УДК 004.048
DOI: 10.52190/2073-2600_2024_2_17
EDN: ICEKIG

Технология "слепой" обработки привлекаемых данных в системах машинного обучения

[1, 2, 3] *В. А. Конявский*, д-р техн. наук; [1, 2] *С. В. Конявская-Счастная*, канд. филолог. наук; [3] *Г. В. Росс*, д-р техн. наук; [1, 7, 8, 9] *А. М. Райгородский*, д-р физ.-мат. наук; [1] *С. А. Тренин*; [1, 4] *А. В. Леонидов*, д-р физ.-мат. наук; [1, 4] *Е. Е. Васильева*, канд. физ.-мат. наук; [1, 4, 5] *С. Б. Васильев*, канд. физ.-мат. наук; [6] *М. Ю. Коновалихин*, д-р техн. наук

[1] Московский физико-технический институт (НИУ), Московская обл., г. Долгопрудный, Россия; [2] АО «ОКБ САПР», Москва, Россия; [3] РЭУ им. Г. В. Плеханова, Москва, Россия; [4] Физический институт им. П. Н. Лебедева РАН, Москва, Россия; [5] НИУ «ВШЭ», Москва, Россия; [6] Банк ВТБ (ПАО), Санкт-Петербург, Россия; [7] МГУ им. М. В. Ломоносова, Москва, Россия; [8] Кавказский математический центр Адыгейского государственного университета, Республика Адыгея, г. Майкоп, Россия; [9] Бурятский государственный университет им. Доржи Банзарова, Республика Бурятия, г. Улан-Удэ, Россия

Статья посвящена разрешению противоречия между потребностью в обработке объединенных данных при построении прогнозных моделей и отсутствием технических решений по обеспечению достаточного уровня защищенности информации, сложившегося к настоящему времени. Вводятся понятия "слепой" обработки данных, неизвлекаемых данных и моделей, привлекаемых данных. Приведены примеры косвенных утечек, возникновение которых, как правило, связано с дифференциальными атаками. Для повышения уровня защищенности данных предлагается при слепой обработке данных ограничить запросы к системе машинного обучения только проверенными последовательностями команд, не приводящими к утечкам защищаемых данных — ИТ-конвейерами. Предложена формальная модель обработки данных "вслепую", референсная архитектура автоматизированной системы "слепой" обработки данных, структура и меры по фиксации и обеспечению целостности и легальности применения ИТ-конвейеров, особенности их формирования и применения. На основе полученных результатов создан программно-аппаратный комплекс "Крипто-Анклав". Результаты также могут применяться при создании национальной системы антифрода, в медицине, промышленности.

Ключевые слова: машинное обучение, безопасность, "слепая" обработка данных, конфиденциальное машинное обучение, ИТ-конвейер, объединение данных, защищаемые данные, неизвлекаемость данных, прямые утечки, косвенные утечки, дифференциальные атаки, конфиденциальные вычисления, национальный оператор антифрода.

Конявский Валерий Аркадьевич, зав. кафедрой защиты информации, научный руководитель, главный научный сотрудник.
E-mail: konyavskiy@gospochta.ru
Конявская-Счастная Светлана Валерьевна, заместитель генерального директора, доцент кафедры защиты информации.
E-mail: cd@okbsapr.ru
Росс Геннадий Викторович, профессор, главный научный сотрудник.
E-mail: ross-49@mail.ru
Райгородский Андрей Михайлович, директор ФПМИ, профессор кафедры математической статистики и случайных процессов, руководитель, профессор.
E-mail: mraigor@yandex.ru
Тренин Сергей Алексеевич, вед. инженер, руководитель ключевых научно-технических направлений лаборатории прикладных исследований.
E-mail: s.trenin@gmail.com
Леонидов Андрей Владимирович, профессор, профессор кафедры дискретной математики, ведущий научный сотрудник лаборатории физики высоких энергий, заведующий лаборатории математического моделирования сложных систем.
E-mail: leonidovav@lebedev.ru
Васильева Екатерина Евгеньевна, научный сотрудник лаборатории математического моделирования сложных систем отделения теоретической физики, инженер лаборатории прикладных исследований.
E-mail: serebryannikovaee@lebedev.ru
Васильев Сергей Борисович, заместитель начальника отдела сопровождения проектной работы, младший научный сотрудник, инженер лаборатории прикладных исследований.
E-mail: svasilev@hse.ru
Коновалихин Максим Юрьевич, старший вице-президент, руководитель Департамента анализа данных и моделирования.
E-mail: konovalihin@vtb.ru

Статья поступила в редакцию 27 марта 2024 г.

Цифровые данные и искусственный интеллект

Известно, что при осмысленной обработке больших данных возможно извлечение знаний, способных оказать существенное влияние на качество принимаемых решений в различных задачах общества, государства и бизнеса [1]. Автоматизированные системы (АС), предназначенные для поддержки такого рода деятельности, относятся к системам искусственного интеллекта (СИИ), системам машинного обучения и другим системам этого типа. Успех получаемого решения существенно зависит от того, насколько полными и качественными являются массивы данных, используемые в СИИ. Возникает потребность в обогащении накопленных данных [2] данными, находящимися в распоряжении других операторов.

При этом значительная часть накапливаемых операторами больших данных относится к категории персональных и других защищаемых данных [3], оборот которых регулируется законодательством, в том числе 152-ФЗ [4] и связанными с ним подзаконными актами [5], в совокупности существенно ограничивающими возможности обмена данными указанного типа.

Таким образом, сформировалось противоречие между потребностями хозяйствующих субъектов в сборе и обработке больших данных, с одной стороны, и ограниченными возможностями осуществлять эту деятельность в рамках существующих технических решений и норм правового регулирования, с другой стороны.

Преодоление данного противоречия является научной задачей и возможно путем разработки специализированных технологий, методов и средств обработки защищаемых данных без ознакомления с ними, "вслепую".

Этот тип технологий назовем технологиями "слепой" обработки данных (ТСОД), а средства вычислительной техники (СВТ), реализующие ТСОД, — СВТ ТСОД. Здесь и далее используем термин "слепая обработка данных" в отличие от известного из радиотехники термина "слепая обработка сигналов" [6].

Нужно отметить, что целью извлечения знаний из больших данных является установление закономерностей, характеризующих предметную область [7, 8]. Построенные в результате машинного обучения модели и результаты их применения носят прогнозный характер, не относятся к персональным данным, но могут при этом быть защищаемыми данными.

Исследование отдельных вопросов конфиденциальности машинного обучения давно привлекают внимание исследователей [9, 10], однако целостное решение пока отсутствует.

Накопленный специалистами опыт манипуляций с данными для извлечения знаний позволяет автоматически выполнять многие процессы, например, за счет использования методов ИИ [11, 12], в том числе без участия оператора, и, следовательно, без ознакомления с накопленными и объединенными данными.

Отметим также, что в соответствии со ст. 71 Конституции [13] к вопросам ведения Российской Федерации отнесены вопросы, касающиеся обеспечения "безопасности личности, общества и государства при применении информационных технологий, обороте цифровых данных". Тем самым впервые в нормативной правовой базе использовано понятие "оборот цифровых данных" и само понятие "цифровые данные". Пока это понятие не определено, но сам факт его использования демонстрирует, что законодатель различает понятия "информация" и "цифровые данные", что при должной проработке может создать правовые условия для привлечения и совместной обработки защищаемых данных в цифровой форме.

Бизнес-процесс и подходы к технической защите информации. В перечень работ при построении прогнозных моделей с использованием СИИ в самом общем виде включают подготовку наборов данных (НД) и их обогащение, машинное обучение (построение модели), валидацию модели и применение модели, завершающиеся публикацией отчетов. Схематично возможный процесс показан на рис. 1.

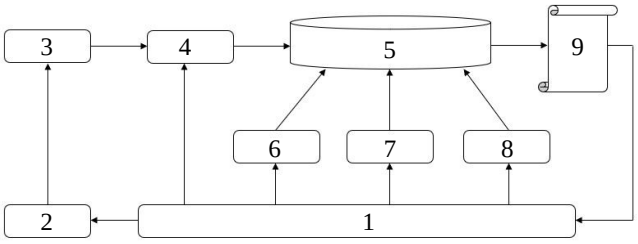

Рис. 1. Построение прогнозной модели:
1 — участники проектов, поставщики данных; 2 — инициирование проекта; 3 — подготовка НД; 4 — загрузка данных; 5 — обогащенный набор данных; 6 — построение модели; 7 — валидация модели; 8 — применение модели; 9 — публикация отчетов

Все этапы может выполнять как один человек, так и разные, специализирующиеся на особенностях этапов. Очевидным образом выделяются роли поставщика данных, инициатора проекта, специалиста по загрузке НД в систему, разработчика модели, валидатора модели, менеджера по применению обученной модели (см. рис. 2).

Рис. 2. Роли участников процесса машинного обучения и применения обученной модели

Наборы данных, как правило, представляют собой плоские таблицы (матрицы "объекты-признаки"), столбцы (признаки) которых поименованы, а строки содержат соответствующие данные о субъектах данных (об объектах) [14]. При этом для каждого столбца может быть указана дата формирования, а некоторые столбцы могут быть использованы как параметры (признаки) для объединения данных для обогащения НД. Различают горизонтальное и вертикальное обогащения (рис. 3).

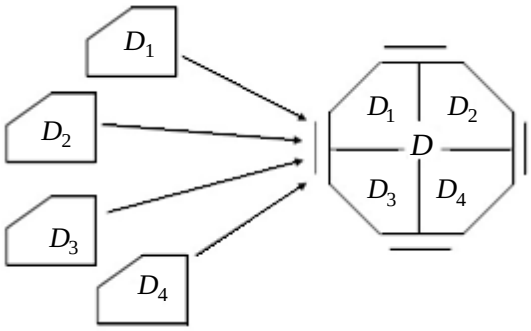

Рис. 3. Объединение данных. Формирование обогащенного НД:
D_1–D_2 и D_3–D_4 — горизонтальное обогащение;
D_1–D_3 и D_2–D_4 — вертикальное обогащение

Обогащенные (в общем случае) НД искусственным образом делятся на две части — для обучения (обучающая выборка) и для валидации — то есть для проверки адекватности полученной при машинном обучении модели. В данном случае рассмотрим обучение по прецедентам, основанное на выявлении в НД эмпирических закономерностей. Разбиение НД, как правило, осуществляется формированием т. н. "целевого столбца", который формирует разработчик модели (рис. 4). Выбирается тип модели, и в результате машинного обучения формируется алгоритм, представляемый, как правило, коэффициентами модели выбранного типа.

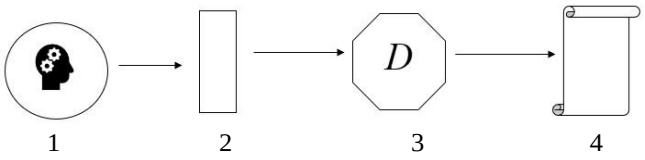

Рис. 4. Обучение модели:
1 — разработчик модели; 2 — целевой столбец;
3 — объединенный НД; 4 — отчет об обучении

Качество полученного алгоритма проверяется на этапе валидации применением входных данных за пределами обучающей выборки. Эта операция выполняется валидатором модели (рис. 5). При этом формируются отчеты, содержащие достаточные для оценки качества характеристики.

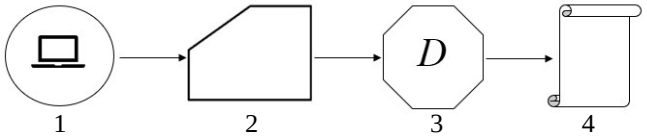

Рис. 5. Валидация модели
1 — валидатор модели; 2 — данные валидации; 3 — модель на объединенном НД; 4 — отчет о валидации

Представленный обученной моделью предсказательный алгоритм может менять свои характеристики, например, устаревать с течением времени. Поэтому валидацию, как правило, необходимо проводить периодически, оценивая адекватность обученной модели.

Если на этапе валидации установлена достаточная адекватность обученной модели, то ее можно использовать для получения прогнозных значений с применением новых данных о новых объектах (в данном случае — о субъектах персональных данных). Это функция менеджера по применению модели (рис. 6).

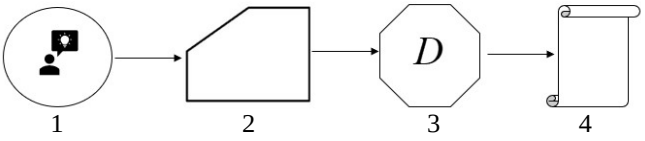

Рис. 6. Применение модели:
1 — менеджер применения; 2 — данные для применения; 3 — обученная модель на объединенных данных; 4 — отчет

Как правило, НД можно охарактеризовать как актуальные или как исторические. Этим определяется область их применения — так, исторические данные наряду с актуальными вполне можно использовать для построения модели и ее валидации, а при применении модели можно использовать в соответствии с буквой закона только актуальные НД.

Описанная технология основана на том, что обработка ведется уполномоченными сотрудниками оператора персональных данных, которые имеют права на ознакомление и обработку персональных данных в соответствии с информированным согласием субъектов этих данных.

Нормативные правовые акты (НПА) в сфере защиты персональных данных[1] ограничивают сбор и обработку ПДн наличием информированного согласия субъекта данных. Фактически, объединение данных в целях обогащения НД при этом значительно затруднено, так как использовать наборы данных, накопленные другими операторами, практически невозможно. Это негативно влияет на качество прогнозных моделей, в частности на результаты кредитного скоринга. Как результат, кредитные организации несут значимые потери, которых вполне можно было бы избежать при учете всех накопленных данных.

Если для обучения моделей использовать НД других операторов, объединяя данные кредитных организаций, ритейла, связи и других можно заметно повысить качество моделей, но это можно делать при условии строгого соответствия требованиям НПА. В первую очередь, это означает, что трансформация данных должна выполняться "вслепую", без ознакомления с самими данными, любые варианты утечки защищаемых данных исключаются.

Наборы данных, формируемые другими операторами и используемые для обогащения основного НД, будем называть привлекаемыми данными.

Разработчик модели привлекает для обогащения НД данные других операторов — привлекаемые данные.

Рассмотрим один из вариантов "слепого" машинного обучения, обращая особое внимание на задачи предотвращения утечек (рис. 7).

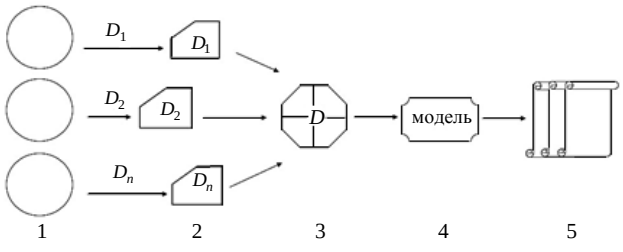

Рис. 7. Обработка привлекаемых данных:
1 — подготовка НД; 2 — загрузка НД; 3 — обогащение НД; 4 — машинное обучение; 5 — публикация отчетов

Итак, для использования привлекаемых данных должны быть созданы условия, исключающие ознакомление с ними и их распространение.

Поскольку характеристики и объемы привлекаемых данных могут быть различными, уровень их защиты целесообразно выбрать наиболее высокий, обеспечивающий всю совокупность требований как федерального законодательства, так и ведомственной нормативной базы [15].

Все информационно-вычислительные процессы машинного обучения, включая построение модели, валидацию и применение, целесообразно выполнять в некотором средстве вычислительной техники, изготовленном в виде "черного ящика", уже упомянутом СВТ ТСОД. СВТ ТСОД выполняет все вычислительные операции, в том числе расшифрование данных, поступающих извне, их обработку и зашифрование данных, подготовленных для выдачи в качестве результирующих отчетов. Вне СВТ ТСОД все данные всегда находятся в зашифрованном виде, исключающем доступ и ознакомление с ними, что, в свою очередь, исключает при проведении этих операций утечку защищаемых данных и данных, подлежащих защите.

Неконтролируемые (неограниченные) запросы операторов и публикуемые отчеты также могут быть источником утечек, так как на этой основе могут быть организованы дифференциальные атаки. Для предотвращения таких утечек нужно ограничить возможности взаимодействия оператора с системой, а также перечень отчетов, который должен включать конечный согласованный набор, контролируемый защитными механизмами. Кроме этого, публикуемые данные в ряде случаев целесообразно контролировать также методами DLP [16, 17].

Утечки через запросы оператора можно наблюдать как при использовании произвольных команд обработки данных, так и при нарушении порядка доступа к данным и манипулировании данными, например, если при запросах используют данные, которые не предназначены для этих целей и в отношении применения которых нет информированного согласия субъекта.

Утечки могут возникать также при анализе обученной модели, так как защищаемые данные вполне могут оказаться среди коэффициентов, получаемых при обучении.

Таким образом, в целях преодоления сложившегося противоречия и обеспечения возможности применения в СИИ привлекаемых данных, необходимо провести исследования и выработать предложения, обеспечивающие доверие участников, в том числе:

– слепую обработку данных в СИИ;
– неизвлекаемость данных и моделей;
– ограничения последовательностей команд и

[1] Например, ФЗ-152 и другие.

применения данных, позволяющие осуществлять трансформацию данных без ознакомления с ними и их распространения.

Одним из важных механизмов обеспечения доверия может быть применение только проверенных и согласованных последовательностей команд, которые обозначим термином "ИТ-конвейер".

Конвейером назовем зафиксированную последовательность команд обработки данных, не приводящую к утечкам, включая ознакомление с данными.

Искомая последовательность фиксируется разработчиком и регуляторами посредством электронной подписи (ЭП) [18], правильность которой контролируется перед началом обработки данных [19].

Необходимо также подготовить примерную, референсную архитектуру системы, основанной на использовании технологий слепой обработки данных.

Неизвлекаемость данных

Методы защиты систем искусственного интеллекта до настоящего времени не разработаны в достаточной степени, хотя потребность в системах этого класса осознана [20, 21]. Опыт, на который можно опираться, очень невелик.

Все стандартные методы технической защиты информации (ТЗИ) для корпоративных систем могут и должны применяться и в системах ИИ, но при этом их следует дополнить новыми, специфическими методами, учитывающими особенности технологий машинного обучения и обработки больших данных, в том числе содержащих персональные.

Данные, обрабатываемые в СИИ, ни при каких обстоятельствах не должны извлекаться из нее. Появляется необходимость обеспечить выполнение новой функции безопасности — функции неизвлекаемости данных (ФНД). Реализация этой функции может быть основана на принципе функции безопасности "неизвлекаемый ключ" [22—24]. Опыт разработки механизмов работы с неизвлекаемым ключом и успешной сертификации позволяет рассчитывать на успех в реализации и ФНД.

Суть неизвлекаемости сводится к тому, что в составе СВТ ТСОД есть возможность использовать данные в соответствии с потребностями технологии, но отсутствует возможность передачи защищаемых данных, ознакомления с ними и их модификации. При наличии достаточных средств защиты информации и правильной их настройке, а также при контроле входных последовательностей команд, такой режим обеспечивается и проверяется довольно несложно. Конечно, ограничение "свободы действий" специалиста по разработке моделей несколько усложняет его работу, но при этом дает возможность использовать привлекаемые данные, что существенно повышает качество моделей.

Отметим, что обеспечение ФНД является необходимым условием использования СВТ ТСОД, но не является достаточным. Достаточность защиты должна обеспечиваться реализацией мер безопасности, предусмотренных требованиями регуляторов, устанавливающих правила работы с защищаемыми данными, а именно ФСБ России, ФСТЭК России, Минцифразвития, ЦБ.

Неизвлекаемость данных — свойство, обеспечиваемое специальным режимом доступа к данным, при котором исключается раскрытие персональных и других защищаемых данных, ознакомление с ними, копирование, модификация, но допускается обработка и использование для извлечения знаний.

Свойство неизвлекаемости данных не является абсолютным, как и другие свойства данных в предметной области ТЗИ — целостность, доступность, конфиденциальность. Наиболее распространенной количественной оценкой таких свойств является уровень доверия, устанавливаемый при испытаниях [25]. По аналогии, оценка свойства неизвлекаемости может быть дана в форме уровня доверия, устанавливаемого на основе выполненного анализа в процессе сертификационных испытаний. Для осуществления этой оценки должна быть выработана и обоснована система координат, в которой могут быть сформулированы параметры свойства неизвлекаемости и оценка его реализации. Это новая задача научного уровня.

Для обеспечения неизвлекаемости защищаемые данные загружают в СВТ ТСОД в зашифрованном виде. Перемещение защищаемых данных в СВТ ТСОД не приводит к их утечке. Извлечь данные из хранилища и ознакомиться с ними невозможно, даже при том, что извлекать знания о закономерностях в автоматическом режиме можно.

Начальная обработка выполняется автоматически — без участия человека. Здесь под начальной обработкой понимается обеспечение безопасного "криптографического скачка" — расшифрование с использованием ключа поставщика и зашифрование на ключе хранения. После безопасной загрузки данных можно реализовывать технологию "слепого" машинного обучения, при которой использование данных осуществляется автоматически, без участия человека, чем также обеспечивается неизвлекаемость данных.

Для защиты от дифференциальных атак участие оператора необходимо ограничить применением только тех последовательностей команд, которые не приводят к утечкам.

Ограничение команд, обеспечивающих неизвлекаемость данных, должно выполняться по результатам изучения конкретных информационных технологий, предназначенных для слепой обработки данных. Термин "информационная технология" (ИТ) ниже будем понимать как процесс, состоящий из последовательности информационных операций над данными [26]. ИТ, обеспечивающая неизвлекаемость, должна быть согласована и зафиксирована.

Согласование подтверждается, а фиксация обеспечивается — электронной подписью регулятора. Согласование возможно в случае, если для данной ИТ (последовательности операций, конвейера операций) выполняется основное требование: в составе публикуемых результатов полностью автономной обработки (без участия человека) не содержатся защищаемые данные. В этом случае ИТ подписывается электронной подписью (ЭП) регулятора.

Зафиксированная безопасная последовательность операций трансформации данных без ознакомления с ними может быть подписана ЭП разных регуляторов — если предполагается обработка наборов данных, относящихся к разным видам тайн. Например, при обработке наборов, содержащих сведения о персональных данных, тайне связи и банковской тайне соответствующая ИТ должна быть подписана ЭП уполномоченных представителей Роскомнадзора (РКН) и ЦБ, а наборы данных должны быть маркированы соответствующим образом.

При положительном результате проверки ЭП проверенная ИТ исполняется, при отрицательном — отклоняется. Результат проверки отражается в журнале аудита. Проверка ЭП выполняется сертифицированными средствами электронной подписи (СЭП) с использованием ключа проверки подписи ответственного сотрудника регулятора.

Таким образом, обеспечить безопасную работу систем ИИ с большими данными, относящимися к категории защищаемых, в том числе персональными данными и данными, относимыми к банковской тайне, тайне связи и другим видам тайн — технически и организационно вполне возможно.

Основы проектирования системы слепой обработки данных с ФНД

Как было показано, неизвлекаемость данных связана с блокированием возможности как прямых, так и косвенных утечек.

Прямые утечки — пользователь получает несанкционированный доступ на ознакомление с защищаемыми данными, и/или в публикуемых отчетах содержатся персональные данные и/или данные, с использованием которых можно идентифицировать человека, а также данные, содержащиеся в составе защищаемых данных.

Косвенные утечки — в публикуемых отчетах содержатся данные, с использованием которых можно получить данные из состава защищаемых данных. Анализ возможности косвенных утечек пока в достаточной мере не разработан и не систематизирован в открытой печати, и поэтому требует детального изучения.

Для обеспечения защиты данных рассмотрим формальное описание работы с данными в СВТ ТСОД.

Формальное описание работы с данными

Участники

1. Множество организаций $\mathcal{J} = \{1, \dots, I\}$, индекс принадлежности $i \in J$. Организация $i = 1$ — собственник Анклава. Каждая организация i является оператором персональных данных.

2. В каждой организации есть множество сотрудников, работающих в той или иной роли с Анклавом. Множество этих сотрудников $\varepsilon_i = 1, \dots, \mathrm{E}_i$, будем обозначать сотрудников индексом $e_i \in \varepsilon_i$.

3. Множество ролей $\mathcal{R} = \{1, \dots, R\}$, будем обозначать роли индексом $r \in \mathcal{R}$. Под ролью понимается возможность выполнения тех или иных операций с данными. На данный момент имеется следующий список ролей (R = 6):

- поставщик данных (r = 1),
- оператор загрузки данных (r = 2),
- инициатор проекта (r = 3),
- разработчик модели (r = 4),
- валидатор модели (r = 5),
- менеджер применения модели (r = 6).

4. Каждый сотрудник может выполнять одну или несколько ролей. Множество ролей, выполняемых сотрудником e_i организации i — это отображение $f_i(e_i): \varepsilon_i \to \mathcal{R}$. При этом один сотрудник может выполнять несколько разных ролей, т. е. $f_i(e_i) \subset \mathcal{R}$.

Данные участников

1. Множество столбцов-идентификаторов $\mathcal{ID} = \{id_1, ..., id_N\}$, индекс принадлежности $id \in \mathcal{ID}$. По этим столбцам может производиться

сопоставление и объединение данных разных участников-поставщиков данных. Примерами столбцов такого рода могут быть ФИО, телефон, ИНН и т. п.

2. Множество дат формирования столбцов $\mathcal{IDT} = \{id_1, ..., idt_{D_i}\}$.

3. Каждая организация i является поставщиком массива данных $\mathcal{D}_i = \{id_1, ..., i, x_{i1}, x_{i2}, ..., x_{iD_i}\}$, где x_{iD} — произвольный столбец данных (признак), которым владеет организация i, $d = \{1, ..., D_i\}$, D_i — количество столбцов данных, поставляемых организацией i. Все эти данные относятся к категории персональных данных и собираются организацией i только с разрешения клиентов. Количество строк в поставляемых организацией i данных $\mathcal{D}_i$ обозначим S_i, а множество строк $S_i = \{S_{i1}, ..., S_{iS_i}\}$. Предполагается, что одной сущности отвечает одна строка (см. рис. 8). Таким образом, S_i — мощность множества сущностей организации i, на обработку которых дали согласие клиенты организации i (в рамках проекта).

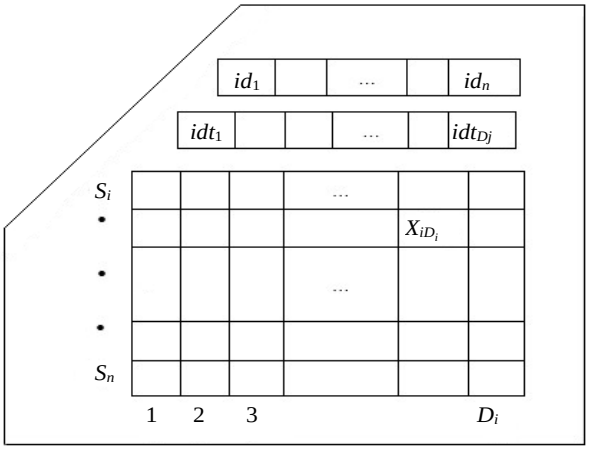

Рис. 8. Структура данных участника

Используемые данные

1. При обучении и валидации модели допустимо использование исторических данных

2. При применении модели допустимо использование только актуальных данных

3. В системе хранится словарь *providerDict*, сопоставляющий каждому столбцу индекс принадлежности поставщика данных:

$$\text{providerDict} = \{x_{11}{:}1, x_{12}{:}1, ..., x_{1D_1}{:}1, ..., x_{i1}{:}i, x_{i2}{:}i, ..., x_{iD_i}{:}i, x_{I1}{:}I, x_{12}{:}I, ..., x_{ID_I}{:}I\}.$$

Используя этот словарь, по имени каждого столбца можно восстановить индекс принадлежности поставщика данных.

4. Для каждой организации i в системе хранится словарь *actualDict*$_i$, характеризующий каждую строку, как актуальную (значение 1) или историческую (значение 0).

$$\text{actualDict} = \{s_{i1}{:}a_{i1}, ..., s_{is_i}{:}a_{i1}\},$$
$$a_{ij} \in \{0,1\} \forall j = 1, ..., S_i.$$

5. Сотрудники каждой организации, обладающие ролью инициатора проекта $(f(e_i) = 3)$, могут инициировать разработку проекта. Множество проектов организации i обозначим $\mathcal{P}_i = \{1, ..., P_i\}$, будем обозначать проект индексом принадлежности $p_i \in \mathcal{P}_i$.

6. Каждому проекту p_i однозначно соответствует целевой столбец t_{ip_i}. Этот столбец формируется инициатором проекта и содержит разметку строк массива данных $\mathcal{D}_i$, на основе которой:

- формируется обучающая выборка строк $S_i^{train\ p_i} \subset S_i$ (те субъекты данных организации i, для которых известно значение целевого признака (объясняемой переменной) в рамках проекта p_i и которые отнесены разработчиком модели $(f(e_i) = 4)$ к обучающим примерам). Для этих строк столбец t_{ip_i} содержит значения целевого признака: принадлежность к классу в случае задачи классификации, значение целевой (объясняемой) переменной в случае задачи регрессии);
- формируется тестовая выборка строк $S_i^{test\ p_i} \subset S_i$ (те субъекты данных организации i, для которых известно значение целевого признака в рамках проекта p_i и которые отнесены разработчиком модели $(f(e_i) = 4)$ к тестовым примерам). Для этих строк, так же как для строк из $S_i^{train\ p_i}$ столбец t_{ip_i} содержит значения разметки (принадлежность к классу в случае задачи классификации, значение целевой переменной в случае задачи регрессии).

Для применения модели могут быть использованы неразмеченные строки данных, для которых планируют применение модели в рамках проекта p_i. Обозначим множество этих строк $S_i^{apply\ p_i} \subset S_i$. Для этих строк столбец t_{ip_i} содержит некоторое специально установленное, зафиксированное правилами системы, значение, говорящее о том, что по данному объекту данных есть информация, но нет разметки и что для таких объектов в рамках проекта в результате применения модели должен

быть получен ответ (оцененный класс или значение в зависимости от типа задачи);

- выделяются строки, которые не рассматривают в рамках проекта $p_i : S_i^{\text{exclude } p_i} \subset S_i$. Для этих строк столбец t_{ip_i} содержит некоторое специально установленное, зафиксированное правилами системы, значение, говорящее о том, что по данному объекту данных есть информация, но нет разметки, и что эти объекты в рамках проекта p_i не рассматриваются.

Отметим, что

$$S_i^x \cap S_i^y = \varnothing, x \neq y, x, y \in$$
$$\in \{train\ p_i, test\ p_i, apply\ p_i, exclude\ p_i\}.$$

Количество строк в столбце t_{ip_i} равно числу строк S_i в данных организации i. Важно отметить, что организация i может получить какую-либо дополнительную информацию (в ходе процессов обучения, верификации и применения модели) только относительно s_i собственных клиентов, от которых получено согласие на обработку персональных данных.

7. Объединенный массив данных $\mathcal{D} = \{id_1, ..., id_N, x_{11}, x_{12}, ..., x_{1D_1}, ..., x_{I1}, x_{I2}, ..., x_{ID_I}\}$. Мощность данного множества (суммарное количество столбцов данных) $N + \sum_{i=1}^{I} D_i$. Количество строк S в данном массиве определяется, во-первых, количеством строк в массиве данных каждой организации (S_i) и, во-вторых, тем насколько пересекаются клиентские базы организаций. Так, если клиентские базы совсем не пересекаются $(\forall i, j \in \mathcal{J} : S_i \cap S_j = \varnothing)$, то количество строк в объединенном массиве данных $\mathcal{D}$ равняется $\sum_{i=1}^{I} S_i$. В противоположном крайнем случае, когда есть некоторая организация, клиентами которой являются также клиенты всех остальных организаций $(\exists i \in \mathcal{J} : \forall j \in \mathcal{J}\ S_i \cap S_j = S_j)$, имеем оценку числа строк $\max_i S_i$. Таким образом, $\max_i S_i \leq S \leq \sum_{i=1}^{I} S_i$.

8. Сотрудники каждой организации, обладающие ролью разработчика модели $(f(e_i) = 4)$, могут разрабатывать и сохранять в Анклаве множество моделей $\mathcal{M}_{ip_i} = 1, ..., M_{ip_i}$, индекс принадлежности $m_{ip_i} \in \mathcal{M}_{ip_i}$ для реализации проекта p_i. Каждая модель обучается на объединенном наборе данных $\mathcal{D}$, но с использованием целевой переменной t_{ip_i}, предоставленной данной организацией при инициации проекта p_i.

Взаимодействие сущностей системы

1. Каждая организация $i \in \mathcal{I}$ поставляет данные $\mathcal{D}_i$ в систему Анклав, находящуюся у собственника системы. Загрузка данных осуществляется сотрудником e_i организации i, обладающим ролью оператора загрузки данных $(f(e_i) = 2)$.

2. Внутри системы Анклав производится сопоставление и объединение данных по набору столбцов идентификаторов $\mathcal{JD}$. В результате внутри системы Анклав имеется объединенный массив данных вида $\mathcal{D} = \{id_1, ..., id_N, x_{11}, x_{12}, ..., x_{1D_1}, ..., x_{I1}, x_{I2}, ..., x_{1D_1}\}$.

3. Работники каждой из организаций $(e_i \in \varepsilon_i)$, обладающие ролью инициатора проекта $(f(e_i) = 3)$, могут инициировать создание нового проекта $p_i \in \mathcal{P}_i$. При этом каждому проекту p_i однозначно соответствует столбец целевой переменной t_{ip_i}.

4. Работники каждой из организаций $(e_i \in \varepsilon_i)$, обладающие ролью разработчика модели $(f(e_i) = 4)$, могут обучить новую модель $m_{ip_i} \in \mathcal{M}_{ip_i}$ в рамках реализации проекта p_i. Для этого используется

- объединенный массив данных $\mathcal{D}$;
- строки $S_i^{train\ p_i} \subset S_i$ — доля строк организации i, для которых в целевом столбце t_{ip_i} задана разметка организации i в рамках проекта p_i, отвечающая обучающей выборке, т. е. некоторая заданная доля тех строк, для которых имеется разметка в столбце t_{ip_i};
- настройки параметров обучения этого сотрудника при реализации данного проекта p_i. Каждой модели m_{ip_i} однозначно соответствует некоторый набор значений настроечных параметров.

В результате проведения этой операции работник e_i получает отчет с метриками качества обучения модели m_{ip_i} на обучающей выборке. Данный отчет включает в том числе информацию о полезности информации, предоставленной организацией i, и информации, предоставленной другими организациями, для получения модели.

На основе информации о качестве построенной модели работник может принять решение об из-

менении части настроечных параметров и повторном запуске процесса обучения

Результатом данного процесса является модель $m_{ip_i} \in \mathcal{M}_{ip_i}$.

5. Работники каждой из организаций (e_i), обладающие ролью валидатора модели $(f_i(e_i)=5)$, могут провести верификацию модели m_{ip_i}, разрабатываемой в рамках проекта p_i на части целевого столбца t_{ip_i}, организации i, которая была специально загружена для валидации применимости ранее построенной модели либо отложена для валидации модели, разрабатываемой в настоящий момент в рамках проекта p_i, т. е. на строках $S_i^{test\ p_i} \subset S_i$. В результате выполнения этой операции работник получает отчет о верификации модели.

6. Для применения модели m_{ip_i} сотрудник организации i (e_i), обладающий ролью менеджера применения $(f_i(e_i)=6)$, передает в Анклав идентификатор сохраненной в Анклаве модели m_{ip_i} и идентификаторы тех строк, для которых модель m_{ip_i} должна быть применена (строки $S_i^{apply\ p_i} \subset S_i$). Применение осуществляется на наборе данных, содержащем все столбцы из $\mathcal{D}$, но ограниченном на строки, определяемые переданным списком идентификаторов строк. На выходе этой операции сотрудник получает разметку для переданного списка идентификаторов.

Референсная архитектура АС на базе СВТ ТСОД и обеспечение технической защиты информации

Референсная архитектура — это предварительно разработанная примерная (референс — это пример, образец) архитектура, предназначенная для копирования или адаптации под особенности имеющихся ресурсов или фактических задач. Как правило, референсная архитектура разрабатывается для платформы (чипсета или какого-либо устройства, предназначенного для использования в качестве основы решения), однако к автоматизированной системе этот подход также применим и целесообразен.

Опишем референсную архитектуру АС (назовем ее АС ТСОД), с тем чтобы определить угрозы, актуальные для ее структурных и функциональных компонентов.

АС ТСОД строится по принципу матрешки (рис. 9).

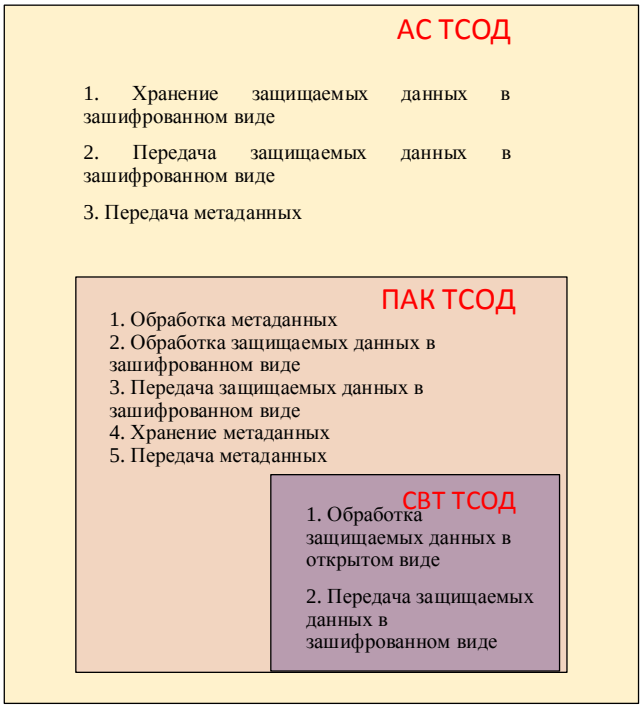

Рис. 9. Распределение функций АС ТСОД по контурам

Собственно обработкой данных в составе АС занимается программно-аппаратный комплекс (ПАК ТСОД), за рамками которого находятся терминалы Участников, возможно — СХД для хранения данных ПАК ТСОД в зашифрованном виде, а также хранилище, предназначенное для загрузки данных в систему извне, снаружи АС.

В свою очередь, ключевым элементом ПАК ТСОД является СВТ ТСОД, в границах которого обрабатывают защищаемые данные в открытом виде. За границы СВТ ТСОД, даже в пределах ПАК ТСОД, защищаемые данные не передаются, там не обрабатываются и не хранятся.

Соответствующим образом построена и инфраструктура защиты информации в АС ТСОД: каждый следующий контур усиливает защиту, и данные на более глубоких уровнях защищены всеми внешними контурами.

Исходя из этого располагаются компоненты функциональных подсистем АС: те, что работают с более критичными данными располагаются на структурных компонентах в более глубоко вложенном контуре.

Функциональная схема АС ТСОД выглядит следующим образом (рис. 10).

Все подсистемы, кроме ПМО — подсистемы машинного обучения, непосредственно обрабатывающей защищаемые данные в открытом виде, реализованы распределенно, то есть их части, обрабатывающие данные разной степени критичности, располагаются на структурных элементах разных уровней вложенности.

Рис. 10. Функциональная схема АС ТСО

Стрелка, показывающая ввод в подсистему зашифрованных данных, на представленной схеме идет из ниоткуда, поскольку реализация подсистемы загрузки данных и ее взаимодействия с подсистемой взаимодействия с пользователем существенно зависит от того, как будет построена в конкретной целевой АС ключевая система

ПМО полностью реализована в СВТ ТСОД, компоненте самого глубокого контура защиты, максимально защищенном.

Наложение функциональной схемы и контуров элементов (совпадающих с контурами защищенности) дает следующую картину (рис. 11).

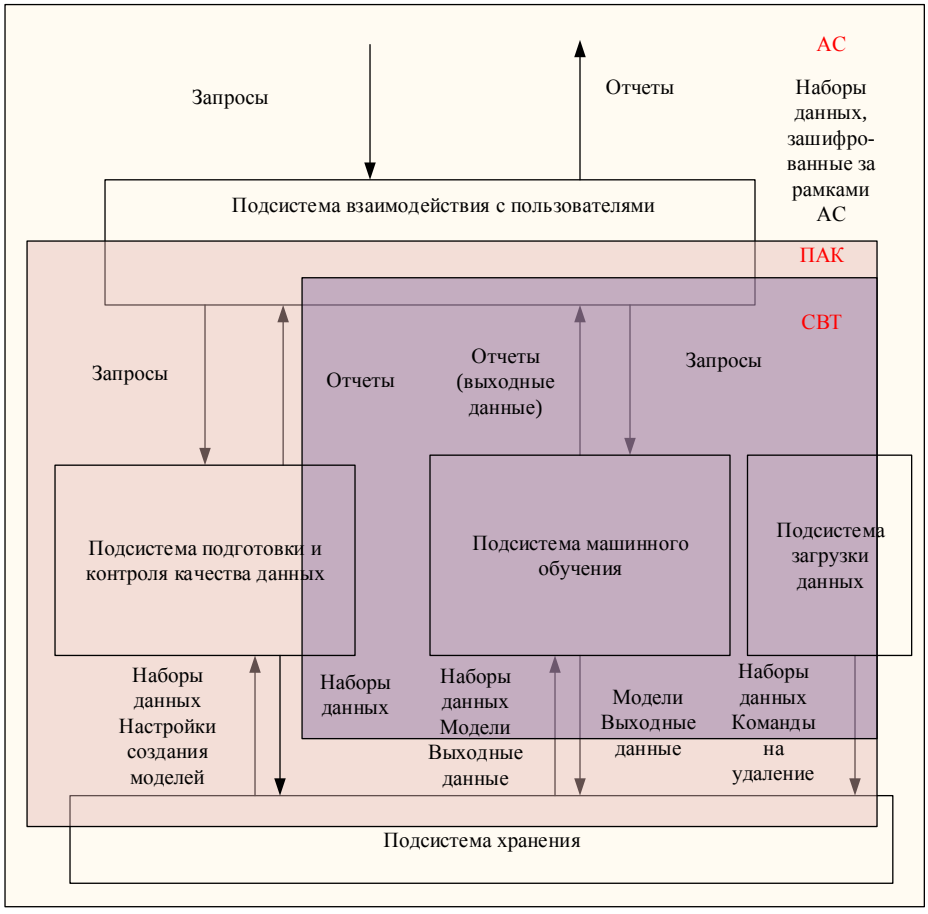

Рис. 11. Наложение функциональной схемы и контуров защиты АС ТСОД

Из функциональной схемы вытекает примерная структурная схема АС. Очевидно, что она должна включать в себя следующее.

1. СВТ ТСОД
- Сервера вычислений,
- Сервера безопасности (под общим названием "Крипто-сервер" будем понимать совокупность сервера СКЗИ, Крипто-шлюза, Канального шифратора (в случае, если архитектура включает кластер серверов вычислений), Сервера регистрации, Сервера мониторинга, Сервера инженерной (физической) защиты).

2. ПАК ТСОД
- СВТ ТСОД (потенциально — кластер из нескольких СВТ ТСОД, так как система должна быть масштабируемой),
- сервер работы с метаданными (далее — сервер среды метаданных, ССМ),
- сервер сетевой защиты (VPN, МЭ, СОВ).

3. АС ТСОД
- ПАК ТСОД,
- СХД,
- АРМ Пользователей (далее — Клиентские терминалы).

В соответствии с функциональной и структурной архитектурой АС проектируется подсистема защиты информации (ПЗИ), которая в свою очередь тоже строится по принципу "матрешки".

Каждый более глубокий контур защиты защищен всеми средствами защиты, защищающими предыдущие уровни, плюс теми средствами защиты, что свойственны именно ему.

Косвенные утечки. ИТ-конвейеры и исполняемые задания

Многие приемы, которые применяют в работе специалисты по машинному обучению, могут быть использованы для выявления *закономерностей* — на чем, собственно, и основана их деятельность. Но эти же приемы могут привести к возможности косвенных утечек — если из публикуемых отчетов можно выделить *данные* субъекта персональных данных. Возможность выделения данных субъекта и следует ограничивать.

Рассмотрим некоторые примеры таких утечек.

1. Близкие параметры измерений — если вычислить среднюю зарплату n сотрудников и $n + 1$ сотрудников, то не представляет труда узнать зарплату сотрудника с номером $n + 1$.

2. Даже при обезличивании данных заменой на корреляционный аналог, данные зачастую можно достаточно точно восстановить, зная математическое ожидание и дисперсию.

3. Возможно также восстановление при использовании только графических отображений, так как оценку некоторых параметров можно вполне извлечь и из графиков.

4. Анализ данных учета рабочего времени может дать информацию о должностных обязанностях сотрудников и об особенностях их личных отношений.

5. Возможна и наиболее опасна целевая атака — атака на получение данных о конкретном человеке. При этом внедренные коды вредоносного ПО (ВрПО) могут остаться незамеченными, так как для огромного большинства измерений будут выдаваться адекватные данные, и на основе их анализа утечка не может быть обнаружена.

6. Доступ к онтологиям позволяет оценить реальные данные при знании некоторых опорных значений.

7. Могут быть использованы также хорошо изученные методы имитационного моделирования. Например, если известно, что человек из п. А в п. Б добрался за время, которое не обеспечивается общественным транспортом, и при этом такси не вызывал — то должна быть собственная машина или сообщник.

8. Утечки, связанные с категориальными данными. Если разработчик модели знает названия столбцов и может настроить модель так, чтобы основным объясняющим признаком был интересующий его категориальный признак, то, подавая итеративно изменяемый столбец объясняемой переменной, пользователь может идентифицировать категорию, к которой принадлежит каждый объект данных.

9. Осмысленность имен автоматически сгенерированных признаков при слепой разработке моделей машинного обучения, также как и возможность добавления пользовательских новых признаков, являющихся функциями от имеющихся признаков, предоставленных разными поставщиками, также является потенциальным местом для возникновения косвенных утечек. Так, допустим, пользователь является поставщиком столбца a, он комбинирует его с неизвестным ему признаком b. Допустим, как новый объясняющий признак он получает $f(a,b)$, такую, что обратная функция $f^{-1}(x|a)$ известна. Настраивает модель так, что $f(a,b)$ является наиболее информативным признаком (специально или случайно). В результате, имея результаты моделирования и информацию столбца a, пользователь может восстановить неизвестный ему столбец b.

Если это делается специально (с использованием итеративного процесса подбирается нужный вид объясняемой переменной), то защититься от такого рода утечки можно путем ограничения на количество обращений. Однако такого рода ситу-

ация теоретически может возникнуть и случайно, когда $f(a,b)$ это, действительно, очень хороший объясняющий фактор. В этом случае защита от полного восстановления может состоять только в зашумлении результатов, которое, однако, снижает качество модели.

Этот перечень вполне может быть продолжен, что свидетельствует о сложности задачи обеспечения неизвлекаемости данных при слепом машинном обучении.

Базовое правило технической защиты информации состоит в том, что правильные результаты дает только [27]:

- проверенная программа,
- использующая правильные данные,
- функционирующая в проверенной среде.

Программа проверяется.

Данные подготавливаются.

Среда функционирования создается.

Как правило, среда функционирования создается[2] для решения не одной, а многих задач. То есть среда функционирования должна быть инвариантна задачам класса, для которого она создается, и должна обеспечивать контроль программ, предназначенных для исполнения.

Проверенная программа должна быть зафиксирована (например, ЭП или средствами КЦ).

Проверенные программы могут:

1) храниться в составе среды (на ресурсах СВТ в защищенном исполнении);

2) поступать из среды, внешней по отношению к СВТ, на которых выполняется обработка.

Первый вариант — это традиционный, привычный вариант. Он вполне подходит для ограниченного набора задач, но при изменении набора задач потребует внесения изменений в состав информационной системы, что в свою очередь приведет к новым затратам на мероприятия по обеспечению безопасности. Второй вариант организации взаимодействия намного более универсальный, и поэтому более предпочтительный[3].

Если рассматривать СВТ как универсальный вычислитель, решающий задачи заданного класса, то нужно обеспечить поступление в него из внешней среды заданий в целом — включающих как программы обработки, так и данные.

В этом случае необходимо обеспечить контроль заданий, имея ввиду совокупность программ, данных и их взаимное соответствие.

Ниже в целях упрощения текста, без потери общности, не будем различать объекты и ссылки на них, имея ввиду, что, например, под словом "данные" можно понимать как непосредственно данные, так и ссылки на них.

Исполняемое задание (ИЗ) в общем случае можно представить как кортеж: *данные—описание данных—последовательность операций обработки—последовательность операций проверки результата*. В каждом конкретном случае некоторых частей кортежа может и не быть. Например, для значительного числа задач последовательность операций проверки результата может быть постоянной, и тогда она может размещаться в составе ресурсов СВТ.

В свою очередь, последовательность операций обработки должна содержать не менее одной операции, и не пустым должно быть множество данных.

В общем случае ИЗ является ничем иным, как особого рода программой, исполняемой на СВТ в доверенной среде. Эта программа имеет свою структуру, а суть обработки данных — информационная технология обработки, определяется последовательностью операций обработки.

Для дальнейших рассуждений зафиксированную последовательность операций обработки данных будем называть ИТ-конвейером (ИТ-к).

Применение ИТ-конвейеров и их легализация

Рассмотрим теперь особенности формирования и применения ИТ-к и ИЗ на их основе.

ИТ-к определяет технологию обработки защищаемых данных. Именно исполнение несовершенных ИТ-к несет угрозу утечки данных через результаты обработки. Следовательно, ИТ-к должен быть тщательно изучен, и в случае положительного результата изучения (т. е. установлено — утечек нет) он должен быть зафиксирован, предпочтительно — регулятором.

При этом имеются объективные различия в трактовке разными регуляторами безопасности этапов информационного взаимодействия. Так, например, выше уже упоминалось, что регулятор в сфере безопасности считает безопасным перемещение защищаемых данных и их хранение в зашифрованном виде, а с точки зрения регулятора в области персональных данных, в части, их касающейся, это передача персональных данных, что может трактоваться как нарушение нормативных правовых актов (НПА).

В общем случае зона безопасности, соответствующая представлению всех регуляторов одно-

[2] В составе и на базе средств вычислительной техники (СВТ) и средств защиты информации (СЗИ).

[3] Он является предпочтительным также потому, что в этом случае не требуется дополнительная сертификация СВТ, а можно ограничиться только проверкой конкретной программы.

временно, должна быть определена. Эта зона отличается для различных наборов данных и последовательности операций обработки. Поэтому единственным вариантом "слепой" обработки защищаемых данных[4] является согласование процедур обработки со всеми регуляторами, контролирующими обработку данных того или иного содержания. Совокупность согласований ИТ-к должна быть "не уже" ограничений на использование данных. Требования НПА **ограничивают** использование данных, а **разрешения** регуляторов подтверждают возможность применения именно такой последовательности обработки, так как она не приводит к утечкам, в рамках ограничений, заданных требованиями. То есть разрешения должны быть получены от всех регуляторов, в чью область ответственности входят данные, планируемые к обработке.

Мы говорим о совместной обработке защищаемых данных, которые могут относиться к персональным данным (регулятор — РКН), к банковской тайне (регулятор — ЦБ), тайне связи (регулятор — Минцифразвития).

Объем разрешений регуляторов может не совпадать. В этом случае легальным будет набор операций, согласованный одновременно всеми регуляторами.

Таким образом, формируются требования к фиксации ИТ-конвейеров, подготовке данных и возможности обработки, а именно: итоговое ограничение ИЗ должно быть не шире пересечения всех разрешений регуляторов (см. рис.12).

Пусть П, Б и С — метки Персональных данных, Банковской тайны и тайны Связи соответственно для данных, и метки разрешений применения в согласовании ИТ-к. Обозначим через 0 — отсутствие метки, через 1 — наличие метки. Возможность применения показана в таблице, где "-" означает, что к указанным данным конвейер с указанными метками ИТ-к применять нельзя, и "+" — можно.

Отметим теперь, что чем жестче ограничения, тем легче противостоять утечкам информации. Очевидно, что для различных ролей пользователей в общем случае будут использоваться разные ИЗ. Стало быть, ИЗ должно формироваться с учетом ролей, но ИЗ вполне может не содержать указания на роль, так как это упрощает СВТ и не интересует регулятора — нет разницы, кто именно ознакомится с защищаемыми данными — на этом этапе ознакомление должно быть полностью исключено.

С другой стороны, ИЗ может содержать указание на участника с тем, чтобы обеспечить функционирование ключевой системы.

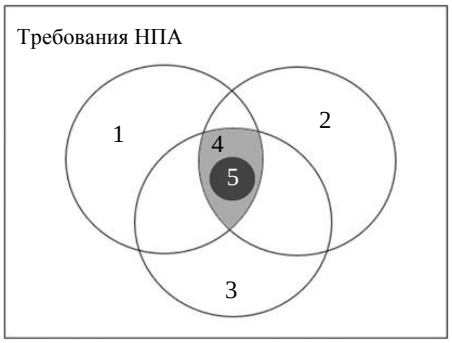

Рис. 12. Наложение разрешений регуляторов определяет область итогового разрешения:
1, 2, 3 — разрешения регуляторов;
4 — итоговое разрешение;
5 — зона действия ИЗ

Таблица

Возможность применения конвейеров к помеченным данным

Конвейеры ПБС		**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**
Данные ПБС									
000		+	+	+	+	+	+	+	+
001		-	+	-	+	-	+	-	+
010		-	-	+	+	-	-	+	+
011		-	-	-	+	-	-	-	+
100		-	-	-	-	+	+	+	+
101		-	-	-	-	-	+	-	+
110		-	-	-	-	-	-	+	+
111		-	-	-	-	-	-	-	+

[4]Напомним, это обработка, при которой оператор не имеет возможности ознакомиться с защищаемыми данными, данные хранятся в зашифрованном виде, во время обработки их извлечь невозможно (режим неизвлекаемости данных).

Таким образом, ИЗ принимает вид:

{*D*, O*d*, *K*, ИТ-к, *P*},

где *D* — данные;
O*d* — описание данных;
K — пользователь (роль, ключ, ссылка на ключ — уточняется при разработке ключевой системы);
ИТ-к — последовательность операций обработки данных;
P — последовательность операций публикации (вывода) данных.

Эта информационная структура должна быть дополнена средствами фиксации ИЗ и ИТ-к.

Средством фиксации волеизъявления ВСЕГДА выступает электронная подпись (ЭП). В нашем случае возможен следующий процесс.

1. Разработчик готовит ИТ-к и предъявляет регулятору по схеме добровольной оценки соответствия.

2. Уполномоченная структура регулятора изучает ИТ-к на предмет возможных утечек защищаемых данных по предмету регулирования.

3. В случае положительного заключения регулятор принимает решение о возможности использования ИТ-к для безопасной обработки защищаемых данных и подписывает ИТ-к своей ЭП, в сертификате ключа подписи (СКП) которой указывается класс разрешенных для обработки данных (П, Б, С) [5].

В этом случае ИТ-к может оформляться так:

(ИТ-к)[ЭП^{п}; ЭП^{б}; ЭП^{с}],

где ЭП^{п} — ЭП регулятора в области персональных данных;
ЭП^{б} — ЭП регулятора в области банковской тайны;
ЭП^{с} — ЭП регулятора в области тайны связи.

Отметим, что ИТ-к может сопровождаться не полным набором ЭП, и даже может не содержать ни одной ЭП. В этом случае решение о возможности обработки совокупности данных принимается в соответствии с таблицей.

В свою очередь, ИЗ, кроме ИТ-к, должен содержать также и другие данные, уточняющие задание. Эта подготовка выполняется вне СВТ ТСОД, и должна поступать в СВТ ТСОД в виде, позволяющем убедиться в подлинности источника ИЗ. Это может быть достигнуто применением установки ЭП в автоматическом режиме, которая в этом случае выполняет функцию защитного кода аутентификации — ЗКА. Полное ИЗ при этом выглядит так (в случае размещения сервиса публикации результатов в составе СВТ ТСОД):

{*D*, O*d*, *K*, (ИТ-к) [ЭП^{п}; ЭП^{б}; ЭП^{с}]} [ЗКА]

где *D* — ссылка на данные;
O*d* — ссылка на описание данных;
K — ссылка на ключ.

Описанное таким образом ИЗ позволяет обеспечить организационную, техническую и юридическую безопасность процессов слепой обработки защищаемых данных, и, главное, создать предпосылки для структурирования программно-аппаратного комплекса и разработки требуемой ключевой системы.

Использование технологии "слепой" обработки привлекаемых данных

Предложенные подходы к обеспечению слепой обработки данных, включая реализацию принципа неизвлекаемости данных, референсной архитектуры, структуры ИТ-конвейеров позволяют выполнять слепое машинное обучение на основе совместной обработки (привлекаемых) защищаемых данных. Результаты исследования использованы при разработке экспериментального образца ПАК "Крипто-Анклав" и АС "Анклав", которая проводилась совместно банком ВТБ (ПАО), МФТИ и ГК "Иннотех" при участии АО "ОКБ САПР".

Основная техническая задача, которая ставилась при проектировании — при высокой производительности обеспечить выполнение требований регуляторов (ФСБ, ФСТЭК, ЦБ РФ, Минцифразвития) и сформировать доверие участников.

Доверие формируется государственными документами, применением сертифицированных СЗИ НСД и СКЗИ, дополнительными мерами блокирования актуальных рисков.

Источниками рисков являются:
1. Доступ из облака.
2. Физический доступ к СВТ.
3. Утечки по техническим каналам
4. Утечки памяти.

Риски доступа из облака блокируются применением сертифицированных:

СКЗИ по классу КС3 на процессорах с архитектурой ARM;

ОС по 2-му профилю;

СЗИ НСД по 4-му профилю;

СЗИ виртуальных инфраструктур по 4-му профилю;

МЭ, криптомаршрутизаторов и однонаправленных шлюзов.

При этом разрабатываются механизмы безопасной загрузки данных.

[5] Предварительное обсуждение показывает, что регулятор вполне готов организовать такую работу, с учетом возможности организации новой необходимой гос. функции.

Риски физического доступа к СВТ блокируются механизмами контроля вскрытия, активного аудита и непрерывным контролем за возникновением НШС с защищенной фиксацией, в частности, в распределенном реестре. При этом разрабатывают механизмы управления доступом в специальных ситуациях и реакцию на проникновения.

Утечки по техническим каналам блокируются криптографическими методами, сертифицированными фильтрами и развязками.

Риски утечки памяти блокируются сертифицированными средствами защиты виртуальных инфраструктур и механизмами изоляции виртуальных машин.

Разработанный ПАК "Крипто-Анклав" подготовлен к сертификации.

Работы на экспериментальном образце показали существенный рост качества обучения моделей.

Нужно отметить, что результаты исследования выходят далеко за рамки решения конкретной задачи машинного обучения, и могут быть применены для многих задач данного класса — например, кредитного и страхового скоринга, антифрода, многих практических задач обработки привлекаемых защищаемых данных.

Заключение

В последнее время наблюдается повышенный интерес со стороны российских компаний к технологиям конфиденциальных вычислений и к платформам слепого машинного обучения, что подчеркивает актуальность решения задачи.

В работе предложены и обоснованы понятия слепой обработки данных, ИТ-к и исполняемых заданий, разработана формальная модель работы с данными в системах слепой обработки, подготовлена референсная архитектура информационной системы данного вида, рассмотрены меры по фиксации разрешенных ИТ-к и по блокированию прямых и косвенных утечек. На этой основе банк ВТБ (ПАО) совместно с МФТИ и ГК "Иннотех" разработали экспериментальный образец ПАК "Анклав", использующий ТСО и позволяющий решать широкий спектр задач в различных отраслях экономики:

- в финансовом секторе прорабатывают возможность создания национальной платформы антифрода на базе ПАК "Анклав", где важно консолидировать данные банков и телеоператоров и комплексно оценивать совершаемые финансовые операции;
- в здравоохранении накапливают, хранят и обрабатывают защищаемые данные, относящиеся к медицинской тайне (специальные персональные данные медицинского характера), где ТСО позволит создавать автоматизированные алгоритмы машинного обучения, которые станут помощниками в комплексном анализе множества факторов;
- в промышленности использование таких платформ как ПАК "Анклав" позволит создавать технологии искусственного интеллекта, в том числе на важных государственных объектах.

Литература

1. *Акаткин Ю. М., Карпов О. Э., Конявский В. А., Ясиновская Е. Д.* Цифровая экономика: концептуальная архитектура экосистемы цифровой отрасли // Бизнес-информатика. 2017. № 4(42). С. 17—28.
2. *Берестнева О. Г., Пеккер Я. С.* Выявление скрытых закономерностей в сложных системах // Известия Томского политехнического университета. 2009. Т. 315. № 5. С. 138—143.
3. *Хорев А. А.* Организация защиты конфиденциальной информации в коммерческой структуре // Защита информации. 2015. № 1. С. 14—17.
4. Федеральный закон "О персональных данных" от 27.07.2006 № 152-ФЗ [Электронный ресурс]. URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 21.11.2023).
5. Состав и содержание организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных (утв. приказом ФСТЭК России от 18 февраля 2013 г. № 21) [Электронный ресурс]: URL: https://fstec.ru/dokumenty/vse-dokumenty/prikazy/prikaz-fstek-rossii-ot-18-fevralya-2013-g-n-21 (дата обращения: 21.11.2023).
6. *Горячкин О. В.* Методы слепой обработки сигналов и их приложения в системах радиотехники и связи. — М.: Радио и связь, 2003. — 230 с.
7. *Агравал А.* Искусственный интеллект на службе бизнеса. Как машинное прогнозирование помогает принимать решения. — М.: "Манн, Иванов и Фербер", 2019. — 336 с.
8. Искусственный интеллект и машинное обучение [Электронный ресурс]. URL: https://azure.microsoft.com/ru-ru/overview/artificial-intelligence-ai-vs-machine-learning/#introduction/ (дата обращения: 20.05.2022).
9. *Морозова В. И., Логунова Д. И.* Прогнозирование методом машинного обучения // Молодой ученый. 2022. № 21(416). С. 202—204.
10. *Пилецкая А. В.* Искусственный интеллект и большие данные // Молодой ученый. 2019. № 50(288). С. 20—22.
11. *Запечников С. В.* Модели и алгоритмы конфиденциального машинного обучения // Безопасность информационных технологий. 2020. Т. 27. № 1. С. 51—67.
12. *Запечников С. В., Щербаков А. Ю.* Конфиденциальное машинное обучение с нулевым разглашением // Вестник современных цифровых технологий. 2021. № 7. С. 15—25.
13. Конституция Российской Федерации [Электронный документ]. URL: http://www.constitution.ru (дата обращения: 13.03.2023).
14. *Воронцов К. В.* Математические методы обучения по прецедентам (теория обучения машин) [Электронный ресурс]. URL: https://mathprofi.com/uploads/files/4210_f_41_lekcii-voronocva-k.v.-mashinnoe-obuchenie.pdf?key=77680f456097da8ff038c97ac64842b8/ (дата обращения: 05.01.2024).
15. Меры защиты информации в государственных информационных системах. Методический документ (утв. Фе-

деральной службой по техническому и экспортному контролю 11 февраля 2014 г.) [Электронный ресурс]. URL: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-11-fevralya-2014-g (дата обращения: 21.11.2023).

16. *Блинов А.* Российский рынок DLP-систем 2021. Проблемы и решения. Ч. 1 [Электронный ресурс]. URL: https://ict-online.ru/analytics/obzor-rossiyskiy-rynok-dlp-sistem-2021-problemy-i-resheniya-chast-1-osobennosti-sovremennogo-14787 (дата обращения: 21.11.2023).

17. *Блинов А.* Российский рынок DLP-систем 2021. Проблемы и решения. Ч. 2 [Электронный ресурс]. URL: https://ict-online.ru/analytics/obzor-rossiyskiy-rynok-dlp-sistem-2021-problemy-i-resheniya-chast-2-gramotnoe-vnedrenie-13062 (дата обращения: 21.11.2023).

18. Федеральный закон "Об электронной подписи" от 06.04.2011 № 63-ФЗ (последняя редакция) [Электронный ресурс]. URL: https://www.consultant.ru/document/cons_doc_LAW_112701/ (дата обращения: 21.11.2023).

19. *Конявский В. А.* Идентификация и применение ЭЦП в компьютерных системах информационного общества // Безопасность информационных технологий. 2010. № 3. С. 6—13.

20. *Аветисян А. И.* Разработка доверенных систем. Искусственный интеллект. Доклад на XXVII науч.-практ. конф. "Комплексная защита информации" [Электронный ресурс]. URL: https://kzi.su/files/files/materials2022/Avetisyn.pdf (дата обращения: 21.11.2023).

21. *Артамонов В. А., Артамонова Е. В., Сафонов А. Е.* Безопасность искусственного интеллекта // Защита информации. 2022. № 6. С. 2—11.

22. *Конявская С. В.* Защита банкомата согласно Закону о КИИ: как избежать уязвимости традиционной архитектуры // Расчеты и операционная работа в коммерческом банке. 2020. № 1(155). С. 13—25.

23. *Агафьин С., Смирнов П., Смышляев С.* Неизвлекаемые ключи в облаке — путь через криптопровайдер // BIS Journal. 2017. № 3(26). С. 48—50.

24. *Конявский В. А.* Смысл и безопасность // Информационная безопасность. 2020. № 5. С. 54—56.

25. Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий (утв. приказом ФСТЭК России от 2 июня 2020 г. № 76). Выписка [Электронный ресурс]. URL: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/trebovaniya-po-bezopasnosti-informatsii-utverzhdeny-prikazom-fstek-rossii-ot-2-iyunya-2020-g-n-76 (дата обращения: 21.11.2023).

26. *Конявский В. А., Медведев В. В., Росс Г. В.* Защищенные информационные технологии в цифровой экономике // Вопросы защиты информации. 2022. № 2(137). С. 34—44.

27. *Конявский В. А., Конявская С. В.* Доверенные информационные технологии: от архитектуры к системам и средствам. — М.: URSS, 2019. — 264 с.

Technology of "blind" processing of attracted data in machine learning systems

[1, 2, 3] *V. A. Konyavskiy*, [1, 2] *S. V. Konyavskya-Schastnaya*, [3] *G. V. Ross*, [1, 7, 8, 9] *A. M. Raigorodskii*, [1] *S. A. Trenin*, [1, 4] *A. V. Leonidov*, [1, 4] *E. E. Vasilyeva*, [1, 4, 5] *S. B. Vasilyev*, [6] *M. Yu. Konovalyhin*

[1] Moscow Institute of Physics and Technology (National Research University), Moscow Region, Dolgoprudny, Russia; [2] SC "OKB SAPR", Moscow, Russia; [3] Plekhanov Russian Economic University, Moscow, Russia; [4] P. N. Lebedev Physical Institute of the Russian Academy of Sciences, Moscow, Russia; [5] National Research University "Higher School of Economics", Moscow, Russia; [6] VTB Bank (PJSC), St. Petersburg, Russia; [7] Lomonosov Moscow State University, Moscow, Russia; [8] Caucasus Mathematical Center, Adyghe State University, Republic of Adygea, Maykop, Russia; [9] Banzarov Buryat State University, Ulan-Ude, Buryat Republic, Russia

The article is devoted to resolving the contradiction that has arisen to date between the need to process combined data when building predictive models and the lack of technical solutions to ensure a sufficient level of information security. This created a contradiction between the need to process combined data and the lack of technical solutions to ensure a sufficient level of information security. This article is devoted to resolving this contradiction. The concepts of "blind" data processing, unretrievable data and models, and attracted data are introduced. Examples of indirect leaks are given, the occurrence of which is usually associated with differential attacks. To increase the level of data security, it is proposed to limit requests to the machine learning system during "blind" data processing only to proven command sequences that do not lead to leaks of data that is to be protected — IT pipelines. The formal model of "blind" data processing, the reference architecture of an automated "blind" data processing system, the structure and measures for fixing and ensuring the integrity and legality of the use of IT pipelines, features of their formation and application are proposed. Based on the results obtained, the "Crypto-Enclave" software and hardware complex was created. The results can also be used in creating a national antifraud system, in medicine, and industry.

Keywords: machine learning, security, "blind" data processing, confidential machine learning, IT Pipeline, data combining, data that is to be protected, unretrievable data, direct leaks, indirect leaks, differential attacks, confidential computing, national antifraud operator.

Bibliography — 27 references. *Received March 27, 2024*