(19) Евразийское патентное ведомство

(11) 047223 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.06.21

(21) Номер заявки
202393324

(22) Дата подачи заявки
2023.12.19

(51) Int. Cl. ***G06F 21/00*** (2013.01)
G06F 21/62 (2013.01)
G06N 3/08 (2023.01)

(54) СПОСОБ И СИСТЕМА ДЛЯ ГЕНЕРАЦИИ СИНТЕТИЧЕСКИХ ДАННЫХ

(31) **2023126703**
(32) **2023.10.18**
(33) **RU**
(43) **2024.06.20**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Кочетков Сергей Борисович, Поздняков Илья Николаевич, Фаттахова Юлдуз Зуфаровна, Руднев Александр Сергеевич, Кочетков Максим Дмитриевич, Хабибуллина Ляйсан Наилевна (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-B1-11720709
US-B2-10713384
US-A1-20220180234
WO-A1-2023080994

(57) Настоящее изобретение относится к области искусственного интеллекта, и в частности к способу и системе для генерации синтетических данных. Техническим результатом, достигаемым при выполнении вышеуказанной цели, является обеспечение возможности генерации синтетических данных, структура которых соответствует оригинальным данным, за счет использования первичных и внешних ключей, связывающих информацию о субъектах данных и набор данных. Указанный технический результат достигается благодаря осуществлению способа генерации синтетических данных, выполняемого по меньшей мере одним вычислительным устройством, содержащего этапы, на которых получают оригинальный сэмпл данных, содержащий информацию о субъектах данных, и по меньшей мере один набор данных, связанный с упомянутыми субъектами данных; классифицируют информацию, содержащуюся в сэмпле данных, для выявления атрибутов, относящихся к чувствительным данным (ЧД), и атрибутов, не относящихся ЧД; обучают по меньшей мере один генератор синтетических данных с использованием алгоритмов машинного обучения на основе сэмпла данных; значений классов ЧД, полученных на предыдущем этапе; метаданных о данных для обучения и первичных и внешних ключей, связывающих информацию о субъектах данных и набор данных; генерируют посредством обученного генератора по меньшей мере один набор синтетических данных, в которых набор атрибутов субъектов данных отличается от наборов атрибутов субъектов данных, содержащихся в оригинальном сэмпле данных, с сохранением первичных и внешних и ключей, содержащих информацию о субъектах данных.

047223 B1

047223 B1

Область техники

Настоящее изобретение относится к области искусственного интеллекта и, в частности, к способу и системе для генерации синтетических данных.

Синтетические данные - это искусственные данные, имитирующие наблюдения реального мира и используемые для подготовки моделей машинного обучения, когда получение реальных данных невозможно из-за сложности или дороговизны или когда реальных данных недостаточно.

Уровень техники

В последние годы стала очень актуальной проблема утечки персональных данных, так как все больше информации хранится и обрабатывается в цифровом виде. Это может произойти по разным причинам, включая хакерские атаки, уязвимости в системах безопасности, небрежное отношение с данными.

Последствия утечки персональных данных могут быть очень серьезными. Люди могут стать жертвами мошенничества, кражи личности, распространения ложной информации. Кроме того, утечка информации может нанести существенный ущерб репутации компании или организации.

Поэтому защита персональных данных является важной задачей для компаний, организаций и даже отдельных лиц.

На данный момент наиболее популярное решение защиты персональных данных - это обезличивание данных. Методы обезличивания помогают защищать персональные данные (данные, которые явно или косвенно определяют субъект персональных данных), используя различные процессы изменения или удаления информации.

В уровне техники также известна анонимизация, как один из видов обезличивания, при этом методе идентифицирующие атрибуты (имена, адреса, номера телефонов и другая личная информация) заменяются на псевдонимы или удаляются из наборов данных.

Также существуют различные техники маскировки данных, которые могут быть применены для обезличивания. Например, хэширование, токенизация, шифрование и т.д.

Еще одним из способов защиты конфиденциальных данных являются инструменты ручной генерации синтетических данных на основе справочников и словарей. Такие решения сохраняют смысловую структуру реальных данных, при этом защищая клиентские данные.

Помимо защиты данных инструменты генерации синтетических данных решают задачи машинного обучения, когда возникает необходимость в достаточном количестве валидных данных.

Технология генерации синтетических данных является одним из способов быстрого получения выборок больших объемов с сохранением свойств и взаимосвязей оригинальных данных. При использовании синтетических данных оригинальные персональные данные заменяются сгенерированными данными, которые имитируют статистические характеристики и общую структуру реальных данных, но не содержат прямой идентифицирующей информации. Они могут быть использованы в тестировании, разработке и анализе без риска утечки персональных данных.

Для реализации генерации синтетических данных существует множество инструментов и библиотек, таких как, например, Faker.

Существенным недостатком известных решений в области ручной генерации синтетических данных является трудоемкость и трудозатратность, особенно при работе с большим объемом данных и обеспечении разнообразия, реалистичности данных; низкая эффективность в выявлении чувствительных данных; также ручная генерация может привести к внесению человеческой предвзятости в данные при анализе входных данных или непредсказуемости при создании синтетических данных. Сгенерированные вручную данные могут быть сложными для поддержки и обновления в дальнейшем, если изменятся требования к данным.

В области обезличивания данных - это отсутствие валидного датасета для решения задач машинного обучения, нарушение целостности при рандомизации и применения других методов обезличивания, зависимость от исходных данных, возможность восстановления оригинальных данных из обезличенных, также в процессе обезличивания данных могут возникать потери информации или искажение статистических свойств данных.

Наиболее близким к представленному решению является система облачных вычислений, раскрывая в заявке US 20200012933 A1, опубл. 09.01.2020. В данном документе раскрывается генератор наборов данных системы облачных вычислений, выполненный с возможностью генерации синтетических наборов данных для обучения модели данных. Вычислительные ресурсы могут обучать модель данных с помощью синтетического набора данных. Оптимизатор модели может хранить модель данных и метаданные модели данных в хранилище моделей. Система облачных вычислений может получать производственные данные из источника данных производственным экземпляром системы облачных вычислений с использованием общей файловой системы. Производственные данные могут быть обработаны производственным экземпляром с использованием модели данных. Вычислительные ресурсы, генератор наборов данных и оптимизатор модели могут размещаться в отдельных виртуальных вычислительных экземплярах системы облачных вычислений. Недостатком известного решения является отсутствие возможности сохранения взаимосвязей в исходных данных между данными, относящихся к одному наборы данных, например, к одному объекту. Также в известном решении отсутствуют средства, позволяющие оценить

схожесть оригинальных и синтетических данных.

Раскрытие изобретения

Технической проблемой или задачей, поставленной в данном изобретении, является создание нового эффективного, простого и надежного сервиса генерации синтетических данных.

Техническим результатом, достигаемым при выполнении вышеуказанной цели, является обеспечение возможности генерации синтетических данных, структура которых соответствует оригинальным данным, за счет использования первичных и внешних ключей, связывающих информацию о субъектах данных и набор данных.

Указанный технический результат достигается благодаря осуществлению способа генерации синтетических данных, выполняемого по меньшей мере одним вычислительным устройством, содержащего этапы, на которых

получают оригинальный сэмпл данных, содержащий информацию о субъектах данных, и по меньшей мере один набор данных, связанный с упомянутыми субъектами данных;

классифицируют информацию, содержащуюся в сэмпле данных, для выявления атрибутов, относящихся к чувствительным данным (ЧД) и атрибутов, не относящихся ЧД;

обучают по меньшей мере один генератор синтетических данных с использованием алгоритмов машинного обучения на основе: сэмпла данных; значений классов ЧД, полученных на предыдущем этапе; метаданных о данных для обучения и первичных и внешних ключей, связывающих информацию о субъектах данных и набор данных;

генерируют посредством обученного генератора по меньшей мере один набор синтетических данных, в которых набор атрибутов субъектов данных отличается от наборов атрибутов субъектов данных, содержащихся в оригинальном сэмпле данных, с сохранением первичных и внешних и ключей, содержащих информацию о субъектах данных.

В одном из частных примеров осуществления способа дополнительно выполняют этап, на котором перекладывают синтетические данные или генератор синтетических данных по меньше мере в один внутренний или внешний контур. В другом частном примере осуществления способа метаданные данных для обучения: характеризуют структуру данных, количество структур данных, типы данных для обучения, структуру БД.

В другом частном примере осуществления способа генератор синтетических данных состоит из генератора ЧД, и генератора данных, не относящийся к ЧД, причем этап обучения по меньшей мере одного генератора синтетических данных содержит этапы, на которых

извлекают из оригинального сэмпла данных атрибуты, имеющие класс ЧД, и метаданные, относящиеся к заданным атрибутам, не имеющие класс ЧД;

обучают генератор ЧД на основе извлеченных на предыдущем этапе данных;

извлекают из оригинального сэмпла данных атрибуты, не имеющие класса ЧД, и метаданные, относящиеся к заданным атрибутам, не имеющие класс ЧД;

обучают генератор данных, не относящихся к ЧД, на основе извлеченных на предыдущем этапе данных.

В другом частном примере осуществления способа генератор синтетических данных состоит из генератора данных, не относящихся к ЧД, и генератора ЧД, причем этап генерации синтетических данных содержит этапы, на которых

направляют команду для генерации данных в генератор данных, не относящихся к ЧД, который при получении упомянутой команды выполняет генерацию атрибутов, не относящихся к ЧД, а также метаданные об атрибутах ЧД, представленных в оригинальном сэмпле данных;

направляют команду в генератор ЧД, который при получении упомянутой команды для генерации атрибутов, относящихся к ЧД, обращается к памяти, которой он оснащен, в которой сохранены наборы атрибутов, относящихся к ЧД, и из упомянутого набора атрибутов, с учетом сгенерированных на предыдущем этапе метаданных о атрибутах ЧД, случайным образом выбирает атрибуты, которые будут выданы генератором ЧД в качестве сгенерированных атрибутов, относящихся к ЧД.

объединяют атрибуты, сгенерированные генератором ЧД, с атрибутами, сгенерированными генератором данных, не относящихся к ЧД, по меньшей мере в один набор данных, в соответствии с метаданными о данных для обучения, характеризующими структуру данных, количество структур данных, типы данных для обучения, структуру БД.

В другом частном примере осуществления способа дополнительно выполняют этапы для оценки качества генератора синтетических данных, применяемых для решения задач машинного обучения по предсказанию заданной пользователем целевой переменной, на которых:

разделяют оригинальный сэмпл данных на тренировочный сэмпл данных и тестовый сэмпл данных;

обучают новый генератор синтетических данных, описанным в п.1 способом на тренировочном сэмпле данных;

генерируют тренировочные синтетические данные с помощью нового генератора тренировочного синтетического сэмпла;

обучают на тренировочном сэмпле данных первую модель с помощью алгоритмов машинного обу-

чения для прогнозирования целевой переменной;

обучают на тренировочном синтетическом сэмпле данных вторую модель прогнозирования целевой переменной;

прогнозируют на основе тестового сэмпла данных при помощи первой модели первый набор значений целевой переменной;

прогнозируют на основе тестового сэмпла данных при помощи второй модели второй набор значений целевой переменной;

на основе спрогнозированных первого и второго наборов целевых переменных и фактического значения целевых переменных, содержащихся в тестовом сэмпле данных, определяют метрику качества синтетических данных для данной задачи машинного обучения (ML).

В другом частном примере осуществления способа дополнительно выполняют этапы для оценки безопасности сгенерированных синтетических данных, на которых:

извлекают из синтетического сэмпла данных по каждому субъекту данных набор атрибутов, относящиеся к ЧД, причем типы атрибутов, подлежащих включению в набор, заданы разработчиком;

сравнивают извлеченные данные с наборами атрибутов для каждого субъекта данных, содержащихся в оригинальном сэмпле данных;

на основе результатов сравнения определяют метрику безопасности синтетических данных.

В другом частном примере осуществления способа дополнительно выполняют этапы для оценки схожести данных, на которых:

извлекают из оригинального сэмпла данных атрибуты, не имеющие класса ЧД;

определяют статистическое распределение атрибутов, извлеченных на предыдущем этапе;

извлекают из синтетических данных атрибуты, не относящиеся к ЧД;

определяют статистическое распределение упомянутых атрибутов, извлеченных на предыдущем этапе;

на основе полученных статистический распределений атрибутов, извлеченных из оригинального сэмпла данных и из синтетических данных, определяют метрику схожести данных.

В другом частном примере осуществления способа дополнительно выполняют этапы для оценки качества сгенерированных синтетических данных, на которых:

определяют инвертированный коэффициент силуэта между синтетическими и оригинальными данными и коэффициент воспроизведения корреляционных взаимосвязей для расчета метрик схожести данных.

В другом предпочтительном варианте осуществления заявленного решения представлена система генерации синтетических данных, содержащая по меньшей мере одно вычислительное устройство и по меньшей мере одну память, содержащую машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют вышеуказанный способ.

Краткое описание чертежей

Признаки и преимущества настоящего изобретения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей, на которых:

На фиг. 1 представлена общая схема взаимодействия элементов системы интеллектуальной обработки данных.

На фиг. 2 представлен процесс подключения к хранилищу источника и сбор метаданных

На фиг. 3 представлен процесс классификации персональных данных

На фиг. 4 представлен процесс обучения данных

На фиг. 5 представлена работа модуля оценки качества

На фиг. 6 представлен процесс генерации синтетических данных

Осуществление изобретения

Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройств хранения данных. В роли устройства хранения данных могут выступать, но не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

База данных (БД) - совокупность данных, организованных в соответствии с концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, причем такое собра-

ние данных, которое поддерживает одну или более областей применения (ISO/IEC 2382:2015, 2121423 "database").

Сигнал - материальное воплощение сообщения для использования при передаче, переработке и хранении информации.

Логический элемент - элемент, осуществляющий определенные логические зависимости между входными и выходными сигналами. Логические элементы обычно используются для построения логических схем вычислительных машин, дискретных схем автоматического контроля и управления. Для всех видов логических элементов, независимо от их физической природы, характерны дискретные значения входных и выходных сигналов.

Субъект данных - это идентифицированное или поддающееся идентификации лицо или объект.

В настоящем решении под термином "генерация синтетических данных" здесь и далее по тексту будет пониматься процесс создания новых данных, которые имеют схожие характеристики с реальными данными, выполняемый с помощью машинного обучения. Под синтетическими данными следует понимать класс данных, которые генерируются искусственно, т.е. не получены в результате прямых наблюдений и фиксаций реальных событий (реальных данных) и не являются их модификацией. Каждая синтетическая запись - это новое наблюдение, не существующее в реальном мире, тем не менее повторяющее свойства и природу реальных данных. Именно за счет того, что синтетические наблюдения - это новые объекты данных и достигается конфиденциальность реальных данных.

Важно иметь ввиду, что система представлена как наглядный вариант осуществления. Целью этого описания является исключительно оказание помощи в понимании, а не определение объема и границ настоящего решения.

В соответствии со схемой, представленной на фиг. 1, заявленная система (1) для генерации синтетических данных состоит из следующих элементов:

хранилище источника входных данных (далее хранилище источника) (11),

устройство (12) маршрутизации данных,

устройство (13) ввода данных, которое позволяет пользователю взаимодействовать с системой (1),

модуль (14) классификации чувствительных данных (далее ЧД),

модуль ML (15),

модуль (16) оценки качества данных,

хранилище (17) приемника выходных данных (далее хранилище приемника),

хранилище (18) данных и моделей,

перекладчик (19) данных и моделей (далее перекладчик).

Например, хранилище (11) источника данных может быть представлено в виде памяти, содержащей информацию или данные, которые используются или могут использоваться для получения информации. Данные в источнике могут быть сохранены в виде файлов разных форматов: -csv, excel или в виде реляционной базы данных таких как PostgreSQL, MySQL, Oracle, в виде базы данных, предназначенной для работы с большими данными (Big Data) Hadoop.

Устройство (12) может быть реализовано на базе вычислительного устройства с использованием языка программирования Python и других.

Устройство (12) работает в связке с устройством ввода (13), отвечающим за отображение информации на клиентской стороне.

Python - высокоуровневый язык программирования общего назначения с динамической строгой типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости кода и его качества, а также на обеспечение переносимости написанных на нём программ.

Устройство ввода (13) - интерфейс, с которым взаимодействует пользователь посредством устройства пользователя.

Модуль (14) классификации ЧД может быть реализован на базе вычислительного устройства, исполняющего программный код, написанный на языке программирования Python, содержащий описание алгоритма обработки данных и выполнения модулем (14) приписанных ниже функций.

Программный код представляет собой представляет собой набор строк, написанных на языке программирования.

Модуль ML (15) может быть реализован на базе вычислительного устройства, исполняющего программный код, написанный на языке программирования Python, содержащий описание алгоритма обработки данных и выполнения модулем (15) приписанных ниже функций.

Модуль (16) оценки качества данных может быть реализован на базе вычислительного устройства, исполняющего программный код, написанный на языке Python, содержащий алгоритмы машинного обучения.

Хранилище (17) приемника выходных данных (далее хранилище приемника) может быть реализовано как база данных в виде Hadoop (база данных (database), предназначенная для работы с большими данными), hive (система управления базами данных на основе платформы Hadoop), файлового хранилища данных и любой другой базы данных.

Хранилище (18) данных и моделей может быть представлено в виде файлового хранилища данных.

Перекладчик (19) данных и моделей (далее перекладчик (19)) может быть реализован на базе системы хранения результатов работы устройства (12) (далее - артефакты).

В соответствии со схемой, представленной на фиг. 2, пользователь вручную задает информацию об источнике данных, в котором хранятся данные для обучения, путем ввода параметров подключения в соответствующие поля через устройство ввода (13), например, указывает адрес (url), логин и пароль подключения к хранилищу источника. Также данные для обучения могут быть переданы пользователем через устройство ввода (13) в виде файлов, содержащие таблицы данных. Пользователь может установить связи между таблицами, указав первичные и внешние ключи через устройство ввода (13) (см. рисунок 1). Данные для обучения могут содержать, например, информацию о субъектах данных, в частности ФИО, номер документа, дату рождения и пр., которые могут быть представлены в первой таблице, и по меньшей мере один набор данных, связанный с упомянутыми субъектами данных, например, данные о транзакциях лиц, который может быть представлен во второй таблице. Также данные для обучения могут содержать заданные для каждого субъекта данных фактические значения целевой переменной, предназначенные для обучения моделей прогнозирования и проверки качества синтетических данных для данной задачи машинного обучения.

В частности, данные для обучения могут представлять:

первую таблицу "Клиенты", которая содержит информацию о клиентах: ФИО, паспортные данные, дата рождения и т.д., также первичный ключ "clientjd" -уникальный идентификатор каждого клиента в таблице "Клиенты"; каждая запись в этой таблице имеет уникальное значение "clientjd", которое является первичным ключом;

Таблица "Клиенты"

client_id	FIO	Document	BirthDate
1	Иванов Сергей Петрович	8011 111111	01.01.1990
2	Петров Сергей Иванович	8011 222222	02.02.1990

вторую таблицу "Кредитные продукты", содержащую информацию о кредитных продуктах клиентов с информацией по кредиту, суммой и задолженностью; каждый кредитный продукт обладает уникальным ключом id и ссылается на клиента, уникальный идентификатор которого хранится в колонке client_id; чтобы установить связь между кредитными продуктами и клиентами используется внешний ключ "clientjd"; фактические значения целевой переменной (колонка IsDebt) могут указывать, например, на то, что клиент 1 выплатил кредит, а клиент 2 не выплатил кредит к дате окончания кредитного договора;

Таблица "Кредитные продукты"

id	BeginDate	EndDate	FullSum	DebtSum	IsDebt	client_id
1	25.08.2022	25.08.2023	50000.00	00.00	0	1
2	10.09.2022	10.09.2023	30000.00	1287.89	1	2
3	11.10.2022	11.10.2023	100000.00	12000.00	1	2

"clientjd" в таблице "Кредитные продукты" является внешним ключом, ссылающимся на первичный ключ "clientjd" в таблице "Клиенты". Каждая запись о кредитных продуктах в таблице "Кредитные продукты" связана с соответствующим клиентом из таблицы "Клиенты" через внешний ключ. Таким образом, внешний ключ "clientjd" в таблице "Кредитные продукты" ссылается на первичный ключ "clientjd" в таблице "Клиенты".

Соответственно, через устройство ввода (13) пользовательская информация о подключении (url) поступает в устройство (12) маршрутизации данных, в котором выполняется обработка информации, в частности, устройство (12) подключается к хранилищу источника (11) и собирает-информацию о первичных и внешних ключах и метаданные о данных для обучения, хранимых в источнике (11), которые могут характеризовать структуру данных (например, таблица), количество структур данных (информация о количестве таблиц, столбцов), типы данных для обучения, структуру БД. Полученные данные для обучения, метаданные и информация о первичных и внешних ключах и другое, из хранилища источника (11) сохраняются в хранилище (18) данных и моделей.

В соответствии со схемой, представленной на фиг. 3, устройство (12) для получения входных данных (далее оригинальных сэмплов) для классификации ЧД запускает сбор данных из хранилища (11) в объеме, необходимом для запуска классификации ЧД (14), в соответствии с полученной командой, например, от оператора системы (1), либо при вызове устройством 12.

Устройство (12) забирает оригинальные сэмплы данных из хранилища источника (11) и размещает оригинальные сэмплы данных в хранилище (18) данных и моделей. Далее устройство (12) направляет команду в модуль (14) классификации ЧД для классификации информации, содержащейся в оригинальном сэмпле данных, для выявления чувствительной и не чувствительной информации. Для классификации информации модуль 14 может быть оснащен нейронной сетью или любым другим алгоритмом машинного обучения, заранее обученного на размеченной выборке данных. Соответственно, информация,

содержащаяся в оригинальном сэмпле данных, подается на вход нейронной сети или другого алгоритма машинного обучения, а на выходе модуль 14 выдает значения класса ЧД для информации, причем значение класса ЧД присваивается каждому атрибуту и указывает на то, относятся ли данные к ЧД или не относятся к ЧД. Например, значение класса ЧД может быть: "Фамилия", "Имя", "Отчество", "Номер ДУЛ", "Электронная почта", "Номер телефона", "ИНН" и т.д.

Пример:

surname_ col	Класс ПДн	name_ col	Класс ПДн	middlename_ col	Класс ПДн	phone_col	Класс ПДн
Иванов	Фамилия	Иван	Имя	Иванович	Отчество	791712312312	Телефон
Петров		Петр		Петрович		791723423423	

Также чувствительной информацией, помимо персональных данных, может быть различная конфиденциальная информация, содержащая коммерческую тайну, банковскую тайну, служебная информация ограниченного распространения, данные платежных карт.

Алгоритмы классификации ЧД могут использовать, как и подходы, основанные на правилах (например, регулярные выражения, которые используют последовательности специальных символов, формирующих паттерн или шаблон, который сопоставляется со строкой), так и различные алгоритмы классификации для форматно-логического контроля и выявления класса персональных данных, так и оба подхода совместно. Например, алгоритм определения электронной почты может использовать регулярное выражение следующего вида:

r'[\w.+-]+@[\w-]+\ [\w.-]+'.

Если большинство значений в колонке удовлетворяет данному регулярному выражению, то вся колонка будет размечена как электронная почта. Атрибуты, которые промаркировались как персональные данные (чувствительные), например, "Фамилия", "Имя", "Отчество", "Номер ДУЛ", "Электронная почта", "Номер телефона", "ИНН", будут в дальнейшем генерироваться в соответствии с общей природой оригинальных данных, используя справочники, собранные для конкретного класса персональных данных. По завершению классификации ЧД, сформированный системой и утвержденный пользователем отчет классификации ЧД, и значения классов данных, указывающие на то, являются ли данные чувствительными данными, сохраняется в хранилище (18) данных и моделей. Далее система (1) может перейти к этапу обучения модели, описания алгоритма, который хранится в модуле ML (15). Для запуска процесса обучения в соответствии со схемой, представленной на фиг. 4, устройство (12), в соответствии с полученной командой, например, от оператора системы (1) или устройства 12, запускает процесс извлечения данных из хранилища (11) источника в объеме, необходимом для обучения, который сохраняется в хранилище (18) данных и моделей для запуска процедуры обучения. В частности, входными данными для запуска модуля ML (15) являются, оригинальные сэмплы данных для обучения, первичные и внешние ключи, значения классов ЧД и метаданные из хранилища (18) данных и модулей. Далее устройство (12) извлекает из хранилища (18) оригинальные сэмплы данных, содержащие информацию по каждому субъекту данных, набор данных, связанный с упомянутым субъектом (например, таблицы), значения классов ЧД, метаданные и первичные и внешние ключи и передает извлеченные данные в модуль ML (15) для обучения известными методами генератора синтетических данных, состоящего из модуля генерации ЧД (генератор ЧД) и модуля генерации данных, не относящихся к ЧД (генератор данных, не относящихся к ЧД), реализованных на базе нейронных сетей и других классических алгоритмов машинного обучения. Для обучения модуля генерации ЧД модуль ML (15) извлекает из оригинального сэмпла данных атрибуты, имеющие класс ЧД, и вместе с заданными разработчиком метаданными, относящимися к атрибутам, не имеющими класс ЧД (например, пол, регион и др), и первичными и внешними ключами подает их на вход модуля генерации ЧД для его обучения известными методами. Соответственно, для обучения модуля генерации данных, не относящихся к ЧД, модуль ML (15) извлекает из оригинального сэмпла данных атрибуты, не имеющие класса ЧД, и вместе с метаданными, относящимися к заданным (например, разработчиком) атрибутам, не имеющие класс ЧД (например, пол, регион и др) и первичными и внешними ключами подает их на вход упомянутого модуля генерации для его обучения.

По завершению процедуры обучения модуль ML (15) уведомляет устройство (12) о завершении задачи. В свою очередь устройство (12) сохраняет артефакты обучения в перекладчике (19), в частности информация о параметрах генератора, например, значения весовых коэффициентов генератора, и метаинформации (информации о структуре синтетической базы данных).

Далее в соответствии со схемой, представленной на фиг. 5, устройство (12) запускает процесс оценки качества модели, передав ссылки на обученный генератор синтетических данных и на оригинальный сэмпл данные, на основе которых был обучен генератор.

Для оценки качества обученной модели модуль (16) оценки качества из хранилища (18) данных и моделей извлекает оригинальные сэмплы данных, на основе которых был обучен генератор, а также обученный генератор, посредством которой упомянутый модуль (16) производит пробную генерацию синтетических данных, в частности синтетического сэмпла данных, для расчета метрик качества алгоритма

генерации синтетических данных и сгенерированной синтетики. В частности, модуль (16) осуществляет запуск генератора синтетических данных, который при его запуске генерирует синтетическую информацию о субъектах данных и по меньшей мере один набор синтетических данных, связанный с упомянутыми субъектами данных, а также первичные и внешние ключи, причем для осуществления генерации данных генератор отдельно производит генерацию не ЧД с метаданными о ЧД и отдельно производит генерацию ЧД. В частности, упомянутый генератор направляет команду для генерации данных в модуль генерации данных, не относящихся к ЧД, который при получении упомянутой команды выполняет генерацию атрибутов, не относящихся к ЧД, а также метаданные об атрибутах ЧД, представленных в оригинальном сэмпле данных. Соответственно, для генерации ЧД генератор направляет команду в модуль генерации ЧД, который при получении упомянутой команды для генерации атрибутов, относящихся к ЧД, обращается к памяти, которой он оснащен, в которой сохранены наборы атрибутов, относящихся к ЧД (например, списки ФИО и пр.) и из упомянутого набора атрибутов, с учетом сгенерированных на предыдущем этапе метаданных об атрибутах ЧД, случайным образом выбирает атрибуты, которые будут выданы модулем генерации ЧД в качестве сгенерированных атрибутов, относящихся к ЧД.

Далее генератор синтетических данных объединяет атрибуты, сгенерированные генератором ЧД, с атрибутами, сгенерированными генератором данных, не относящихся к ЧД, по меньшей мере в один набор данных, в соответствии с метаданными о данных для обучения, характеризующими структуру данных, количество структур данных, типы данных для обучения, структуру БД. После того, как синтетический сэмпл данных сгенерирован, модуль (16) оценки качества данных начинает расчет метрик (метрики качества, метрики безопасности, метрики схожести) на основе анализа данных из оригинального сэмпла данных и синтетического сэмпла данных.

В случае если синтетические данные предполагаются использовать для решения некоторой задачи машинного обучения (кредитный скоринг, прогнозирование стоимости недвижимости и тд) рассчитываются метрики качества моделирования на синтетических данных. Метрики качества на синтетических данных для задач машинного обучения подразумевают расчет метрик качества оценки ML алгоритмов по поставленной задаче (кредитный скоринг, прогнозирование стоимости недвижимости и тд). Среди наиболее популярных методов оценки качества можно выделить следующие ROC-AUC, F1, Precision, Recall, MSE, RMSE и др.

Для этого пользователь передает в модуль 16 данные о наименовании колонки с фактическими значениями (т.е. с фактическими значениями целевой переменной) для решения задачи машинного обучения (кредитный скоринг, прогнозирование стоимости недвижимости и тд). Далее модуль (16) оценки качества данных разделяет оригинальный сэмпл данных (оригинальный датасет) случайным образом в соответствии с распределением фактической целевой переменной или с заданным разработчиком алгоритмом на две части: тренировочный сэмпл данных и тестовый сэмпл данных. Например, если сэмпл данных содержит записи о 100 субъектах данных, то он может быть разделен в пропорции, например, 70 или 80 записей для тренировочного сэмпла данных и 30 или 20 записей для тестового сэмпла данных. Далее тренировочный сэмпл данных направляется модулем (16) в модуль ML (15) для обучения нового генератора синтетических данных для последующего тестирования качества сгенерированных синтетических данных.

Данные нового генератора направляются модулем ML (15) в модуль (16), после чего с помощью данного генератора модулем (16) производится генерация тренировочного синтетического сэмпла (датасета) по объему, равному тренировочному сэмплу данных. Далее модуль (16) использует тренировочный сэмпл данных, для обучения модели прогнозирования целевой переменной на поставленной задаче машинного обучения (кредитный скоринг, прогнозирование стоимости недвижимости и тд) и в результате получает первую модель, обученную на оригинальных тренировочных данных. Далее модуль (16) использует синтетический сэмпл данных для обучения модели прогнозирования целевой переменной на поставленной задаче машинного обучения (кредитный скоринг, прогнозирование стоимости недвижимости и тд) и в результате получает вторую модель, обученную на синтетических данных.

В результате получаются две модели ML: первая модель ML, обученная на оригинальном тренировочном сэмпле данных и вторая модель ML, обученная на синтетическом сэмпле данных. Далее модуль (16) направляет на входы первой и второй модели тестовый оригинальный сэмпл данных для получения спрогнозированных значений целевых переменных первой модели и спрогнозированных значений целевых переменных второй модели для решения поставленной задачи машинного обучения, после чего на основе полученных спрогнозированных значений и фактических значений целевых переменных модуль (16) осуществляет определение значения метрики качества моделей ML, например, посредством алгоритмов оценки качества: ROC-AUC, F1, Precision, Recall, MSE, RMSE и др.

Например, спрогнозированные значения целевых переменных первой модели и спрогнозированные значения целевых переменных второй модели, полученные по итогу обработки тестового сэмпла данных указывают на то, что субъекты данных выплатят кредит или не выплатят. Соответственно, модуль (16) сравнивает спрогнозированные значения целевых переменных первой модели и спрогнозированные значения целевых переменных второй модели с фактическими значениями целевых переменных из тестового оригинального сэмпла данных -значениями выплат клиентов, и рассчитывает метрики качества моде-

лей, значения которых характеризуют количество совпадений спрогнозированных моделей с тестовыми фактическими целевыми переменными. Далее модуль 16 переходит к этапу сравнения полученных метрик качеств.

Если метрика качества второй модели (модели обученной на синтетических данных) незначительно ниже, чем метрика первой модели (модели обученной на оригинальных данных), т.е. значение отклонения метрики качества второй модели от метрики качества первой модели находится в диапазоне допустимых значений, то синтетические данные (в частности, синтетический сэмпл данных) считаются валидными, то есть они повторяют все взаимосвязи исходных данных (т.е. сэмпла данных) так, что позволяют обучать синтетические модели с незначительным снижением качества предсказаний на реальных данных.

Для определения метрики безопасности модуль (16) оценки качества данных выполняет расчет количества полнострочных пересечений между данными из сэмпла данных и из синтетического сэмпла данными, когда каждая строка оригинальных данных сравнивается с каждой строкой синтетических данных. Если все значения в строке оригинальных данных точно совпадают со значениями в строке синтетических данных, то это будет считаться полнострочным пересечением. На примере показан частный случай, когда полнострочных пересечений нет, что является успешным результатом.

Глобальные построчные совпадения	0.0%

Пересечения в данных из результатов профилирования

(Имя, Фамилия, Отчество) + КИ	Пересечения
(first_name, last_name, middle_name) + inn2	0
(first_name, last_name, middle_name) + tk	0
(first_name, last_name, middle_name) + pan	0
(first_name, last_name, middle_name) + passport	0
(first_name, last_name, middle_name) + kpp	0
(first_name, last_name, middle_name) + inn1	0
(first_name, last_name, middle_name) + phone	0
(first_name, last_name, middle_name) + ssn	0

Оригинальные данные:

pass_col	surname_col	gender_col	birthdate_col	date_col	price_col
4508 458526	Иванов	м	05 07.1990	05 07.2021	900 000
4524 374323	Петрова	ж	06.08.1992	06.05.2020	700 000
3802 192392	Андреев	м	14 12.1993	24 04.2019	500 000
3602 523123	Ильина	ж	07.07.1987	04.01.2021	1 000 000

Синтетические данные:

pass_col	surname_col	gender_col	birthdate_col	date_col	price_col
4508 458526	Васильева	ж	05 07.1992	04 04.2020	720 000
4524 374323	Постов	м	04.04.1990	06.07.2021	910 000
3802 192392	Симонова	ж	12 07.1989	07 02.2021	1 100 000
3602 523123	Госов	м	15.11.1994	26.04.2019	502 000

Расчет количества пересечений между данными из оригинальными сэмпла и синтетического сэмпла данных по кортежам классов персональных данных, которые могут однозначно идентифицировать субъект данных (например, ФИО+ИНН, ФИО+паспорт, и другие).

Например, есть пересечения: Паспорта и ФИО пересекаются, то есть совпадают в оригинальных и синтетических данных, таким образом идентифицируют субъект.

Оригинальные данные:

pass_col	fio_col	gender_col	birthdate_col
4508 458526	Иванов Иван Петрович	м	05.07.1990
4524 374323	Петрова Ирина Олеговна	ж	06.08.1992
3802 192392	Андреев Егор Иванович	м	14.12.1993
3602 523123	Ильина Ольга Петровна	ж	07.07.1987

Синтетические данные:

pass_col	fio_col	gender_col	birthdate_col
4508 458526	Иванов Иван Петрович	м	11.12.1989
4524 374323	Петрова Ирина Олеговна	ж	18.01.1993
3802 192392	Андреев Егор Иванович	м	25.10.1990
3602 523123	Ильина Ольга Петровна	ж	17.09.1995

В частности, модуль (16) извлекает из синтетического сэмпла данных набор атрибутов, относящиеся к ЧД, заданного разработчиком типа субъекта данных, и сравнивает его с наборами атрибутов каждого субъекта данных, содержащихся в сэмпле данных, в соответствии с типами атрибутов. Например, извлеченный модулем (16) набор упомянутых атрибутов может содержать номер документа "4508 458526", фамилия "Иванов", имя "Иван" отчество "Петрович", которые будут сравниваться с номерами документов и ФИО субъектов данных, информация о которых содержатся в оригинальном сэмпле данных. Соответственно, аналогичным образом модуль (16) сравнивает наборы атрибутов, относящихся к ЧД, каждого субъекта данных из синтетического сэмпла данных с наборами атрибутами каждого субъекта данных, содержащихся в оригинальном сэмпле данных.

Если извлеченный из синтетических данных набор атрибутов, относящиеся к ЧД, совпал с набором атрибутов по меньшей мере одного субъекта данных, содержащегося в сэмпле оригинальных данных, то модуль (16) назначает синтетическим данным метрику безопасности, указывающую на то, что синтетические данные не прошли проверку на безопасность, после чего блокирует доступ к генератору синтетических данных, посредством которого были получены эти данные, или удаляет данные о генераторе из хранилища (18). Если извлеченный набор атрибутов, относящиеся к ЧД, не совпал с наборами атрибутов субъектов данных, содержащимися в оригинальном сэмпле данных, то модуль (16) назначает синтетическим данным метрику безопасности, указывающую на то, что синтетические данные соответствуют требованиям безопасности.

Метрики схожести данных включают в себя:

визуальные тесты на соответствие распределений признаков оригинальных и синтетических данных;

статистические тесты на равенство распределений в разрезе атрибутов (KS, CS -тесты);

коэффициенты корреляционной схожести данных, показывающие насколько корреляции исходных данных сохранены в синтетических данных;

инвертированные метрики силуэта, показывающие насколько кластер синтетических данных, совпадает с кластером оригинальных данных в пространстве.

Другие тесты.

В частности, модуль (16) извлекает из оригинальных сэмплов данных атрибуты, не относящиеся к ЧД, например, значения возрастов субъектов данных, после чего строит статистическое распределение упомянутых атрибутов для получения информации о статистическом распределении атрибутов, в частности значения интервалов и соответствующих им частот. Например, информация о статистическом распределении атрибутов может указывать на то, что в оригинальном сэмпле данных большинство клиентов в возрасте от 50 до 70 лет (см. фиг. 2).

Далее модуль (16) извлекает из синтетического сэмпла данных атрибуты, не относящиеся к ЧД, аналогичного типа, например, значения возрастов субъектов данных, после чего также выполняет расчет статистического распределения упомянутых атрибутов для получения информации о статистическом распределении атрибутов значения интервалов и соответствующих им частот. Полученная на основе сэмпла данных информация о статистическом распределении атрибутов сравнивается модулем (16) с информацией о статистическом распределении атрибутов, полученной на основе синтетического сэмпла данных для определения того, что упомянутые атрибуты, содержащиеся в оригинальном сэмпле данных, и атрибуты этого же типа, содержащиеся в синтетическом сэмпле данных, соответствуют одному и тому же закону распределения. Сравнение упомянутой информации может осуществляться, например, посредством использования теста Колмогорова-Смирнова. Например, если модулем (16) определено, что пиковый возраст субъектов данных в оригинальном сэмпле данных и синтетическом сэмпле данных соответствует распределению от 50 до 70 лет и распределение частот по другим возрастам также находится в заданном оригинальными данными распределении, то модуль (16) назначает синтетическим данным максимальную метрику схожести. Если модулем (16) определено, что пиковый возраст субъектов данных в синтетическом сэмпле данных не соответствует распределению оригинальных данных в любом из промежутков, например, от 50 до 70 лет, то модуль (16) назначает метрику качества в зависимости от отклонения информации атрибутов, содержащихся в синтетического сэмпле данных, от атрибутов, содержащихся в оригинальном сэмпле данных.

В отчете генерации важным результатом является информация о качественных метриках сгенерированных данных, которые представлены в как коэффициент силуэта, который может быть частично определен известным методом, раскрытым, например, в статье по ссылке:

https://en.wikipedia.org/wiki/Silhouette_(clustering),

и коэффициент воспроизведения корреляционных связей, который может быть частично определен известным методом, раскрытым, например, в статье по ссылке:

(https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D1%8F).

Коэффициент силуэта показывает, насколько синтетические данные похожи на оригинальные данные. Наилучшим значением будет являться показатель в 100%. Методика расчета: оригинальным данным проставляется значение label=0, синтетическим - label=1. Оба набора конкатенируются и рассчитывается инвертированный коэффициент силуэта, (инвертируем чтобы пользователю было легче понимать успешность/неуспешность генерации данных). Коэффициент силуэта показывает, насколько каждый объект "похож" на другие объекты в том кластере, в который он был распределён в процессе кластеризации, и "не похож" на объекты из других кластеров. Чем больше объекты синтетического кластера похожи на объекты оригинального кластера - тем лучше, значит синтетика не отличима от оригинальных данных в пространстве. (Рисунок 3).

В частности, модуль (16) извлекает данные о субъектах данных из оригинального сэмпла данных и данные о субъектах данных из синтетического сэмпла данных и известным методом сравнивает их для получения коэффициента силуэта. Коэффициент силуэта вычисляется с помощью среднего внутрикластерного расстояния (a) и среднего расстояния до ближайшего кластера (b) по каждому образцу. Силуэт вычисляется как (b-a)/max(a, b). Пояснение: b - это расстояние между a и ближайшим кластером, в который a не входит.

Коэффициента воспроизведения корреляционных связей показывает, насколько корреляции в исходных данных сохранены. Коэффициент корреляции характеризует величину отражающую степень взаимосвязи двух переменных между собой. Он может варьировать в пределах от -1 (отрицательная корреляция) до +1 (положительная корреляция). Если коэффициент корреляции равен 0 то, это говорит об отсутствии корреляционных связей между переменными.

Методика расчета: по всем парам колонок рассчитываются корреляции в синтетическом и исходном датасетах. Абсолютные значения разности по всем парам усредняются, а затем инвертируются. Наилучшим значением будет считаться величина в 100%.

Пример расчета корреляционной матрицы.

Необходимо построить матрицу корреляций на оригинальных и синтетических данных: заголовками строк и столбцов будут обрабатываемые атрибуты, а на пересечении строк и столбцов выводятся коэффициенты корреляции для соответствующей пары признаков:

	Сумма	Количество товара	Время на работу
Сумма	1.00	0.82	-0.22
Количество товара	0.82	1.00	
Время на работу	-0.22		1.00

Например, корреляция между "суммой" и "количеством товара" составляет 0.82, что указывает на сильную положительную корреляцию. Чем больше количество товара, тем больше сумма.

А корреляция между "суммой, потраченной на товар" и "временем, потраченным на работу" составляет -0.22, что указывает на более слабую корреляцию, чем в предыдущем примере. Чем больше сумма, тем меньше времени, потраченного на работу.

Качественные метрики.

Метрика	Значение
Коэффициент силуэта	79.83%
Коэффициент воспроизведения корреляционных связей	99.89%

Соответственно, модуль (16) извлекает данные о субъектах данных из сэмпла данных и данные о субъектах данных из синтетического сэмпла данных и известным методом сравнивает их для получения коэффициента воспроизведения корреляционных связей. В результате этапа генерации сформированный архив с синтетическими данными и отчетом качества передается пользователю в необходимом формате в целевое хранилище данных.

По завершению процесса определения метрик модуль (16) формирует отчет о метриках, который сохраняется вместе с обученным генератором в хранилище (18) данных и моделей.

Устройство (12) забирает и сохраняет обученный генератор и сформированный отчет в перекладчик (19).

Следует отметить, что данные примеры являются одними из метрик оценки.

Список метрик может расширяться.

Как представлено на фиг. 6.

Для генерации синтетических данных пользователь инициирует генерацию синтетических данных

через устройство (13) ввода данных указав параметры генерации (объем генерируемых данных и информацию о подключении к хранилищу приемника).

Устройство (13) передает информацию в устройство (12).

Устройство (12) забирает артефакты обучения (обученная модель и отчет) из перекладчика (19).

Устройство (12) передает ссылки на артефакты обучения в модуль ML (15), который генерирует необходимое количество синтетических записей, заданных пользователем.

Сэмплы сгенерированных данных сохраняются в хранилище данных и моделей (18).

Устройство (12) забирает сгенерированные данные.

Устройство (12) загружает синтетические данные пользователю в хранилище приемника (17)

Как было описано раннее для обучения модели на вход алгоритмам машинного обучения пользователем подаются отчет классификации ЧД и наборы данных чувствительной и нечувствительной информации. Запускается процесс обучения, где модель обучается согласно алгоритму обучения, а также автоматически подбираются алгоритмы генерации данных.

В результате обучения формируются обученные модели и отчет, в котором содержатся результаты обучения с метриками качества. В сгенерированных данных сохраняются взаимосвязи между наборами данных в виде первичных и внешних ключей.

В связях между таблицами применяются концепции первичного ключа и внешнего ключа, которые обеспечивают целостность данных и устанавливают связи между таблицами. Поэтому пользователи могут указать связи (или это может определяться автоматически при сборе метаинформации из хранилища источника), как внешний ключ в одной таблице ссылается на запись в другой таблице. Связи между таблицами позволяют организовать данные, сделать их более структурированными и логически связанными. Информация о ключах помогает алгоритмам обучения генератора найти межтабличные взаимосвязи, и воспроизвести их в синтетических данных. Это позволяет представлять сложные структуры данных и устанавливать связи для эффективного извлечения и обработки информации.

Как представлено на фиг. 4, в соответствии с вышеописанным выделяются два сценария генерации синтетических данных относительно сегментов сети: обучение и генерация синтетических данных в одном контуре, обучение и генерация синтетических данных в разных контурах с передачей синтетических данных или модели синтетических данных в другой контур.

На фиг. 4 представлен процесс обучения модели в защищенном контуре, генерация синтетических данных в незащищенном контуре с передачей обученной модели из одного контура в другой (6).

Классификация ЧД, обучение модели, согласно алгоритму обучения, происходит во внутреннем контуре (1). В результате формируются обученные модели и отчет, в котором содержатся результаты обучения с метриками качества. Артефакты этапа обучения помещаются перекладчик (2). Перекладчик передает артефакты этапа обучения из внутреннего контура во внешний контур (здесь важно отметить, что никакие исходные данные, содержащие какие-либо конфиденциальные данные не передаются). Пользователь во внешнем контуре выбирает обученную модель и заказывает синтетические данные в нужном объеме в целевое хранилище пользователя (3).

В некоторых вариантах осуществления, способ включает обучение модели и генерацию синтетических данных в одном контуре без потребности перекладки между сегментами. В рамках такого сценария возможно обучение и генерация модели во внутреннем контуре либо во внешнем контуре: модель обучается согласно алгоритму обучения, формируются артефакты этапа обучения: обученные модели и отчет, в котором содержатся значения метрик качества. Далее обученную модель можно использовать для генерации синтетических данных. Поскольку генерация синтетических данных будет выполняться в том же контуре, где было обучение модели, перекладка артефактов этапа обучения между сегментами не потребуется. Обученная модель загружается из хранилища для последующей генерации. Пользователь выбирает необходимую модель из списка обученных им моделей и заказывает синтетические данные в нужном объеме с последующей перекладкой в целевое хранилище пользователя (см. фиг. 5). В общем виде (см. фиг. 7) вычислительное устройство содержит объединенные общей шиной информационного обмена один или несколько процессоров (101), средства памяти, такие как ОЗУ (102) и ПЗУ (103), интерфейсы ввода/вывода (104), устройства ввода/вывода (105), и устройство для сетевого взаимодействия (106).

Процессор (101) (или несколько процессоров, многоядерный процессор и т.п.) может выбираться из ассортимента устройств, широко применяемых в настоящее время, например, таких производителей, как: Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. Под процессором или одним из используемых процессоров в устройстве (100) также необходимо учитывать графический процессор, например, GPU NVIDIA или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа, а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах. ОЗУ (102) представляет собой оперативную память и предназначено для хранения исполняемых процессором (101) машиночитаемых инструкций для выполнение необходимых операций по логической обработке данных. ОЗУ (102), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом, в качестве ОЗУ (102) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (103) представляет собой одно или более устройств постоянного хранения данных, например, жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др. Для организации работы компонентов устройства (100) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (104). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

Для обеспечения взаимодействия пользователя с устройством (100) применяются различные средства (105) В/В информации, например, клавиатура, дисплей (монитор), сенсорный дисплей, тач-пад, джойстик, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п. Средство сетевого взаимодействия (106) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например, Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (206) может использоваться, но не ограничиваться: Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др,

Дополнительно могут применяться также средства спутниковой навигации в составе устройства (100), например, GPS, ГЛОНАСС, BeiDou, Galileo. Конкретный выбор элементов устройства (100) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

Модификации и улучшения вышеописанных вариантов осуществления настоящего технического решения будут ясны специалистам в данной области техники. Предшествующее описание представлено только в качестве примера и не несет никаких ограничений. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ генерации синтетических данных, выполняемый по меньшей мере одним вычислительным устройством, содержащий этапы, на которых

получают оригинальный сэмпл данных, содержащий информацию о субъектах данных, и по меньшей мере один набор данных, связанный с упомянутыми субъектами данных;

классифицируют информацию, содержащуюся в сэмпле данных, для выявления атрибутов, относящихся к чувствительным данным (ЧД), и атрибутов, не относящихся ЧД;

обучают по меньшей мере один генератор синтетических данных с использованием алгоритмов машинного обучения на основе сэмпла данных; значений классов ЧД, полученных на предыдущем этапе; метаданных о данных для обучения и первичных и внешних ключей, связывающих информацию о субъектах данных и набор данных;

генерируют посредством обученного генератора по меньшей мере один набор синтетических данных, в которых набор атрибутов субъектов данных отличается от наборов атрибутов субъектов данных, содержащихся в оригинальном сэмпле данных, с сохранением первичных и внешних и ключей, содержащих информацию о субъектах данных.

2. Способ по п.1, характеризующийся тем, что дополнительно выполняют этап, на котором перекладывают синтетические данные или генератор синтетических данных по меньше мере в один внутренний или внешний контур.

3. Способ по п.1, характеризующийся тем, что метаданные данных для обучения характеризуют структуру данных, количество структур данных, типы данных для обучения, структуру БД.

4. Способ по п.1, характеризующийся тем, что генератор синтетических данных состоит из генератора ЧД и генератора данных, не относящихся к ЧД, причем этап обучения по меньшей мере одного генератора синтетических данных содержит этапы, на которых

извлекают из оригинального сэмпла данных атрибуты, имеющие класс ЧД, и метаданные, относящиеся к заданным атрибутам, не имеющие класс ЧД;

обучают генератор ЧД на основе извлеченных на предыдущем этапе данных;

извлекают из оригинального сэмпла данных атрибуты, не имеющие класса ЧД, и метаданные, относящиеся к заданным атрибутам, не имеющие класс ЧД;

обучают генератор данных, не относящихся к ЧД, на основе извлеченных на предыдущем этапе данных.

5. Способ по п.1, характеризующийся тем, что генератор синтетических данных состоит из генератора данных, не относящихся к ЧД, и генератора ЧД, причем этап генерации синтетических данных содержит этапы, на которых

направляют команду для генерации данных в генератор данных, не относящихся к ЧД, который при

получении упомянутой команды выполняет генерацию атрибутов, не относящихся к ЧД, а также метаданные об атрибутах ЧД, представленных в оригинальном сэмпле данных;

направляют команду в генератор ЧД, который при получении упомянутой команды для генерации атрибутов, относящихся к ЧД, обращается к памяти, которой он оснащен, в которой сохранены наборы атрибутов, относящихся к ЧД, и из упомянутого набора атрибутов, с учетом сгенерированных на предыдущем этапе метаданных о атрибутах ЧД, случайным образом выбирает атрибуты, которые будут выданы генератором ЧД в качестве сгенерированных атрибутов, относящихся к ЧД;

объединяют атрибуты, сгенерированные генератором ЧД, с атрибутами, сгенерированными генератором данных, не относящихся к ЧД, по меньшей мере в один набор данных, в соответствии с метаданными о данных для обучения, характеризующими структуру данных, количество структур данных, типы данных для обучения, структуру БД.

6. Способ по п.1, характеризующийся тем, что дополнительно выполняют этапы для оценки качества генератора синтетических данных, применяемых для решения задач машинного обучения по предсказанию заданной пользователем целевой переменной, на которых

разделяют оригинальный сэмпл данных на тренировочный сэмпл данных и тестовый сэмпл данных;

обучают новый генератор синтетических данных описанным в п.1 способом на тренировочном сэмпле данных;

генерируют тренировочные синтетические данные с помощью нового генератора тренировочного синтетического сэмпла;

обучают на тренировочном сэмпле данных первую модель с помощью алгоритмов машинного обучения для прогнозирования целевой переменной;

обучают на тренировочном синтетическом сэмпле данных вторую модель прогнозирования целевой переменной;

прогнозируют на основе тестового сэмпла данных при помощи первой модели первый набор значений целевой переменной;

прогнозируют на основе тестового сэмпла данных при помощи второй модели второй набор значений целевой переменной;

на основе спрогнозированных первого и второго наборов целевых переменных и фактического значения целевых переменных, содержащихся в тестовом сэмпле данных, определяют метрику качества синтетических данных для данной задачи машинного обучения (ML).

7. Способ по п.1, характеризующийся тем, что дополнительно выполняют этапы для оценки безопасности сгенерированных синтетических данных, на которых

извлекают из синтетического сэмпла данных по каждому субъекту данных набор атрибутов, относящиеся к ЧД, причем типы атрибутов, подлежащих включению в набор, заданы разработчиком;

сравнивают извлеченные данные с наборами атрибутов для каждого субъекта данных, содержащихся в оригинальном сэмпле данных;

на основе результатов сравнения определяют метрику безопасности синтетических данных.

8. Способ по п.1, характеризующийся тем, что дополнительно выполняют этапы для оценки схожести данных, на которых

извлекают из оригинального сэмпла данных атрибуты, не имеющие класса ЧД;

определяют статистическое распределение атрибутов, извлеченных на предыдущем этапе;

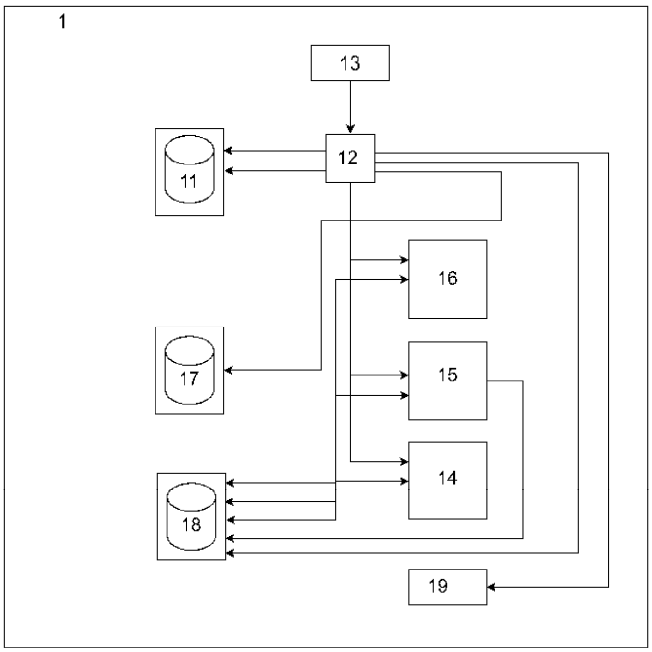
извлекают из синтетических данных атрибуты, не относящиеся к ЧД;

определяют статистическое распределение упомянутых атрибутов, извлеченных на предыдущем этапе;

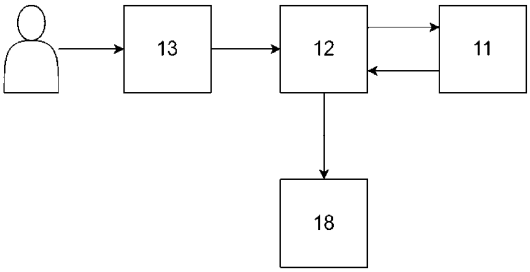
на основе полученных статистических распределений атрибутов, извлеченных из оригинального сэмпла данных и из синтетических данных, определяют метрику схожести данных.

9. Способ по п.1, характеризующийся тем, что дополнительно выполняют этапы для оценки качества сгенерированных синтетических данных, на которых определяют инвертированный коэффициент силуэта между синтетическими и оригинальными данными и коэффициент воспроизведения корреляционных взаимосвязей для расчета метрик схожести данных.

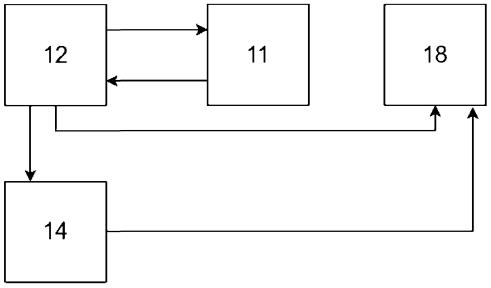
10. Система генерации синтетических данных, содержащая по меньшей мере одно вычислительное устройство и по меньшей мере одно устройство памяти, содержащее машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют способ по любому из пп.1-9.



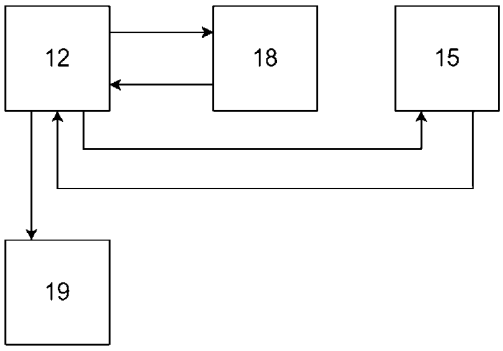
Фиг. 1



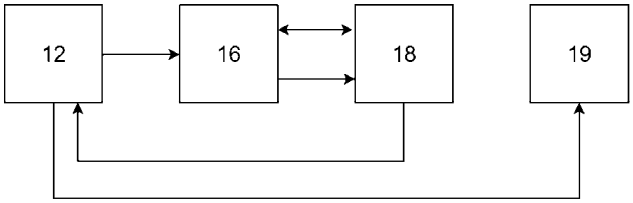
Фиг. 2



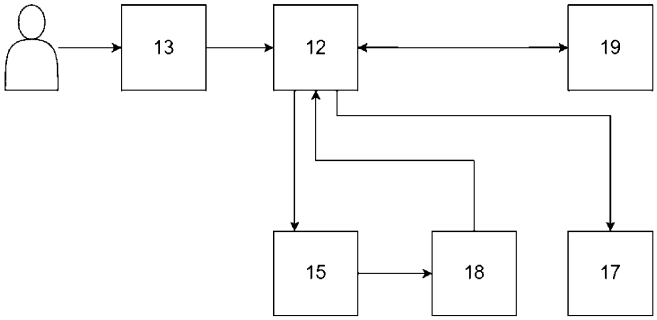
Фиг. 3



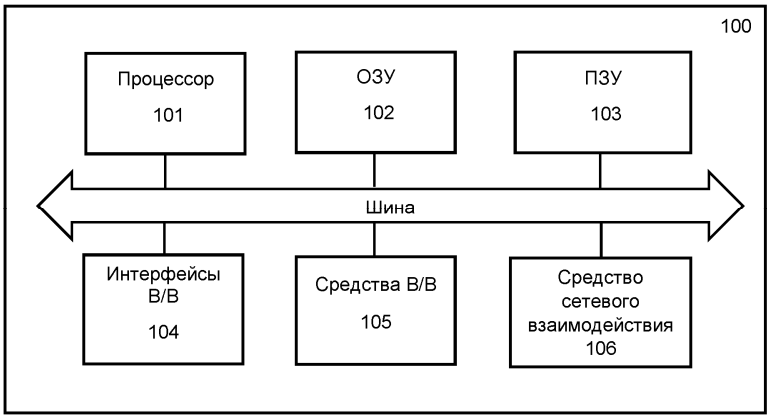
Фиг. 4



Фиг. 5



Фиг. 6



Фиг. 7

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2