предположение о пуассоновости обосновано тем, что в период прогноза общий поток моментов возникновения событий гипотетически представляет собой сумму большого числа составных разнородных потоков. Интенсивность каждого из слагаемых потоков мала по сравнению с интенсивностью суммарного потока – в такой ситуации действует предельная теорема Хинчина – Григолиониса, согласно которой суммарный поток будет близок к пуассоновскому. В свою очередь, экспоненциальное распределение обладает свойством отсутствия последействия. Это означает, что согласно предположению об экспоненциальности остаток времени до реализации угрозы всегда имеет то же распределение с тем же параметром, что и время с момента возникновения угрозы. Это предполагает более тяжелые условия функционирования моделируемой системы.

Принятые предположения и допущения позволяют предложить следующие вероятностные модели для анализа рассматриваемых угроз с использованием показателей рисков.

4. Модель для оценки риска невыявления некорректностей в машинном обучении при разработке ПС

Модель позволяет оценить возможность реализации рассматриваемых угроз при разработке ПС (в частности – при его тестировании) по показателям вероятности получения корректных результатов машинного обучения $P_{корр(1)}$ и риска невыявления некорректностей в машинном обучении. Модель адаптирует разработанные ранее авторские вероятностные подходы, которые доведены до уровня реализации в ГОСТ Р 59341, приложении В.3.7, а также учитывает иные научно-практические взгляды [5-14].

Определение: считается, что при разработке ПС машинное обучение (дообучение) проведено корректно в моделируемой системе, если в процессе контроля обученных ПС до истечения заданного срока его контроля все некорректности выявлены и новые алгоритмические ошибки не допущены. Некорректности при разработке ПС (в параметрах, исходных текстах программ, алгоритмах, обучающих фотографиях, метках и опорных векторах, действиях и др., способных привести к нарушениям нормального функционирования ПС при эксплуатации СИИ) – это в общем случае то, что искажает ожидаемые результаты последующего применения ПС после их машинного обучения в условиях рассматриваемых угроз по сравнению со случаем отсутствия каких-либо угроз. Некорректности появляются в результате реализации угроз, описанных выше в разделе 2, и характеризуют отсутствие корректности машинного обучения в моделируемой системе. Требуемая корректность машинного обучения при разработке ПС в идеале заключается в недопущении злоумышленной модификации адекватной ММО и использования небезопасных версий ПС, а также в исключении искажения обучающих данных. В общем случае под корректностью машинного обучения при разработке ПС для СИИ понимается свойство ПС, получаемых в результате машинного обучения, обеспечивать получение правильных согласованных результатов или эффектов обработки информации в соответствии с целевым назначением этой обработки в моделируемой системе. Корректность обеспечивается на основе применения адекватных способов машинного обучения и контроля результатов обучения, позволяющих выявить все имеющие место некорректности и не допустить алгоритмических ошибок при контроле

обученных ПС. Корректность машинного обучения после контроля информации по обучаемым ПС является следствием приемлемого соотношения между объемом контролируемой информации, частью важной для принятия решения информации, подлежащей учету, скоростью контроля информации, частотой ошибок контролера, длительностью его непрерывной работы и ограничениями на допустимое время контроля. В качестве контролера могут выступать человек – разработчик ПС, учитель, тестировщик или аналитик (в т.ч. лицо, принимающее решение), программно-технические инструментальные средства, ориентированные на выявление некорректностей в машинном обучении при разработке ПС, или их комбинация.

Для моделирования процесса контроля информации в моделируемой системе при разработке ПС приняты следующие обозначения:

V – объем информации по обучаемым (при тестировании – по обученным) ПС, подлежащий контролю (объем измеряется в безразмерных условных единицах – у.е., это могут, например, быть количество параметров, строк текста, алгоритмов, обучающих фотографий, меток и опорных векторов, действий, количество нарушений нормального функционирования ПС при тестировании и др.);

μ – часть важной для принятия решения информации, которая должна быть объективно использована при контроле информации в заданном объеме , измеряемая от 0 до 100% от анализируемого объема информации;

ν – скорость контроля (у.е. в единицу времени);

n – частота ошибок контроля 1-го рода (когда несущественная для принятия решения информация ошибочно воспринимается в качестве важной, влияющей на корректность машинного обучения);

$T_{нар}$ – среднее время наработки на алгоритмическую ошибку (когда объективно важная для принятия решения информация игнорируется, это – аналог ошибки контроля 2-го рода);

$T_{непр}$ – период непрерывной работы контролера;

$T_{зад}$ – задаваемое время на контроль информации.

Возможны 4 варианта соотношений между временем реального контроля всего контролируемого объема , задаваемым допустимым временем на контроль и непрерывным временем работы контролера .

Вариант 1. Задаваемое время на контроль информации не меньше, чем время реального контроля (т. е. $T_{реальн} \le T_{зад}$), а объем контролируемой информации относительно мал, что позволяет проверить его за один период непрерывной работы контролера ($T_{реальн} \le T_{непр}$).

Для экспоненциальной аппроксимации распределений интервалов между ошибками в контролируемой информации, времени до свершения ошибки 1-го рода и времени наработки контролера на ошибку, а также при условии независимости исходных характеристик вероятность $P_{после\ (1)}$ (V, μ, ν, n, $T_{нар}$, $T_{непр}$, $T_{зад}$) отсутствия некорректностей в машинном обучении после контроля для варианта 1 определяется выражением:

$$P_{после\ (1)} = \begin{cases} e^{-nV/\nu}\left[T_{нар}^{-1}\, e^{-\mu V} - \mu\nu e^{-V/(\nu\, T_{нар})}\right]/ \\ /\left(T_{нар}^{-1} - \mu\nu\right), \text{если } T_{нар}^{-1} \neq \mu\nu, \\ e^{-(n+\mu\nu)V/\nu}[1 - V\mu], \text{если } T_{нар}^{-1} = \mu\nu. \end{cases} \quad (1)$$

Вариант 2. Задаваемое время на контроль информации не меньше, чем время реального контроля (т. е. $T_{реальн} \le T_{зад}$), но объем контролируемой информации относительно большой ($T_{реальн} \le T_{непр}$). Это требует нескольких (N) периодов непрерывной работы контролера, в общем случае $N=V/(\nu\, T_{непр})$. Внутри каждого периода проверяют часть всего объема, равную в среднем $V_{части(2)}=V/N$, а допустимое время контроля информации для этой части принимается равным $T_{задчасти\ (2)} = T_{зад}/N$. Тем самым для каждой контролируемой части выполняются условия варианта 1. Вероятность $P_{после\ (2)} = P_{после(2)}$ (V, μ, ν, n, $T_{нар}$, $T_{непр}$, $T_{зад}$) отсутствия некорректностей в машинном обучении после контроля для варианта 2 определяется выражением:

$$P_{после\ (2)} = \\ = \{P_{после\ (1)}(V_{части\ (2)}, \mu, \nu, n, T_{нар}, T_{непр}, \\ T_{зад\ части\ (2)})\}^N. \quad (2)$$

Вариант 3. Задаваемое время на контроль информации меньше, чем время реального контроля ($T_{реальн} > T_{зад}$) при задаваемой средней скорости контроля ν, т. е. объективно может быть проконтролирована лишь часть от всего объема информации при контроле, эта часть равна $V_{части\ (3)} = \nu\, T_{зад}$. В свою очередь, сам объем контролируемой информации относительно мал и может быть проверен за один период непрерывной работы контролера, т. е. $T_{реальн} \le T_{непр}$ и для проверяемого объема $V_{части(3)}$ выполняются условия варианта 1. Вероятность $P_{после\ (3)} = P_{после(3)}$ (V, μ, ν, n, $T_{нар}$, $T_{непр}$, $T_{зад}$) отсутствия некорректностей в машинном обучении после его контроля для варианта 3 определяется выражением:

$$P_{после\ (3)} = [V_{части\ (3)}/V] \\ \cdot P_{после\ (1)}(V_{части\ (3)}, \mu, \nu, n, T_{нар}, T_{непр}, T_{зад}) + \\ +[(V - V_{части\ (3)})/V] \cdot P_{без\ контроля}, \quad (3)$$

где вероятность отсутствия некорректностей в непроверенной части информации, равной $V–V_{\text{части (3)}}$, составляет $P_{\text{без контроля}} = e^{-\mu(V-V_{\text{части (3)}})}$, а вероятность отсутствия некорректностей в объеме проверенной информации равна $P_{\text{после (1)}}$. ($V_{\text{части (3)}}$, μ, ν, n, $T_{\text{нар}}$, $T_{\text{непр}}$, $T_{\text{зад}}$).

Вариант 4. Задаваемое время на контроль информации меньше, чем время реального контроля ($T_{\text{реальн}} > T_{\text{зад}}$), а объем контролируемой информации относительно большой ($T_{\text{реальн}} > T_{\text{зад}}$). Аналогично варианту 3 реально может быть проконтролирована лишь часть от всего объема, равная $V_{\text{части (4)}} = \nu T_{\text{зад}}$. Относительно этой части возможны два подварианта:

- подвариант 4.1: $T_{\text{зад}} \le T_{\text{непр}}$, т. е. проверка будет завершена за один период непрерывной работы контролера;
- подвариант 4.2: $T_{\text{зад}} > T_{\text{непр}}$, т. е. потребуется несколько (N) периодов непрерывной работы контролера, $N = V_{\text{части (4)}} / (\nu T_{\text{непр}})$.

Для подварианта 4.1 вероятность отсутствия некорректностей в машинном обучении после контроля $P_{\text{после(4.1)}} = P_{\text{после(4.1)}}$ (V, μ, ν, n, $T_{\text{нар}}$, $T_{\text{непр}}$, $T_{\text{зад}}$) определяется выражением:

$$P_{\text{после (4.1)}} = [V_{\text{части (4)}}/V] \cdot \cdot P_{\text{после (1)}}\left(V_{\text{части (4)}}, \mu, \nu, n, T_{\text{нар}}, T_{\text{непр}}, T_{\text{зад}}\right) + +[V - V_{\text{части (4)}})/V] \cdot e^{-\mu\left(V-V_{\text{части (4)}}\right)}. \quad (4)$$

Для подварианта 4.2 внутри каждого периода проверяют новую часть, равную в среднем $V_{\text{части (4.2)}} = V_{\text{части (4)}}/N$, и допустимое время контроля для этой новой части принимают равным $T_{\text{зад части (4.2)}} = T_{\text{зад}}/N$.

Вероятность $P_{\text{после(4.2)}} = P_{\text{после(4.2)}}$ (V, μ, ν, n, $T_{\text{нар}}$, $T_{\text{непр}}$, $T_{\text{зад}}$) отсутствия некорректностей в машинном обучении после его контроля определяется выражением:

$$P_{\text{после (4.2)}} = [V_{\text{части (4)}}/V] \cdot \cdot \left\{P_{\text{после (1)}}\left(V_{\text{части (4.2)}}, \mu, \nu, n, T_{\text{нар}}, T_{\text{непр}}, T_{\text{зад части (4.2)}}\right)\right\}^N + +[V - V_{\text{части (4)}})/V] \cdot e^{-\mu\left(V-V_{\text{части (4)}}\right)}. \quad (5)$$

В итоге вероятность отсутствия некорректностей в машинном обучении после контроля $P_{\text{корр(1)}} = P_{\text{после}}$ определяется аналитическими выражениями для $P_{\text{после(1)}}$, $P_{\text{после(2)}}$, $P_{\text{после(3)}}$, $P_{\text{после(4.1)}}$, $P_{\text{после(4.2)}}$ в зависимости от варианта соотношений между исходными данными.

Для формирования исходных данных при моделировании могут использоваться статистические данные, включая данные для систем-аналогов, а также обоснованные гипотетические данные.

Для системного анализа результатов моделирования в оценках интегрального риска (см. раздел 6 статьи) рекомендуется задание допустимого уровня Рдоп корр(1) и условия α. Условие α касается не только обеспечения корректности машинного обучения при разработке ПС, но и возможного ущерба при реализации угроз. Условие α формулируется в виде ограничений: $P_{\text{корр(1)}} \ge P_{\text{доп корр(1)}}$ и возможный ущерб от нарушения не превышает допустимого (это - формулировка условия α). Учет результатов моделирования в оценках интегрального риска осуществляется с использованием индикаторного коэффициента $Z_{\text{корр(1)}}$ корректности машинного обучения при разработке ПС:

$$Z_{\text{корр(1)}} = \begin{cases} 1, \text{ если условие корректности машинного обучения при разработке ПС } \alpha \text{ выполнено,} \\ P_{\text{корр(1)}}, \text{ если условие } \alpha \\ \text{не выполнено или не задано.} \end{cases}$$

Сопоставление с возможным ущербом (или недополученным эффектом) позволяет рассматривать дополнение до единицы этого коэффициента (1 – Zкорр) в качестве вероятностного выражения риска невыявления некорректностей в машинном обучении при разработке ПС.

5. Модель для оценки риска невыявления некорректностей в машинном обучении при эксплуатации ПС

Модель позволяет оценить возможность реализации рассматриваемых угроз УБИ.222 или УБИ.221 при эксплуатации ПС по показателю вероятности получения корректных результатов машинного обучения $P_{\text{корр(2)}}$ и риска невыявления некорректностей в машинном обучении при эксплуатации ПС.

Определение: считается, что машинное обучение (дообучение) характеризуется корректностью при эксплуатации ПС в течение заданного периода прогноза, если в течение этого периода не были реализованы угрозы, связанные с использованием потенциально небезопасных версий ПС, при разработке которых могли быть использованы искаженные («отравленные») нарушителем обучающие данные или осуществлена подмена или модификация ММО. Некорректности при эксплуатации ПС – это в общем случае возникновение на временной оси негативных событий, вызванных допущенными и пропущенными ошибками при разработке ПС, уязвимостями в ПС, способствующих нарушению нормального функционирования СИИ, согласно ее назначению. Некорректности появляются в результате реализации угроз, описанных выше в разделе 2, и характеризуют отсутствие

корректности. Требуемая корректность машинного обучения при эксплуатации ПС достигается противодействием угрозам по факту выявления предпосылок или выявления непосредственного ущерба (недополученного эффекта) от реализации угроз при функционировании моделируемой системы. Корректность при эксплуатации ПС обеспечивается на основе анализа обращений пользователей на нарушения нормального функционирования СИИ с потенциально небезопасной версией ПС и/или на оперативное восстановление приемлемых условий ее функционирования (см. раздел 3). В качестве аналитика могут выступать оператор и пользователи системы, использующей СИИ, разработчик, осуществляющий сопровождение ПС, программно-аналитические инструментальные средства, ориентированные на выявление некорректностей в машинном обучении при эксплуатации ПС, или их комбинация.

Примечание. Нарушение нормального функционирования моделируемой системы должно быть определено формально. Возможно использование экспертных границ с применением универсальной вспомогательной модели показателя (УВМП) – см. раздел 8.

В моделях для оценки риска невыявления некорректностей в машинном обучении (дообучении) при эксплуатации ПС под моделируемой системой понимается множество функциональных действий модели СИИ, выполняемых с использованием потенциально небезопасных версий ПС, получаемых от разработчиков по результатам машинного обучения или дообучения.

Для моделируемой системы возможно либо отсутствие какого-либо контроля, либо периодический системный контроль хода выполнения функциональных действий. Предлагаемые вероятностные модели и методы адаптируют авторские вероятностные подходы, которые доведены до уровня реализации в ГОСТ Р 59341 (из-за изложения модели в этом стандарте, а также в [5,6, 9-14], она не приводится в полном объеме).

Моделируемая система представлена в виде «черного ящика». Специфика состоит в логическом переопределении исходных данных для моделирования. С формальной точки зрения результатом применения модели с учетом возможного ущерба (недополученного эффекта) является расчетный риск невыявления некорректностей в машинном обучении (дообучении) при эксплуатации ПС в моделируемой системе в течение заданного периода прогноза при реализации периодического системного контроля. Для расчета риска в моделируемой системе сложной структуры для каждого элемента используются исходные данные:

σ – частота возникновения источников угроз возникновения небезопасных версий ПС, при разработке которых были использованы искаженные («отравленные») нарушителем обучающие данные или была осуществлена подмена или модификация ММО;

β – среднее время развития угроз с момента их возникновения до нарушения нормального функционирования моделируемой системы;

$T_{меж}$ – среднее время между окончанием предыдущей и началом очередной диагностики целостности моделируемой системы;

$T_{диаг}$ – среднее время системной диагностики целостности моделируемой системы;

$T_{восст}$ – среднее время восстановления нарушаемой целостности моделируемой системы;

$T_{зад}$ – задаваемая длительность периода прогноза.

В итоге расчетная вероятность корректного машинного обучения ПС характеризуется вероятностью отсутствия нарушений целостности моделируемой системы в течение периода прогноза $T_{зад}$ и определяется теми же аналитическими выражениями (В.1) – (В.9), что и в моделях В.2.2, В.2.3, В.2.4 из ГОСТ Р 59341, в зависимости от варианта соотношений между исходными данными.

Сопоставление с возможным ущербом (или недополученным эффектом) позволяет рассматривать расчетную вероятность по формуле (В.1) как риск невыявления некорректностей в машинном обучении при эксплуатации ПС $P_{корр(2)}$ в моделируемой системе при реализации предпринимаемых технологических мер периодического системного контроля и восстановления целостности моделируемой системы. Вероятностное значение этого риска представляет собой дополнение до единицы вероятности корректного машинного обучения ПС в течение заданного периода прогноза.

В частном случае, когда период между диагностиками больше периода прогноза $T_{зад} < T_{меж}$, модель применима для прогноза риска при отсутствии какого-либо контроля.

Для системного анализа результатов моделирования в оценках интегрального риска (см. раздел 6) рекомендуется задание допустимого уровня $P_{доп\ корр(2)}$ и условия α. Условие α касается не только обеспечения корректности машинного обучения при эксплуатации ПС, но и возможного ущерба при реализации угроз. Условие α формулируется в виде ограничений: $P_{корр(2)} \geq P_{доп\ корр(2)}$ и возможный ущерб от нарушения не превышает допустимого (это - формулировка условия α). Учет результатов моделирования в оценках интегрального риска осуществляется с использованием

индикаторного коэффициента $Z_{корр(2)}$ ($T_{зад}$) корректности машинного обучения при эксплуатации ПС:

$$Z_{корр(2)}(Тзад) = \begin{cases} 1, \text{ если условие корректности машинного обучения при эксплуатации ПС } \alpha \text{ выполнено,} \\ P_{корр(2)}, \text{ если условие } \alpha \text{ не выполнено или не задано.} \end{cases}$$

Сопоставление с возможным ущербом (или недополученным эффектом) позволяет рассматривать дополнение до единицы этого коэффициента ($1 - Z_{корр(2)}(T_{зад})$) в качестве вероятностного выражения риска невыявления некорректностей в машинном обучении (дообучении) при эксплуатации ПС.

6. Метод оценки интегрального риска

Показатель интегрального риска нарушения корректности машинного обучения в моделируемой СИИ позволяет оценить способность нормального функционирования системы в условиях потенциальных угроз злоумышленной подмены и/или модификации ММО. Интегральный риск используется для сравнения весомости прогнозируемых частных рисков, выявления существенных угроз и поддержки принятия решений для задач системного анализа при разработке и эксплуатации моделируемой системы.

В качестве интегрального предлагается виртуальный показатель $\boldsymbol{R}_{интегр}(\boldsymbol{T}_{зад})$ риска нарушения корректности машинного обучения в условиях рассматриваемых угроз моделируемой СИИ, учитывающий в течение задаваемого периода прогноза $\boldsymbol{T}_{зад}$ риск невыявления некорректностей в машинном обучении (дообучении) при разработке ПС и риск невыявления некорректностей в машинном обучении (дообучении) при эксплуатации ПС. С учетом дополнительных условий α, а также в условиях независимости случайных событий (см. раздел 3) этот показатель может быть рассчитан с использованием моделей разделов 4 и 5:

$$R_{нтегр}(T_{зад}) = 1 - Z_{корр(1)} \cdot Z_{корр(2)}(T_{зад}).$$

Примечание. Для более общего случая модели угроз безопасности информации, учитывающей различные виды угроз, могут быть использованы другие вероятностные модели – см., например, ГОСТ Р 59341, ГОСТ Р 59346, ГОСТ Р 59349, ГОСТ Р 59989, ГОСТ Р 59991 и др.

7. Пример оценки риска при разработке ПС

Уже сегодня количество систем, использующих СИИ в различных сферах человеческой деятельности, измеряется многими тысячами, а с широким внедрением Интернета вещей и развитием «умных» систем в ближайшем будущем это количество возрастет на порядки. Тем не менее проблематика количественных оценок исследуемых рисков в России только начинает разворачиваться, критичных случаев злоумышленных модификаций ММО в СИИ не наблюдалось (мошенничество в финансовой сфере – это в общем случае комплекс более специфичных угроз, требующих специального исследования). Соответственно статистика для формирования исходных данных в интересах анализа угроз злоумышленной модификации ММО для СИИ на сегодня практически отсутствует. Поэтому в примере используются правдоподобные гипотетические исходные данные для ориентировочной оценки возможностей наличия некорректностей в машинном обучении при разработке ПС для СИИ.

Положим, по одному исследуемому объекту (например, связанному с распознаванием лиц или документов, строений или сооружений и их местоположений) объем контролируемой информации измеряется различными артефактами общим количеством 1010 у.е. (например, это могут быть параметры объектов, количество строк текста, алгоритмов, обучающих фотографий, меток и опорных векторов, действий, количество нарушений нормального функционирования ПС при тестировании и др.). Т.е. объем информации, подлежащий контролю, для определенности может быть оценен числом V =1010 у.е.

Примечание. Должно быть дано формальное содержательное наполнение у.е. контролируемого объема артефактов при машинном обучении.

В качестве контролера выступает человек – один или несколько разработчиков ПС, учитель, тестировщик или аналитик (в т.ч. лицо, принимающее решение). При этом контроль, как правило, осуществляется не только и не столько по результату, сколько в ходе работ, связанных с машинным обучением (например, в режиме разделения времени «обучение-контроль»). С точки зрения математического моделирования контролеры совместно со средствами, ориентированные на выявление некорректностей в машинном обучении при разработке ПС, представляют собой единое целое.

Часть важной для принятия решения информации, которая должна быть объективно использована при контроле информации в заданном объеме V, рассматривается на уровне до 100% от анализируемого объема в у.е., для определенности положим μ=50%, полагая, что при исследованиях возможны изменения до

100%. Скорость контроля для человека положим вполне реальные 20 у.е. в час, т.е. υ =20 у.е. в час. Период непрерывной работы контролера полагаем равным 1 часу, после чего следует восстановительный отдых, т.е. $T_{непр}$ =1 час. Предположим, что наработка контролера на ошибку 2-го рода (пропуск некорректности) составляет 1 год, что свойственно для специалистов квалификации выше средней, т.е. $T_{нар}$=365 суток. На практике при разработке ПС частота ошибок контроля 1-го рода на порядок меньше, нежели частота ошибок 2-го рода, поэтому соответственно положим n =0.00027 раз в сутки. Время на контроль информации задается таким образом, чтобы успеть завершить контроль всего заданного объема артефактов при установленной скорости контроля.

Тем самым все необходимые исходные данные для моделирования сформированы.

Результаты расчетов показывают, что вероятность получения корректных результатов машинного обучения $P_{корр(1)}$ = 0.994. Более того, достигается высокая степень устойчивости этих результатов (см. рис. 1-4) — вероятность получения корректных результатов машинного обучения не опускается ниже 0.988 (при ориентации на обоснование для системы-эталона по ГОСТ Р 59341, приложению Д допустимый уровень составляет не менее 0.95).

С привязкой к единой вероятностной шкале измерений в сравнении с допустимым уровнем это служит научно обоснованным доказательством несущественности рассмотренных типов угроз в рамках рассмотренного сценария.

Необходимо отметить, что эти положительные результаты получены в предположении, что частота ошибок контроля 1-го рода на порядок меньше, нежели частота ошибок 2-го рода. Это – для случая отсутствия целенаправленных действий по искажению («отравлению») обучающих данных (УБИ.221) или подмене или модификации ММО (УБИ.222).

Несколько изменим сценарий развития угроз, представив себе внедрение в состав разработчиков ПС и контролеров потенциального нарушителя (осуществляющего машинное обучение и контроль), злоумышленно реализующего угрозы УБИ.221 или УБИ.222. Сохраняя неизменными все предыдущие исходные данные для моделирования, проведем дополнительные исследования, изменив лишь частоту

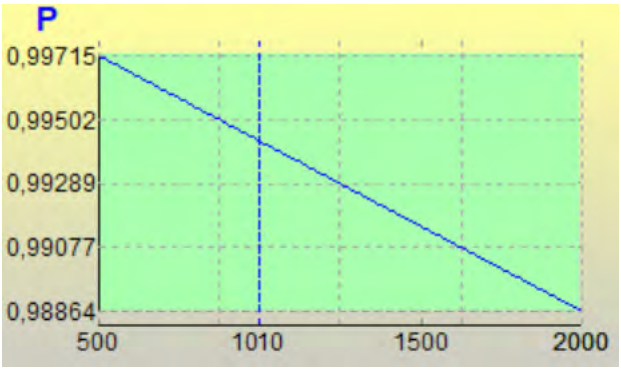

Рис.1. Зависимость вероятности получения корректных результатов машинного обучения от контролируемого объема артефактов (в у.е.)

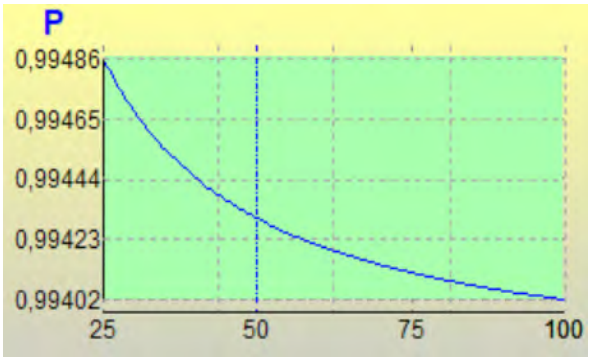

Рис.2. Зависимость вероятности получения корректных результатов машинного обучения от части важной для принятия решения информации (в %)

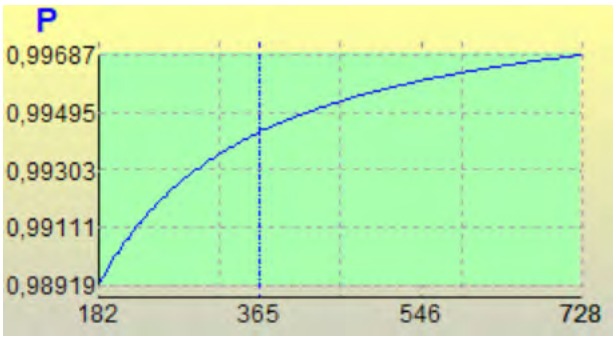

Рис.3. Зависимость вероятности получения корректных результатов машинного обучения от наработки на алгоритмическую ошибку (в сутках)

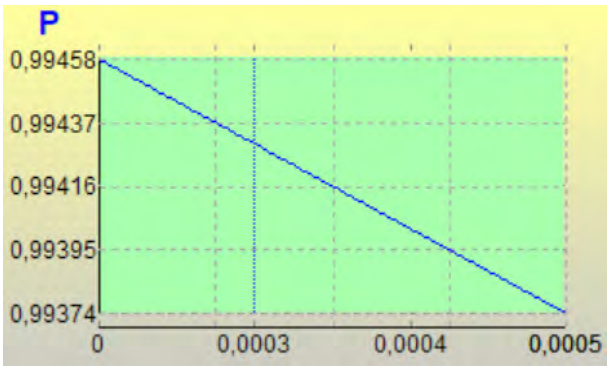

Рис.4. Зависимость вероятности получения корректных результатов машинного обучения от частоты ошибок контроля 1-го рода (раз в сутки)

ошибок контроля 1-го рода (когда несущественная для принятия решения информация ошибочно воспринимается в качестве важной), а именно: сделаем частоту ошибок контроля 1-го рода на порядок больше, нежели частота ошибок 2-го рода, т.е. положим n =0.027 раз в сутки.

Результаты расчетов показывают, что в точке расчета вероятность получения корректных результатов машинного обучения при разработке ПС $P_{корр(1)}$ = 0.939. Это меньше, нежели допустимый уровень 0.95 при ориентации на обоснование для системы-эталона по ГОСТ Р 59341, приложению Д (для вероятности получения корректных результатов обработки информации).

Примечание. При ориентации на прецедентный принцип допустимый уровень для $P_{корр(1)}$ по ГОСТ Р 59341, приложению Д соответствует уровню 0.90.

Более детальные оценки показали следующее. При прочих неизменных условиях контролируемый объем артефактов очень критичен с точки зрения получения корректных результатов машинного обучения – см. рис. 5. Так, при возрастании контролируемого объема до 2000 у.е. вероятность получения корректных результатов машинного обучения падает до 0.88. А допустимый уровень 0.95 будет преодолен, если контролируемый объем артефактов при прочих равных условиях не будет превышать 817 у.е. По этой причине актуальной для снижения риска невыявления некорректностей в машинном обучении при разработке ПС является следующая рекомендация: контролерам качества машинного обучения по возможности следует отбирать для проверки наиболее важные артефакты так, чтобы общее их количество в контролируемом объеме артефактов не превышало 817 у.е. Если этого достичь не удается, следует стараться применять рекомендации, излагаемые далее.

Часть важной для принятия решения информации, которая должна быть объективно использована при контроле информации в заданном объеме артефактов практически не критична – см. рис. 6. Это означает, что в условиях моделирования вся важная информация будет принята контролером во внимание. Скорость контроля и период непрерывной работы контролера практически не критичны. Вместе с тем сравнительно низкое абсолютное значение достигаемой вероятности получения корректных результатов машинного об-

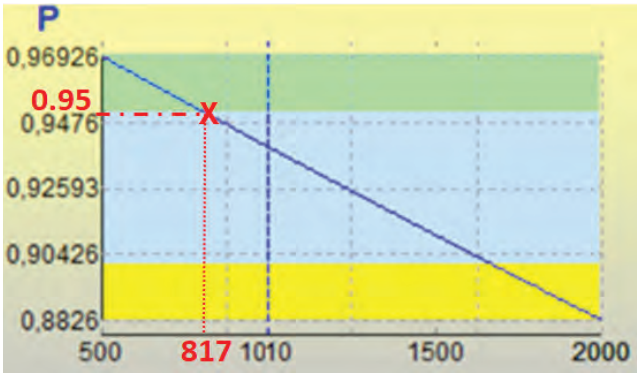

Рис.5. Зависимость вероятности получения корректных результатов машинного обучения от контролируемого объема артефактов (в у.е.)

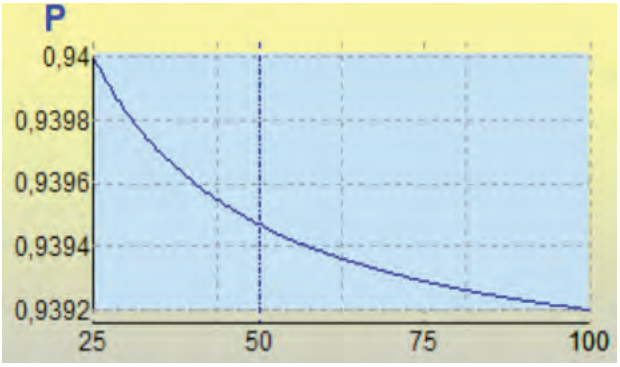

Рис.6. Зависимость вероятности получения корректных результатов машинного обучения от части важной для принятия решения информации (в %)

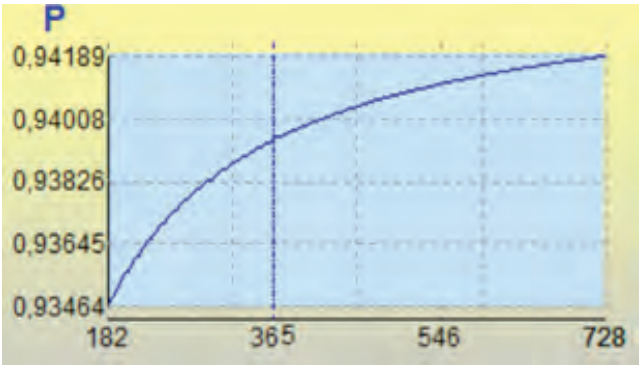

Рис.7. Зависимость вероятности получения корректных результатов машинного обучения от наработки на алгоритмическую ошибку (в сутках)

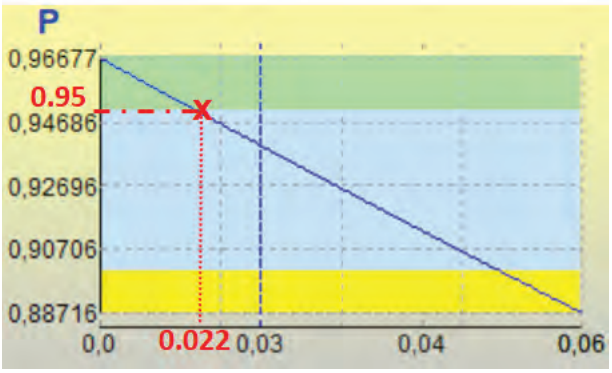

Рис.8. Зависимость вероятности получения корректных результатов машинного обучения от частоты ошибок контроля 1-го рода (раз в сутки)

учения (ниже 0.94) говорит о том, что снижения риска невыявления некорректностей в машинном обучении (дообучении) при разработке ПС следует искать в улучшении значений других параметров.

При прочих неизменных условиях в сравнении с ошибками 2-го рода частота ошибок контроля 1-го рода очень критична для получения корректных результатов машинного обучения – см. рис. 7-8. Так, при возрастании частоты ошибок контроля 1-го рода вдвое с 0.03 до 0.06 раз в сутки вероятность получения корректных результатов машинного обучения монотонно убывает с уровня 0.939 до 0.887. Это подчеркивает актуальность повышения квалификации контролеров машинного обучения. А допустимый уровень 0.95 будет преодолен, если частота ошибок контроля 1-го рода будет не выше 0.022 раз в сутки (что составляет приблизительно 8 раз в год).

Общая рекомендация: целесообразно отслеживать соотношение ошибок контроля 1-го и 2-го рода, не допуская превалирования ошибок 1-го рода (когда несущественная для принятия решения информация ошибочно воспринимается в качестве важной). Заметное превалирование ошибок 1-го рода является явным фактором возрастания риска невыявления некорректностей в машинном обучении при разработке ПС.

8. Предложения по формированию исходных данных для прогнозирования рисков

Выявление некорректностей в машинном обучении при эксплуатации ПС – это очень сложная практическая задача (в народном фольклоре она сродни ситуации, когда на вопрос учителя «Почему у Вас плохие результаты?» следует ответ ученика: «А Вы нас так учили»). В буквальном смысле как обучили ПС, такие прагматические эффекты и будут иметь место с точки зрения применения СИИ по назначению. Формально границы ожидаемых приемлемых эффектов применения СИИ должны быть определены. Например:

- недопустимое время простоя оборудования (использующего СИИ) на объекте с непрерывным производством, влекущее за собой сокращение прибыли или ущербы, должно составлять в среднем не более 0.5 часа за один останов оборудования и не более 4-х раз в месяц (в техническом задании на систему это требование бизнеса преобразуется чисто в техническое требование, к примеру: должна быть обеспечена приемлемая надежность выполнения функций системой в течение года – с вероятностью не ниже 0.995 при среднем времени восстановления после отказа не более 0.5 часа);
- приемлемый эффект применения навигаторов транспортного средства – не менее 99.9% адекватности в навигации на заданной территории;
- приемлемая удовлетворенность клиентов от использования биометрической системы платежей в метрополитене по сравнению с другими средствами платежей - не менее 88%;
- прирост числа пациентов, для которых с применением СИИ установлен верный диагноз на ранней стадии опасного заболевания должен составлять не менее 20% по сравнению с обычным диагностированием;
- число адекватно распознанных номеров транспортных средств нарушителей на автомобильных дорогах должно быть не менее 95%;
- приемлемый уровень экономии энергии в «умном» доме – не менее 25% по сравнению обычными домами, не оснащенными СИИ, и т.п.

Это – системный взгляд с одной стороны (со стороны лиц, ожидающих успешных результатов применения систем, использующих СИИ). При этом даже с использованием СИИ неизбежны случайные ошибки человека.

С другой стороны на практике просматриваются два основных варианта создания и эксплуатации ПС, в которых реализуются результаты машинного обучения:

- вариант 1 (редкий) – разработчики ПС, осуществляющие машинное обучение и дообучение, принадлежат одной и той же головной организации, которая разрабатывает и сопровождает всю систему, использующую СИИ. В этом случае предотвращение внедрения злоумышленников в состав разработчиков ПС, осуществляющих машинное обучение (дообучение), и всесторонний контроль – это прерогатива заказчика и разработчика системы. Угрозы подмены и/или модификации ММО слабоактуальны, риски пренебрежимо малы;
- вариант 2 (распространенный) – разработчики ПС, осуществляющие машинное обучение (дообучение), принадлежат сторонним организациям относительно разработчика системы, использующей СИИ. Взаимоотношения заказчик – разработчик для этого варианта подробно описаны во введении при обосновании актуальности настоящей работы. В этом случае угрозы становятся актуальными, риски могут оказаться недопустимо большими.

В случае варианта 2 становится остро востребованной предложенная модель для оценки риска невыявле-

ния некорректностей в машинном обучении при эксплуатации ПС (см. раздел 5). Однако здесь, если относительно определения средних времен между окончанием предыдущей и началом очередной диагностики целостности моделируемой системы ($T_{меж}$) и непосредственно самой системной диагностики целостности ($T_{диаг}$) трудностей не возникает, то с учетом отсутствия какой-либо статистики у аналитика встает правомерный вопрос – как приблизительно можно определить такие исходные данные, как частота возникновения источников угроз возникновения небезопасных версий ПС, при разработке которых были использованы искаженные («отравленные») нарушителем обучающие данные или была осуществлена подмена или модификация ММО (σ), среднее время развития угроз с момента их возникновения до нарушения нормального функционирования моделируемой системы (β), а также среднее время восстановления нарушаемой целостности моделируемой системы ($T_{восст}$).

Для ответа на этот вопрос предлагается использование универсальной вспомогательной модели показателя (УВМП) по ГОСТ Р 59349 «Системная инженерия. Защита информации в процессе системного анализа».

В любой момент времени у ответственных лиц, принимающих решение, имеет место формальное представление о том, какое состояние эксплуатируемой системы, использующей СИИ, «нормально» и «приемлемо», а какое «неприемлемо» и требует управляющей реакции для улучшения.

Т. е. на любой момент времени по каждому из критичных показателей (или по их совокупности) можно с однозначной уверенностью определить, что его (их) значения находятся в состоянии, которое может быть охарактеризовано как «Приемлемое» или «Приемлемое с отклонением» (когда за счет определенных организационных или обычных технических усилий по улучшению значения критичного показателя можно удерживать систему от перехода этого показателя в зону «Неприемлемого» состояния) или как «Неприемлемое» состояние (когда требуются кардинальные решения по восстановлению условий, которые в существующем виде уже не обеспечивают или в ближайшее время при бездействии не будут гарантировать требуемого уровня эффективности системы) – см. рис. 9. Переход критичного показателя в состояние «Неприемлемое» характеризует подозрение, что в ПС системы были реализованы потенциальные угрозы подмены и/или модификации ММО. Например, в качестве критичных показателей могут быть использованы показатели, перечисленные в начале этого раздела, а их допустимая граница – это вышеуказанные допустимые значения, ухудшение которых с точки зрения прагматического эффекта для системы характеризует зону состояния, именуемого как «Неприемлемое». Выбранные критичные показатели при существенном ухудшении их значений относительно установленных пределов до состояния «Неприемлемое» могут служить показателями возникновения некорректностей в машинном обучении при эксплуатации ПС. А граница «Приемлемое с отклонением» характеризует те некоторые уступки по сравнению с наилучшим достигнутым результатом для критичного показателя, которые могут быть допущены с учетом имеющих место неопределенностей.

При этом становятся определенными недостающие исходные данные σ, β, $T_{восст}$. Эти исходные данные формируются по следующему алгоритму, описанному ниже в привязке к регистрируемым значениям критичного показателя на рис. 9.

Частота возникновения источников угроз возникновения небезопасных версий ПС, при разработке которых были использованы искаженные («отравленные») нарушителем обучающие данные или была осуществлена подмена или модификация ММО, определяется выражением:

$$\sigma = 1/[(\tau_{возн.1} + \tau_{возн.2} + \tau_{возн.3} + \tau_{возн.4})/4].$$

Среднее время развития угроз с момента их возникновения до нарушения нормального функционирования моделируемой системы определяется выражением:

$$\beta = (\tau_{разв.1} + \tau_{разв.2} + \tau_{разв.3} + \tau_{разв.4} + \tau_{разв.5})/5.$$

Среднее время восстановления нарушаемой целостности моделируемой системы определяется выражением:

$$T_{восст} = (\tau_{восст.1} + \tau_{восст.2} + \tau_{восст.3})/3.$$

Здесь $\tau_{возн.i}$ – i-й интервал времени между возникновениями источника угроз; $T_{разв.j}$ – j-й интервал времени развития угроз с момента возникновения источника угроз до нарушения нормальных условий; $T_{восст.m}$ – m-й интервал времени восстановления нарушаемой целостности.

Значения σ, β, $T_{восст}$, получаемые в результате применения предложенного выше алгоритма к статистическим данным контроля на уровне УВМП, являются исходными данными для формального описания моделируемой системы с учетом возможности прогнозирования динамики разнородных событий. Роль в УВМП каждого из учитываемых критичных показателей сводится к определению σ, β, $T_{восст}$ для последующего моделирования.

Рис.9. Элементарные состояния контролируемого показателя УВМП во времени и временные характеристики для прогнозирования рисков

Применительно к конкретной СИИ остается установить границы для зон «Приемлемое», «Приемлемое с отклонением», «Неприемлемое». Тогда УВМП превращается в некий эталон, ухудшение показателя по которому до состояния «Неприемлемое» характеризует подозрение, что в ПС системы были реализованы потенциальные угрозы подмены и/или модификации ММО. Отличительная особенность изначальной версии этого эталона, который должен быть признан и утвержден, в том, что наблюдаемый приемлемый эффект от применения системы должен быть получен в полной уверенности в корректности соответствующих обученных ПС, т.е. в полной уверенности того, что если и были потенциальные угрозы, рассматриваемые в статье, то они не были реализованы в соответствующих ПС, которые применялись при формировании изначального эталона.

Сегодняшний период развития СИИ в России примечателен тем, что потенциальные угрозы могут быть реализованы главным образом из-за случайных ошибок, нежели из злоумышленных намерений нарушителя. Поэтому есть достаточно высокая уверенность в том, что заказчики систем, использующих СИИ, и организации-разработчики соответствующих ПС не только тщательно подходят к отбору кадров, но и многие математические вопросы по машинному обучению в условиях дефицита высококвалифицированных специалистов решаются сообща (создавая тем самым условия взаимоконтроля). В связи с этим предлагается в качестве эталона, ориентированного на УВМП, и начальных границ зон «Приемлемое», «Приемлемое с отклонением», «Неприемлемое» брать те значения, которые характеризуют достижение прагматического эффекта сразу, как только он появляется (по сравнению со случаем функционирования системы без использования СИИ).

С учетом широкомасштабных работ в области искусственного интеллекта эта ситуация может поменяться буквально через несколько лет. И тогда изначальная версия эталона по УВМП послужит отправной пограничной полосой для обоснованных подозрений о наличии или отсутствии реализации угроз подмены и/или модификации ММО. По мере эксплуатации системы и сбора соответствующей статистики этот изначальный эталон может быть усовершенствован.

9. Относительно примера оценки риска при эксплуатации ПС

В качестве примера оценки риска невыявления некорректностей в машинном обучении при эксплуатации ПС читателю рекомендуется работа «Подход к вероятностному прогнозированию защищенности репутации политических деятелей от «фейковых» угроз в публичном информационном пространстве» [15]. Достаточно представить себе, что речь идет об избирательной системе в расширенном ее понимании, где в систему входят средства массовой информации, использующие СИИ, которые целенаправленно реализуют рассмотренные в настоящей статье угрозы (в виде «фейковых» результатов машинного обучения) против репутации политических деятелей. В статье определены количественные границы относительно вероятностей сохранения и дискредитации изначально положительной репутации виртуального политического деятеля в условиях правового законодательства в России в период с конца 90-х по 2023гг. Выявлено, что в условиях отсутствия правовых норм по ограничению длительности рассмотрения исков в защиту репутации политического деятеля в России наблюдается недопустимо низкая степень защищенности изначально положительной репутации от таких «фейков», которые могут быть усилены потенциальными возможностями технологий нейролингвистического программирования и специальных политтехнологий психологического воздействия на электорат. Обоснованы востребованные способы защищенности репутации политических деятелей, включая комплексные меры мониторинга и выявления угроз, развития системы правосудия в защиты репутации политического деятеля с указанием количественных характеристик противодействия «фейковым» угрозам. Иными словами, если «фейки» рассматривать как результат реализации рассматриваемых угроз злоумышленной модификации ММО для СИИ, задействованных в избирательных кампаниях политических деятелей, то материалы статьи [15] могут послужить непосредственным примером оценки риска невыявления некорректностей в машинном обучении при эксплуатации ПС.

Заключение

1. Для систем, использующих СИИ, проведен анализ актуальных угроз подмены ММО (УБИ.222) и модификации ММО путем искажения («отравления») обучающих данных (УБИ.221). В условиях принятых предположений и допущений разработаны вероятностные модели для оценки частных рисков невыявления некорректностей в машинном обучении (дообучении) при разработке и эксплуатации ПС, а также метод оценки интегрального риска нарушения корректности машинного обучения в течение задаваемого периода прогноза.

2. Риск невыявления некорректностей в машинном обучении (дообучении) при разработке ПС предложено оценивать в зависимости следующих исходных данных: объема информации по обучаемым ПС, подлежащего контролю; части важной для принятия решения информации, которая должна быть объективно использована при контроле информации в заданном объеме; скорости контроля; частоты ошибок контроля 1-го рода; среднего времени наработки на алгоритмическую ошибку; периода непрерывной работы контролера; задаваемого времени на контроль информации.

3. Риск невыявления некорректностей в машинном обучении (дообучении) при эксплуатации ПС предложено оценивать в зависимости следующих исходных данных: частоты возникновения источников угроз возникновения небезопасных версий ПС, при разработке которых были использованы искаженные («отравленные») нарушителем обучающие данные или была осуществлена подмена модели машинного обучения; среднего времени развития угроз с момента их возникновения до нарушения нормального функционирования моделируемой системы; среднего времени между окончанием предыдущей и началом очередной диагностики целостности моделируемой системы; среднего времени системной диагностики целостности моделируемой системы; среднего времени восстановления нарушаемой целостности моделируемой системы; задаваемой длительности периода прогноза. В интересах формирования необходимых исходных данных для последующего моделирования предложено использовать универсальную вспомогательную модель показателя по ГОСТ Р 59349, адаптированную для анализа рассматриваемых угроз.

4. Интегральный риск предложено оценивать через виртуальный показатель риска нарушения корректности машинного обучения в условиях рассматриваемых угроз в течение задаваемого периода прогноза в зависимости от рисков невыявления некорректностей в машинном обучении (дообучении) при разработке и эксплуатации ПС, а через них – в зависимости от исходных данных, обеспечивающих расчет соответствующих рисков.

5. Предложенный методический аппарат, позволяет осуществлять вероятностную оценку корректности обучаемых ПС в системах, использующих СИИ, при их

разработке и эксплуатации в условиях потенциальных угроз злоумышленной подмены и/или модификации ММО. Работоспособность подхода проиллюстрирована на примерами.

Литература

1. Эртель В., Введение в искусственный интеллект.-М. «Эксмо», 2019. – 448с.
2. Лекун Ян, Как учится машина (революция в области нейронных сетей и глубокого обучения). – М. Альпина PRO, 2021. – 335с.
3. Арлазаров В. В., Мобильное распознавание и его применение к системе ввода идентификационных документов. – Диссертация на соискание ученой степени доктора технических наук. -М. ФИЦ ИУ РАН, 2023. – 358с.
4. Chakraborty A., Alam M., Dey V., Chattopadhayay A.U., Yay D.M., Adversarial attacks and defences: A survey //arXiv preprint arXiv:1810.00069. – 2018
5. Probabilistic modeling in system engineering. InTechOpen, 2018, 279p. – URL: http://www.intechopen.com/books/probabilistic-modeling-in-system-engineering
6. Климов С. М. Модели анализа и оценки угроз информационно-психологических воздействий с элементами искусственного интеллекта. / Сборник докладов и выступлений научно-деловой программы Международного военно-технического форума «Армия-2018». 2018. С. 273–277.
7. Манойло А. В., Петренко А. И., Фролов Д. Б. Государственная информационная политика в условиях информационно-психологической войны. 4-е изд., перераб. и доп. — Горячая линия-Телеком Москва, 2020. — 636 с.
8. Костогрызов А. И. Прогнозирование рисков по данным мониторинга для систем искусственного интеллекта / БИТ. Сборник трудов Десятой международной научно-технической конференции – М.: МГТУ им. Н.Э. Баумана, 2019, с. 220-229.
9. A. Kostogryzov and V. Korolev, Probabilistic Methods for Cognitive Solving of Some Problems in Artificial Intelligence Systems (Вероятностные методы для когнитивного решения некоторых задач в системах искусственного интеллекта). Probability, combinatorics and control / IntechOpen, 2020, pp. 3-34. — URL: https://www.intechopen.com/books/probability-combinatorics-and-control
10. Kostogryzov A., Nistratov A., Nistratov G. (2020) Analytical Risks Prediction. Rationale of System Preventive Measures for Solving Quality and Safety Problems. In: Sukhomlin V., Zubareva E. (eds) Modern Information Technology and IT Education. SITITO 2018. Communications in Computer and Information Science, vol 1201. Springer, pp.352-364. https://www.springer.com/gp/book/9783030468941
11. Kostogryzov A, Nistratov A., Probabilistic methods of risk predictions and their pragmatic applications in life cycle of complex systems. In “Safety and Reliability of Systems and Processes”, Gdynia Maritime University, 2020. pp. 153-174. DOI: 10.26408/srsp-2020
12. Нистратов А.А., Аналитическое прогнозирование интегрального риска нарушения приемлемого выполнения совокупности стандартных процессов в жизненном цикле систем высокой доступности. Часть 1. Математические модели и методы // Системы высокой доступности. 2021. Т.17 №3, с. 16–31, Часть 2. Программно-технологические решения. Примеры применения // Системы высокой доступности. 2022. Т.18 №2, с. 42–57.
13. Костогрызов А.И. О моделях и методах вероятностного анализа защиты информации в стандартизованных процессах системной инженерии //Вопросы кибербезопасности. 2022, №6(52), с.71-82.
14. Kostogryzov A., Makhutov N., Nistratov A., Reznikov G., Probabilistic predictive modeling for complex system risk assessments (Вероятностное упреждающее моделирование для оценок рисков в сложных системах). Time Series Analysis - New Insights. IntechOpen, 2023, pp. 73-105. http://mts.intechopen.com/articles/show/title/probabilistic-predictive-modelling-for-complex-system-risk-assessments
15. Костогрызов А. И. Подход к вероятностному прогнозированию защищенности репутации политических деятелей от «фейковых» угроз в публичном информационном пространстве // Вопросы кибербезопасности. 2023, №3. С. 114–133. DOI:1021681/2311-3456-2023-3-114-133

THREAT ANALYSIS OF MALICIOUS MODIFICATION OF THE MACHINE LEARNING MODEL FOR ARTIFICIAL INTELLIGENCE SYSTEMS

Kostogryzov A.I[7], Nistratov A.A.[8]

Objective: *to propose a methodological approach for probabilistic analysis of the correctness of the trained software tools (SW) for artificial intelligence systems (AIS) during their development and operation in conditions of potential threats of malicious modification of the machine learning model (MLM).*

7 Andrey I. Kostogryzov, Dr.Sc., Professor, Chief Researcher, Federal Research Center «Informatics and Control» of the Russian Academy of Sciences. Moscow, Russia. E-mail: Akostogr@gmail.com

8 Andrey A. Nistratov, PhD, Senior researcher, Federal Research Center «Informatics and Control» of the Russian Academy of Sciences. Moscow, Russia. E-mail: andrey.nistratov.job@yandex.ru

Research methods include *methods of probability theory, methods of system analysis. The approach is based on the adaptation of the author's probabilistic models developed earlier to assess the quality of the information used and risk management, which are fixed to the level of implementation in GOST R 59341-2021 "System engineering. Protection of information in system information management process".*

Result: *Under the conditions of accepted suppositions and assumptions, probabilistic models have been developed to assess the particular risks of non-detection of inaccuracies in machine learning during the development and operation of SW, as well as a method for assessing the integral risk of violation of the correctness of machine learning during specified period of prediction. Actual threat of MLM spoofing and the threat of MLM modification by poisoning the training data are analyzed. Proposals have been developed for the formation of input for risks prediction using the proposed models. The approach is illustrated by calculation examples with quantitative assessments, risk dependencies on the input and the rationale of recommendations.*

Scientific novelty: *For the conditions of potential threats of malicious MLM modification, models and methods for assessing the particular risks of non-detection of incorrectness in machine learning during AIS development and operation and the integral risk are proposed.*

Keywords: *probability, poisoning of training data, model, risk, system, threats.*

References

1. E`rtel` V., Vvedenie v iskusstvenny`j intellekt.-M. «E`ksmo», 2019. – 448s.
2. Lekun Yan, Kak uchitsya mashina (revolyuciya v oblasti nejronny`x setej i glubokogo obucheniya). – M. Al`pina PRO, 2021. – 335s.
3. Arlazarov V. V., Mobil`noe raspoznavanie i ego primenenie k sisteme vvoda identifikacionny`x dokumentov. – Dissertaciya na soiskanie uchenoj stepeni doktora texnicheskix nauk. -M. FICz IU RAN, 2023. – 358s.
4. Chakraborty A., Alam M., Dey V., Chattopadhayay A.U., Yay D.M., Adversarial attacks and defences: A survey //arXiv preprint arXiv:1810.00069. – 2018
5. Probabilistic modeling in system engineering. InTechOpen, 2018, 279p. – URL: http://www.intechopen.com/books/probabilistic-modeling-in-system-engineering
6. Klimov S. M. Modeli analiza i ocenki ugroz informacionno-psixologicheskix vozdejstvij s e`lementami iskusstvennogo intellekta. / Sbornik dokladov i vy`stuplenij nauchno-delovoj programmy` Mezhdunarodnogo voenno-texnicheskogo foruma «Armiya-2018». 2018. S. 273–277.
7. Manojlo A. V., Petrenko A. I., Frolov D. B. Gosudarstvennaya informacionnaya politika v usloviyax informacionno-psixologicheskoj vojny`. 4-e izd., pererab. i dop. — Goryachaya liniya-Telekom Moskva, 2020. — 636 s.
8. Kostogry`zov A. I. Prognozirovanie riskov po danny`m monitoringa dlya sistem iskusstvennogo intellekta / BIT. Sbornik trudov Desyatoj mezhdunarodnoj nauchno-texnicheskoj konferencii – M.: MGTU im. N.E`. Baumana, 2019, s. 220-229.
9. A. Kostogryzov and V. Korolev, Probabilistic Methods for Cognitive Solving of Some Problems in Artificial Intelligence Systems (Veroyatnostny`e metody` dlya kognitivnogo resheniya nekotory`x zadach v sistemax iskusstvennogo intellekta). Probability, combinatorics and control / IntechOpen, 2020, pp. 3-34. — URL: https://www.intechopen.com/books/probability-combinatorics-and-control
10. Kostogryzov A., Nistratov A., Nistratov G. (2020) Analytical Risks Prediction. Rationale of System Preventive Measures for Solving Quality and Safety Problems. In: Sukhomlin V., Zubareva E. (eds) Modern Information Technology and IT Education. SITITO 2018. Communications in Computer and Information Science, vol 1201. Springer, pp.352-364. https://www.springer.com/gp/book/9783030468941
11. Kostogryzov A, Nistratov A., Probabilistic methods of risk predictions and their pragmatic applications in life cycle of complex systems. In "Safety and Reliability of Systems and Processes", Gdynia Maritime University, 2020. pp. 153-174. DOI: 10.26408/srsp-2020
12. Nistratov A.A., Analiticheskoe prognozirovanie integral`nogo riska narusheniya priemlemogo vy`polneniya sovokupnosti standartny`x processov v zhiznennom cikle sistem vy`sokoj dostupnosti. Chast` 1. Matematicheskie modeli i metody` // Sistemy` vy`sokoj dostupnosti. 2021. T.17 №3, s. 16–31, Chast` 2. Programmno-texnologicheskie resheniya. Primery` primeneniya // Sistemy` vy`sokoj dostupnosti. 2022. T.18 №2, s. 42–57.
13. Kostogry`zov A.I. O modelyax i metodax veroyatnostnogo analiza zashhity` informacii v standartizovanny`x processax sistemnoj inzhenerii //Voprosy` kiberbezopasnosti. 2022, №6(52), s.71-82.
14. Kostogryzov A., Makhutov N., Nistratov A., Reznikov G., Probabilistic predictive modeling for complex system risk assessments (Veroyatnostnoe uprezhdayushhee modelirovanie dlya ocenok riskov v slozhny`x sistemax). Time Series Analysis - New Insights. IntechOpen, 2023, pp. 73-105. http://mts.intechopen.com/articles/show/title/probabilistic-predictive-modelling-for-complex-system-risk-assessments
15. Kostogry`zov A. I. Podxod k veroyatnostnomu prognozirovaniyu zashhishhennosti reputacii politicheskix deyatelej ot «fejkovy`x» ugroz v publichnom informacionnom prostranstve // Voprosy` kiberbezopasnosti. 2023, №3. S. 114–133. DOI:1021681/2311-3456-2023-3-114-133