



Республика Таджикистан

(19) **TJ** (11) **1601**
(51) **МПК** G06N 3/04, G06N 3/08,
G06N 3/082, G06N 3/088, G06F 17/18

ГОСУДАРСТВЕННОЕ ПАТЕНТНОЕ ВЕДОМСТВО

(12) **Описание изобретения** К МАЛОМУ ПАТЕНТУ

1

(21) 2502087

(22) 14.04.2025

(46) Бюл. 219, 2025

(71)(73) Компания «ЗЕХН.АЙ» ЛТД («ZENH.AI» Ltd).
(TJ)

(72) Азиззон Азими (TJ)

(54) **СПОСОБ ГЕНЕРАЦИИ РЕАЛИСТИЧНЫХ
ТАБЛИЧНЫХ ДАННЫХ С ВЫБРОСАМИ НА
ОСНОВЕ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ
СЕТИ zGAN**

(56) 1. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *arXiv preprint arXiv:1907.00503*. Доступно по ссылке: <https://arxiv.org/abs/1907.00503>.

2. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

3. Laurens de Haan and Ana Ferreira. Extreme Value Theory. Springer New York, 2006.

4. Zheng, H., Liu, Y., Ji, M., Wu, F., & Fang, L. (2015). Learning High-level Prior with Convolutional Neural Networks for Semantic Segmentation. *arXiv preprint arXiv: 1511.06988*. <https://arxiv.org/abs/1511.06988>

5. Kristina Bluwstein, Marcus Buckmann, Andreas Joseph, Sujit Kapadia, and Ozgur Simsek. Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 145:103773, 2023.

6. M. Ghil, P. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. W. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin. Extreme events: dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3):295-350, 2011.

7. Sunday Adeola Ajagbe and Matthew O. Adigun. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multi-media Tools and Applications*, 83:5893-5927, 1 2024.

8. Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein.

2

SAINT: improved neural networks for tabular data via row

attention and contrastive pre-training. *CoRR*, abs/2106.01342, 2021.

9. Blaz Skrlj, Saso Dzeroski, Nada Lavrac, and Matej Petkovic. Feature importance estimation with self-attention networks. *CoRR*, abs/2002.04464, 2020.

10. Penghui Zhao, Zhongjun Ding, Yang Li, Xiaohan Zhang, Yuanqi Zhao, Hongjun Wang, and Yang Yang. Sgad-gan: Simultaneous generation and anomaly detection for time-series sensor data with generative adversarial networks. *Mechanical Systems and Signal Processing*, 210:111141, 3 2024.

(57) Изобретение относится к области искусственного интеллекта, в частности к способу генерации реалистичных табличных данных с выбросами с использованием генеративно-состязательной сети zGAN, включающей генератор, дискриминатор и модуль генерации ковариаций (covGEN). Генератор инициализируют случайным шумом, обучают дискриминатор различать синтетические и реальные данные, а генератор — создавать реалистичные данные с применением метода обратного распространения ошибки. Для генерации выбросов вычисляют ковариационные матрицы на основе реальных данных через модуль covGEN и интегрируют их в синтетические данные в соответствии с заданными распределениями (нормальное, Лапласовское, Вейбулловское, Гумбелевское, Леви). Приватность данных обеспечивается фильтром правдоподобия с использованием хеширования.

Ковариационные матрицы могут формироваться с использованием условного вариационного автоэнкодера (cVAE) для улучшения соответствия структуры данных априорному распределению. Предусмотрена оценка качества данных с помощью метрик (например, AUC) и их использование для обучения моделей машинного обучения.

Способ обеспечивает создание приватных, реалистичных и структурированных синтетических данных с возможностью целенаправленного формирования выбросов..

Изобретение относится к области искусственного интеллекта, а именно к генеративно-состязательным сетям (GAN), для генерации реалистичных синтетических табличных данных с выбросами. Выбросы в задаваемых признаках табличных данных могут применяться для моделирования редких или сложных событий, таких как экономические кризисы или пандемии.

Известен способ CopulaGAN [1], который комбинирует в своей архитектуре возможности классических копул и принципы игрового обучения характерные для GAN в целом [2]. По назначению CopulaGAN является аналогом zGAN.

CopulaGAN — это генеративно-состязательная сеть, предназначенная для синтеза табличных данных с сохранением зависимостей между переменными. Её ключевая особенность заключается в использовании классического статистического инструмента, известного как копулы, для моделирования сложных взаимосвязей между признаками. Копулы позволяют разделять совместное распределение данных на независимые маргинальные распределения и функцию зависимости, что способствует восстановлению корреляций в сложных многомерных данных.

При обучении CopulaGAN генератор создаёт синтетические данные, а дискриминатор пытается отличить их от реальных. В то же время, копулы используются для обеспечения точного воспроизведения статистических зависимостей, которые характерны для табличных данных. Гибридный подход позволяет улучшить реализм сгенерированных данных, особенно в случаях, когда данные содержат смешанные типы признаков.

CopulaGAN активно используется для задач синтеза табличных данных в финансовой и медицинской сферах, где важно сохранять взаимосвязи между признаками для анализа и моделирования. Несмотря на перечисленные достоинства, модель уступает другим архитектурам GAN, таким как CTGAN и zGAN, в точности сохранения корреляций, возможности целенаправленной генерации данных с выбросами, а также воспроизведении сложных нелинейных зависимостей.

Несмотря на свою эффективность в сохранении зависимостей, CopulaGAN имеет ряд ограничений. Одним из таких ограничений является сложность моделирования высокоразмерных данных или данных с большим числом категориальных признаков, что может приводить к ухудшению качества генерируемых данных. Кроме того, использование копул накладывает ограничения на типы зависимостей, которые могут быть воспроизведены, что требует выбора подходящей копулы для каждой конкретной задачи, что в свою очередь может быть сопряжено с высокими вычислительными затратами.

Известен способ TVAE [1], который по назначению является аналогом zGAN. Данная генеративная модель также предназначена для генерации табличных данных, однако основана на принципе действия вариационного автоэнкодера, без использования принципа действия характерного для GAN.

TVAE (Tabular Variational AutoEncoder) — это

генеративная модель, предназначенная для создания синтетических табличных данных. Основой модели является вариационный автоэнкодер (VAE), который обучается формировать представление данных в латентном пространстве и восстанавливать их с минимальной потерей информации. Такой подход позволяет способу TVAE моделировать сложные вероятностные зависимости в данных, делая её особенно эффективной для табличных наборов с разными типами признаков.

В отличие от генеративно-состязательных сетей, TVAE строится на основе принципа максимизации доказательного нижнего порога (ELBO), что делает процесс её обучения более стабильным. Модель способна учитывать категориальные и числовые признаки, а также справляться с отсутствующими данными путём их вероятностного восстановления. Латентное пространство позволяет упростить структуру данных, делая процесс генерации более эффективным.

Способ TVAE имеет свои ограничения. Из-за вероятностного характера генерации модель может хуже сохранять точные взаимосвязи между признаками, особенно в данных с высокой корреляцией между признаками или наличием сложных зависимостей. Кроме того, качество генерируемых данных может снижаться при наличии большого числа категориальных признаков или при дисбалансе классов в исходных данных.

TVAE находит применение в задачах, где важен баланс между реализмом и стабильностью генерации, таких как анализ конфиденциальных данных в медицинских или финансовых приложениях. Данный способ уступает специализированным архитектурам GAN, таким как CTGAN и zGAN, в точности воспроизведения корреляций и способности генерации данных с выбросами.

Наиболее близким к заявленному техническому решению по технической сущности и достигаемому техническому результату является CTGAN [3], который является прототипом zGAN. Концептуально архитектура CTGAN приближена к части архитектуры zGAN отвечающей за обучение и генерацию синтетических табличных данных.

CTGAN (Conditional Tabular Generative Adversarial Network) — это генерати

вно состязательная сеть, разработанная специально для генерации синтетических табличных данных, включающих как числовые, так и категориальные признаки. Основной особенностью прототипа является использование механизма условной генерации, который позволяет модели учитывать распределения категориальных переменных при синтезе новых данных. Это делает прототип особенно эффективным в задачах, где данные имеют смешанные типы признаков со сложными зависимостями между ними.

Способ CTGAN включает специализированные компоненты для обработки табличных данных, такие как Gumbel-Softmax для категориальных признаков и методы перестройки дискретных данных для точного моделирования их распределений. Инициализируемый случайным шумом генератор прототипа создаёт синтетические данные, а дискриминатор обучается отличать их от реальных. Условная генерация позво-

ляет модели фокусироваться на специфических подмножествах данных, сохраняя редкие категории и сложные зависимости между признаками.

При этом CTGAN не лишен ограничений. Модель может испытывать трудности при работе с большими наборами данных или с наборами данных с большим числом уникальных категорий. Кроме того, для достижения оптимального качества синтетических данных требуется тщательная настройка гиперпараметров, а стабильность обучения может зависеть от характеристик исходного набора данных.

CTGAN широко используется в финансовых, медицинских и других индустриях, где важно сохранять корреляции между признаками данных. Несмотря на свою эффективность, прототип уступает специализированной модели zGAN, в задачах генерации данных с выбросами, но остаётся эталоном для большинства сценариев синтеза табличных данных.

zGAN существенно отличается как от прототипа, так и от аналогов наличием возможности целенаправленной генерации синтетических выбросов в соответствии с теорией экстремальных значений [4]. Такая возможность позволяет моделировать редкие экономические события, обнаруживать и обрабатывать выбросы, что способствует повышению показателей качества моделей машинного обучения в различных прикладных областях [5, 6, 7].

Целенаправленная генерация синтетических выбросов в zGAN осуществляется с применением синтетических или вычисляемых по реальным данным ковариационных матриц. При этом существует возможность генерации как синтетических выбросов, так целевых переменных, соответствующих уже сгенерированным выбросам. Другим отличием zGAN от прототипа и аналога является наличие фильтра правдоподобия, который позволяет zGAN получать повышенную приватность генерируемых данных без сохранения клиентских реальных данных.

Технической задачей способа является генерация реалистичных табличных данных, с сохранением в сгенерированных данных корреляций между признаками присущих реальным обучающим данным, а также обеспечение возможности целенаправленной генерации синтетических выбросов для повышения предсказуемости редких событий и стабильности моделей машинного обучения.

Поставленная задача решена следующим образом: применяется архитектура zGAN, которая обладает существенными отличиями от указанного прототипа и аналогов, а именно наличием модуля вычисления ковариаций (covGEN) на основе реальных данных, хешированного фильтра правдоподобия, условного вариационного автоэнкодера (cVAE) и модулей, использующихся для генерации синтетической целевой переменной для строк данных со сгенерированными синтетическими выбросами.

Модуль covGEN позволяет генерировать синтетические выбросы в выбранных признаках с использованием задаваемых законов распределений. Фильтр правдоподобия, работающий с применением алгоритмов хеширования, позволяет сохранять приватность генерируемых данных благодаря отсут-

Оценка качества синтетических данных осуществляется по показателю AUC на примере решения задачи бинарной классификации.

Краткое описание чертежей

Фиг. 1 иллюстрирует структуру zGAN

Фиг. 2 иллюстрирует структуру описывающую последовательность действий в поставленных экспериментах

Фиг. 3 иллюстрирует разность коэффициентов корреляций между синтетическими и реальными данными набора данных Титаник

Краткое описание таблиц

Таблица 1 содержит показатели приватных наборов данных

Таблица 2 содержит медианные значения AUC полученные в эксперименте «out-of-sample»

Таблица 3 содержит медианные значения AUC при «out-of-time» эксперименте с zGAN на 50% наборе данных

Таблица 4 содержит медианные значения AUC при «out-of-time» эксперименте с CTGAN на 50% наборе данных

Таблица 5 содержит медианные значения AUC для различных процентов выбросов в макроданных для набора данных A9 в «out-of-time» эксперименте

Генератор генеративно-состязательной сети zGAN инициализируется случайным шумом, формируя синтетические данные на выходе. Дискриминатор получает на вход как сгенерированные синтетические, так и реальные данные и выдает двоичный код, соответствующий классам «реальные» или «синтетические» данные.

Процесс обучения дискриминатора заключается в улучшении его способности различать синтетические данные от реальных, в то время как генератор, наоборот, обучается создавать данные неотличимые от реальных.

Приватность синтетических данных обеспечивается с помощью фильтра правдоподобия, который использует хеши реальных данных для фильтрации сгенерированных синтетических данных, похожих на реальные клиентские данные. Подход с использованием хеширования гарантирует, что реальные клиентские данные не сохраняются после обучения модели.

zGAN, обученная на реальных табличных данных, генерирует синтетические данные, неотличимые от реальных, в которые интегрируются макровыбросы. Модуль, который интегрирует макровыбросы в синтетические данные, обозначается символом «+».

Макровыбросы синтезируются генератором условной ковариации (covGEN) путем выборки из многомерного распределения, выбираемого вручную, для чего требуется либо ковариация, полученная из реальных данных, либо синтетические ковариационные матрицы, сгенерированные условным вариационным автоэнкодером (cVAE) [8]. Для генерации выбросов в соответствии с теорией экстремальных значений (ТЭЗ) в настоящее время доступны нормальное, Лапласовское, Вейбулловское, Гумбелевское и Леви распределения.

cVAE генерирует синтетические ковариационные матрицы, соответствующие структуре реальных данных. Соответствие сгенерированных ковариационных матриц определенной структуре достигается путем минимизации ошибки восстановления и дивергенции Кульбака-Лейблера между распределением в латентном пространстве и априорным распределением реальных данных. Задача оптимизации решается с помощью декодирующей сети для восстановления реальных данных из латентного пространства, в то время как вариационный автоэнкодер обучается представлениям данных в этом латентном пространстве.

Синтетические табличные данные могут быть необходимы для задач, связанных с дальнейшим анализом отдельных признаков набора данных, в частности целевых переменных [9, 10, 11]. Для формирования синтетических целевых переменных, интегрированных в синтетические данные с выбросами, используется модуль Целевая Модель, основанный на модели классификации, например, такой как Catboost. Модель классификации обучается на выборке реальных данных для получения слабой объяснительной информации из столбцов набора данных, содержащих макро- и микроданные.

Способность zGAN генерировать реалистичные синтетические данные может быть проверена экспериментально и сравнена с метриками качества других GAN.

Предложенные эксперименты по сравнению zGAN с аналогами и прототипом демонстрируют устойчивость модели, а также способность zGAN генерировать реалистичные синтетические данные. Эксперименты основаны на задаче классификации, где имеющаяся целевая переменная в табличных данных используется для вычисления метрики AUC. Данная метрика характеризует качество решения классификационной задачи при сравнении реальных и синтетических данных.

Решение классификационной задачи с целевой переменной, сгенерированной вместе с другими признаками, позволяет оценить способность zGAN генерировать синтетические данные, сохраняя связанность признаков обучающего набора данных. Для этого в эксперименты включена базовая модель, обучение и вычисление показателей качества которой проводится исключительно на реальных данных.

Задача классификации выполнялась в двух вариантах: с разделением данных на обучающую и тестовую выборки по отдельным наблюдениям, то есть в режиме «out-of-sample» (OOS), и с разделением данных по времени, то есть в режиме «out-of-time» (OOT). Способность zGAN генерировать синтетические данные позволяющие сохранять центральную тенденцию и дисперсию AUC при решении классификационной задачи с разными пропорциями синтетических и реальных данных характеризуют устойчивость модели. Устойчивость модели демонстрируется, в том числе воспроизведением классификационных экспериментов на разных наборах данных, как в режиме OOS, так и в режиме OOT.

1601

8

Для демонстрации способности GAN сохранять и воспроизводить корреляции признаков был проведен корреляционный

анализ на исходных и синтетических наборах данных.

Классификационная задача Методика проведенных экспериментов предполагает равные условия и одинаковую последовательность операций для каждого набора данных. Основные этапы проведенных экспериментов описываются следующим образом:

- Парсинг данных;
- Предобработка данных;
- Отбор признаков;
- Разделение данных на обучающую и тестовую выборки;
- Вычисление AUC базовой модели;
- Семплирование обучающих данных (опционально);
- Обучение GAN на реальных обучающих данных;
- Генерация синтетических данных;
- Выборка 80% обучающих и тестовых данных (опционально);
- Обучение классификатора на синтетических данных;
- Расчёт AUC на тестовых данных.

Предобработке данных предшествовал их парсинг. Предобработка включала преобразование типов данных, преобразование дат и времени, если в наборе данных присутствовали временные ряды. Для моделей, обучение которых было невозможно при наличии пропусков, применялись методы заполнения недостающих значений.

Отбор признаков основывался на результатах анализа, направленном на предотвращение влияния избыточности данных, утечки данных и информационного шума. Признаки, содержащие макроданные, не использовались в эксперименте. Основной целью отбора признаков было повышение качества классификационной модели. При обучении генеративных моделей и GAN, как и при генерации данных использовались все признаки, представленные в наборах данных. При этом для решения задачи классификации в наборах данных оставались только отобранные признаки.

Эксперименты «Out-of-Sample» В экспериментах OOS, разделение на обучающую и тестовую выборки осуществлялось с учётом необходимости получения одинаковых законов распределения в обеих выборках.

В качестве базовой модели использовался CatBoost с использованием квази-оптимального поиска гиперпараметров для уменьшения переобучения. Настройка гиперпараметров выполнялась методом поиска по сетке, при этом границы поиска по сетке были определены методом случайного поиска. Во всех экспериментах решалась задача бинарной классификации.

Базовая классификационная модель обучалась на обучающей выборке, а AUC вычислялся на тестовой. Для получения разброса метрики применялся подход Монте-Карло, имплементируемый следующим образом: случайным образом формировалась

выборка, составляющая 80% от общего объема данных, отдельно для обучающих и тестовых данных. На каждой из полученных выборок обучалась классификационная модель, и вычислялась соответствующая метрика. Итоговая метрика, представленная в исследовании, — это медианное значение AUC, полученное за 51 эксперимент для каждой модели и набора данных отдельно. 51-е значение в результатах экспериментов — это значение AUC, полученное без 80% выборки, то есть оно было вычислено с использованием всех записей из обучающих и тестовых наборов данных.

Перед обучением генераторов синтетических данных для моделей, чувствительных к объему обучающих данных, набор данных дополнялся семплированием методом случайной выборкой. Для моделей, где реализм синтетических данных был значительно хуже по сравнению с другими моделями, применялась ручная настройка гиперпараметров. При ручной настройке предпочтение отдавалось конфигурациям, которые достигали максимального значения AUC. Во всех случаях обученные генеративные модели и GAN генерировали по 4000 записей.

Для окончательного вычисления AUC модель классификации обучалась на синтетических обучающих данных, а качество классификации оценивалось на реальных тестовых данных. Разброс AUC в этом случае рассчитывался по процедуре, полностью аналогичной описанной для базовой модели. Единственным отличием было то, что выборки брались из синтетических обучающих данных, а не из реальных.

Эксперименты «Out-of-Time» В экспериментах ООТ набор данных был разделен на обучающие и тестовые выборки в порядке возрастания дат. Данные делились на обучающую и тестовую выборки отдельно в двух пропорциях: 50% и 80%.

В классификационном ООТ эксперименте использовались для генерации синтетических данных zGAN и его прототип CTGAN, классификационная модель обучалась на реальном обучающем наборе данных, при этом метрика AUC вычислялась только на реальном тестовом наборе данных, аналогично процедуре производимой в OOS экспериментах.

Для дальнейшего анализа стабильности обучающий набор случайным образом смешивался с синтетическими данными в различных пропорциях. При каждом смешивании проводился классификационный эксперимент, и вычислялось AUC.

В «Out-of-Time» экспериментах была также исследована стабильность и повышение AUC в зависимости от количества выбросов, сгенерированных из нормального распределения на уровне 3σ в макропеременных наборах данных.

Такие эксперименты проводились только для zGAN, поскольку ни аналоги, ни прототип не обладают возможностью целенаправленной генерации выбросов. В этих экспериментах каждый набор данных разделялся на обучающую и тестовую выборки на основе временных меток. Часть набора данных до 2022 года использовалась как обучающая выборка, в то время как данные с 2022 года включительно составляли тестовую выборку. Модель zGAN обучалась

на тренировочном наборе, после чего для каждого процента выбросов в макропеременных, включая вариант без выбросов, генерировалось по 80 синтетических наборов данных.

В задаче бинарной классификации модель классификации обучалась для каждого набора синтетических данных. AUC оценивался с применением метода Монте-Карло, на случайной выборке из тестового набора, которая составляла 80% от всего тестового набора данных. Последний шаг, включал 81ю итерацию вычисления AUC. Для обучения классификационной модели на последней итерации, использовался объединенный набор данных, состоящий из всех 80 сгенерированных наборов для каждого процента выбросов в отдельности, при этом AUC вычислялся на всем тестовом наборе.

Для оценки влияния выбросов на AUC в качестве метрик оценки использовались медианное значение AUC, размах и межквартильный размах (IQR).

Корреляционный анализ. Для оценки способности GAN сохранять и воспроизводить корреляции признаков использовался коэффициент корреляции Пирсона. Как исходный набор данных, так и синтетические данные были предварительно обработаны следующим образом: пропущенные значения в числовых столбцах заполнялись консистентными значениями, которые значительно отличались от среднего значения столбца и не встречались в других местах данных. В категориальных столбцах пропущенные значения были сгруппированы в отдельную категорию, а все записи в категориальных столбцах были закодированы с использованием OrdinalEncoder. Столбцы с датами и временем были преобразованы в формат Unix-time. После этих преобразований набор данных был нормализован с использованием StandardScaler.

Для визуального наблюдения различий в способности различных GAN воспроизводить корреляции использовались тепловые карты (heatmaps) различий между корреляционными матрицами исходного и каждого синтетического набора данных. Для всех тепловых карт была установлена единая цветовая шкала.

Описание данных В экспериментах использовались как приватные наборы данных с доступом, разрешенным компанией zypl.ai, так и открытый набор данных Титаник. В интересах владельцев исходных данных характеристики приватных наборов, включая их названия, скрыты и обозначены пронумерованными латинскими буквами: A₁, ..., A₉. Все наборы данных демонстрируют дисбаланс классов в целевой переменной. Коэффициент k_{im} , использующийся для описания дисбаланса классов, представляет собой отношение количества образцов в миноритарном классе к количеству образцов в мажоритарном классе.

На качество генерации синтетических данных влияют в том числе общее количество столбцов и записей, а также количество категориальных столбцов в каждом из наборов данных.

Отношение миноритарных классов к мажоритарным k_{im} , общее количество столбцов и записей, количество категориальных столбцов и наличие вре-

менного ряда в частных наборах данных приведены в таблице 1.

Таблица 1

Показатели	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈	A ₉
k_{im}	0.07	0.19	0.02	0.08	0.01	0.14	0.21	0.23	0.06
Количество столбцов	21	29	33	18	29	15	30	68	25
Количество записей	347K	62K	1.76M	34K	759K	18K	3189	139K	2721
Количество категориальных столбцов	6	10	16	18	13	6	15	10	9
Количество категорий	56	44	370	60	44	29	242	47	67
Наличие timeseries	V	V	V	X	X	X	V	V	V

Эксперименты были проведены, в том числе, на открытом наборе данных - Титаник. Обучающая выборка оригинального набора данных Титаник была разделена на обучающую и тестовую выборки в соотношении 0.33. Обучающие данные включали 596 записей, а тестовые данные — 295 записей.

Целевая переменная — «Survived», признаки, отобранные для дальнейших экспериментов: «Pclass», «Sex», «Age», «SibSp», «Parch», «Fare», «Cabin», «Embarked». Отношение миноритарного класса к мажоритарному k_{im} составляет 0.62. Набор данных содержит всего 9 признаков, из которых 7 являются категориальными. В общей сложности в этих столбцах присутствуют 173 категории.

Для обучения всех моделей GAN обучающий набор данных, после разделения, был дополнен случайной выборкой до общего количества 10728 записей. Для моделей других типов дополнение данных не применялось, так как качество генерации синтетических данных у этих моделей не зависит от количества обучающих данных.

Описание моделей Эксперименты включали

две GAN модели, а именно CTGAN (прототип) и CopulaGAN (аналог); одну альтернативную модель - TVAE (аналог). CTGAN использует условную вероятность для моделирования сложных зависимостей в табличных данных, что делает его эталоном для задач, включающих как числовые, так и категориальные переменные. TVAE предлагает вероятностный подход, основанный на автоэнкодерах, что позволяет моделировать сложные распределения.

CopulaGAN, использует копулы для моделирования зависимостей между переменными и объединяет статистическое моделирование на основе копул с нейросетевыми подходами. Разнообразие в выборе моделей для сравнения позволяет всесторонне оценить, как zGAN справляется с моделированием зависимостей и генерацией синтетических данных по сравнению с существующими способами.

Результаты экспериментов

Классификационная задача

Out-of-Sample эксперименты В таблице 2 показаны медианные значения AUC полученные в результате экспериментов.

Таблица 2

Набор данных	zGAN	CTGAN	TLAE	CopulaGAN	Baseline
A ₁	0.7553	0.7496	0.7504	0.6681	0.7841
A ₂	0.7972	0.7633	0.7428	0.7959	0.8216
A ₃	0.7230	0.7035	0.6461	0.6219	0.7552
A ₄	0.7095	0.6350	0.6244	0.6663	0.7343
A ₅	0.8311	0.8284	0.6163	0.6763	0.8680
A ₆	0.8120	0.8025	0.7898	0.7942	0.8154
A ₇	0.7704	0.7013	0.7279	0.7093	0.8183
A ₈	0.7118	0.6992	0.7257	0.7010	0.7907
A ₉	0.6525	0.5705	0.6874	0.5955	0.7109
Титаник	0.8163	0.7923	0.7874	0.8154	0.8736

Результаты показывают, что в восьми из десяти экспериментов модель zGAN превосходит все аналоги и прототип, за исключением базовой модели, обученной на реальных данных. В среднем AUC модели zGAN на 0,03 пункта выше, чем у CTGAN, и на 0,05 пункта выше, чем у TVAE или CopulaGAN.

Out-of-Time эксперименты Результаты OOT

экспериментов с моделью zGAN, в которых использовалась пропорциональная смесь реальных и синтетических данных, представлены ниже в таблице 3.

Таблица 3

Набор данных	100% Синтетические	1:1	0.1:1	0.01:1	0.001:1	100% Реальные Данные
A ₁	0.6738 (0.6641 : 0.6812)	0.6684 (0.6335 : 0.6862)	0.6262 (0.6017 : 0.6415)	0.6165 (0.5979 : 0.6320)	0.6138 (0.6014 : 0.6305)	0.6100 (0.5996 : 0.6184)
A ₂	0.7862 (0.7786 : 0.7917)	0.8108 (0.8015 : 0.8156)	0.8157 (0.8087 : 0.8204)	0.8191 (0.8138 : 0.8225)	0.8227 (0.8178 : 0.8264)	0.8238 (0.8195 : 0.8279)
A ₃	0.7146 (0.7101 : 0.7186)	0.7302 (0.7264 : 0.7339)	0.7433 (0.7389 : 0.7472)	0.7435 (0.7395 : 0.7462)	0.7444 (0.7393 : 0.7479)	0.7456 (0.7417 : 0.7488)
A ₇	0.8250 (0.8134 : 0.8517)	0.8219 (0.8050 : 0.8446)	0.8092 (0.7854 : 0.8366)	0.8091 (0.7812 : 0.8295)	0.8076 (0.7830 : 0.8241)	0.8091 (0.7847 : 0.8378)
A ₈	0.7686 (0.7628 : 0.7725)	0.7804 (0.7767 : 0.7834)	0.7853 (0.7787 : 0.7893)	0.7859 (0.7807 : 0.7900)	0.7865 (0.7824 : 0.7903)	0.7859 (0.7799 : 0.7895)
A ₉	0.7117 (0.6674 : 0.7747)	0.7233 (0.6660 : 0.7805)	0.6443 (0.5628 : 0.7069)	0.6627 (0.5789 : 0.7597)	0.7046 (0.6568 : 0.7567)	0.6862 (0.6129 : 0.7429)

Экспериментальные результаты показывают сравнимые показатели предсказательной точности классификационных моделей, обученных на синтетических данных zGAN, комбинации данных zGAN и реальных данных, а также базовой модели, обученной только на реальных данных. В случаях наборов данных A₁, A₇ и A₉ медианное значение AUC существенно возрастает при использовании синтетиче-

ских данных или добавлении пропорционально большого объема синтетических данных к реальным данным. В случае набора данных A₁ AUC увеличивается на 0,064 пункта.

Сравнительные результаты OOT-экспериментов, полученные на основе наборов данных, сгенерированных CTGAN, показаны ниже в таблице 4.

Таблица 4

Набор Данных	100% Синтетические	1:1	0.1:1	0.01:1	0.001:1	100% Реальные Данные
A ₁	0.7077 (0.6959 : 0.7151)	0.7054 (0.6675 : 0.7144)	0.6295 (0.6137 : 0.6678)	0.6177 (0.6047 : 0.6338)	0.6103 (0.6018 : 0.6196)	0.6100 (0.5996 : 0.6184)
A ₂	0.7815 (0.5000 : 0.7933)	0.8063 (0.8016 : 0.8111)	0.8199 (0.8140 : 0.8245)	0.8235 (0.8184 : 0.8275)	0.8237 (0.8189 : 0.8287)	0.8238 (0.8195 : 0.8279)
A ₃	0.7025 (0.6884 : 0.7109)	0.7368 (0.7317 : 0.7396)	0.7452 (0.7425 : 0.7478)	0.7453 (0.7414 : 0.7488)	0.7456 (0.7427 : 0.7491)	0.7456 (0.7417 : 0.7488)
A ₇	0.7707 (0.7456 : 0.7961)	0.8117 (0.7951 : 0.8370)	0.8108 (0.7862 : 0.8406)	0.8090 (0.7830 : 0.8324)	0.8079 (0.7802 : 0.8363)	0.8091 (0.7847 : 0.8378)
A ₈	0.7481 (0.7409 : 0.7527)	0.7761 (0.7700 : 0.7812)	0.7814 (0.7772 : 0.7857)	0.7855 (0.7798 : 0.7892)	0.7862 (0.7822 : 0.7898)	0.7859 (0.7799 : 0.7895)
A ₉	0.6861 (0.6731 : 0.7050)	0.7072 (0.6725 : 0.7427)	0.7136 (0.7039 : 0.7214)	0.7094 (0.6897 : 0.7202)	0.7104 (0.7014 : 0.7283)	0.6862 (0.6129 : 0.7429)

В OOT-эксперименте по смешиванию реальных и синтетических данных с использованием CTGAN было продемонстрировано влияние смешивания на медианное значение AUC. В случае набора данных A₂ AUC, рассчитанное на реальных данных, на 0,001 пункта превышает наивысшее значение AUC, полученное при смешивании данных, тогда как в случае A₃ улучшения AUC при смешивании не было. В случаях A₁, A₇, A₈ и A₉ наблюдалось увеличение AUC при смешивании, при этом

наибольшее увеличение составило 0,0977 пункта для датасета A₁.

OOT-эксперименты по смешиванию реальных обучающих и синтетических данных показывают в большинстве случаев увеличение AUC, что также указывает на улучшение стабильности в задаче классификации. Увеличение медианного AUC на синте

тических данных или при смешивании синтетических данных с реальными данными наблюдается как в экспериментах с CTGAN, так и с zGAN, что указывает на то, что улучшение стабильности в задаче классификации происходит независимо от GAN, используемой для генерации синтетических

Результаты ООТ-экспериментов с синтетическими выбросами в макроданных синтетических наборов данных, сгенерированных zGAN, показаны в таблице 5, на примере набора данных А₉, характеризующего растущую экономику.

Процент выбросов	Медианное AUC (Размах)
100%	0.6862 (0.5706 : 0.7947)
50%	0.6867 (0.5669 : 0.7677)
10%	0.6962 (0.5736 : 0.7769)
7.7%	0.7061 (0.6357 : 0.7819)
7.4%	0.7122 (0.5985 : 0.8048)
7.1%	0.7077 (0.6039 : 0.7770)
7%	0.7095 (0.6170 : 0.7948)
6.9%	0.7074 (0.6163 : 0.7807)
6.6%	0.7079 (0.5985 : 0.7650)
6.3%	0.6966 (0.5838 : 0.7875)
6%	0.7096 (0.6371 : 0.7853)
5%	0.7147 (0.5858 : 0.7890)
3%	0.7116 (0.6130 : 0.7694)
1%	0.7032 (0.6027 : 0.7736)
Без выбросов	0.7020 (0.6154 : 0.8194)

данных.

Таблица 5

Эксперименты с синтетическими выбросами показывают увеличение медианного AUC в большинстве случаев, без явной линейной зависимости от процента выбросов в макроданных. Когда количество синтетических выбросов превышает 10% от общего числа записей в макроданных, медианный AUC снижается. Наибольшее увеличение медианного AUC наблюдается при 3%, 5% и 7% выбросов в столбцах макроданных. Значительное снижение AUC наблюдается, когда количество выбросов в синтетических макроданных достигает 50 и более процентов. Максимальное и статистически значимое увеличение AUC наблюдается при 5% выбросов в синтетических макроданных.

Результаты эксперимента со сгенерированными синтетическими выбросами демонстрируют полезность выбросов для улучшения качества решения задачи классификации, а также характеризуют стабильность zGAN в задаче целенаправленной генерации синтетических данных с синтетическими выбросами.

Корреляционный анализ На Фиг. 3 представлены тепловые карты, иллюстрирующие различия между матрицами корреляции синтетических наборов данных, сгенерированных zGAN, прототипом и аналогами, по сравнению с оригинальным набором данных. Способ TVAE показывает умеренные различия, что указывает на адекватное сохранение оригинальных корреляций. Способ CopulaGAN демонстрирует более значительные изменения, что свидетельствует о менее надежном сохранении оригинальной структуры. Прототип - CTGAN и zGAN показывают наименьшие различия, при этом zGAN превосходит прототип, что указывает на наивысшую степень точности в поддержании оригинальной структуры корреляций. Таким образом, CTGAN и zGAN являются наиболее эффективными в генерации синтетических данных, которые близки к ориги-

нальному набору данных, за которыми следует TVAE, который предлагает адекватный баланс между точностью и качеством сохранения корреляций, а CopulaGAN является наименее эффективным.

Заключение Эксперименты при out-of-sample разделении данных, демонстрируют сравнение качества классификации на синтетических данных, сгенерированных современными открытыми GAN и генеративными моделями, основанными на классических принципах, в том числе прототипом и аналогами, демонстрируют преимущества zGAN по сравнению с исследованными моделями. В среднем, AUC zGAN превышает таковой у прототипа на 0.03 пункта, а также на 0.05 пункта у аналогов.

Применение zGAN для генерации синтетических табличных данных позволяет модели классификации изучать основные закономерности в данных без взаимодействия с реальными данными клиентов. Эксперименты при out-of-time разделении данных показывают, что обучение классификационной модели на синтетических наборах данных улучшает ее способность к обобщению, положительно влияя на качество классификационной задачи при работе с будущими, ранее не виденными данными. В эксперименте по обогащению и смешиванию реальных данных с синтетическими данными для набора данных А₁ было достигнуто увеличение медианного AUC на 0.0638 пункта при использовании только синтетических данных для обучения, и увеличение AUC на 0.0584 пункта было получено при смешивании реальных данных для обучения с синтетическими данными в соотношении 1:1.

Генерация синтетических выбросов в случае ООТ экспериментов, где тренировочные данные были собраны до 2022 года, а тестовые данные — с 2022 по 2023 год, приводит к значительному улучшению качества решения задачи классификации, как показано на наборе данных А₉, представляющем рас-

тущую экономику. В эксперименте по генерации синтетических выбросов в макроданных синтетического набора данных было достигнуто увеличение медианного AUC при умеренном количестве синтетических выбросов до 10%, в то время как увеличение выбросов выше 50% снижало качество. Наибольшее и наиболее значительное повышение AUC на 0.0127 пункта наблюдалось при 5% выбро-

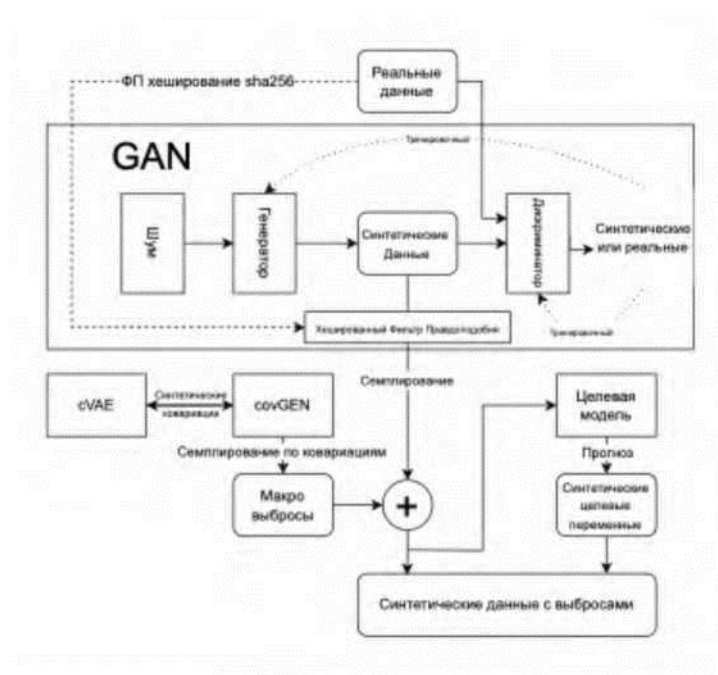
сов. Эти результаты подчеркивают важность тщательной балансировки количества выбросов, особенно в прогнозах out-of-time, где выбросы часто обладают большей предсказательной способностью.

В экспериментах по корреляционному анализу zGAN демонстрирует способность сохранять корреляции реальных данных на уровне прототипа, при этом превосходит аналоги.

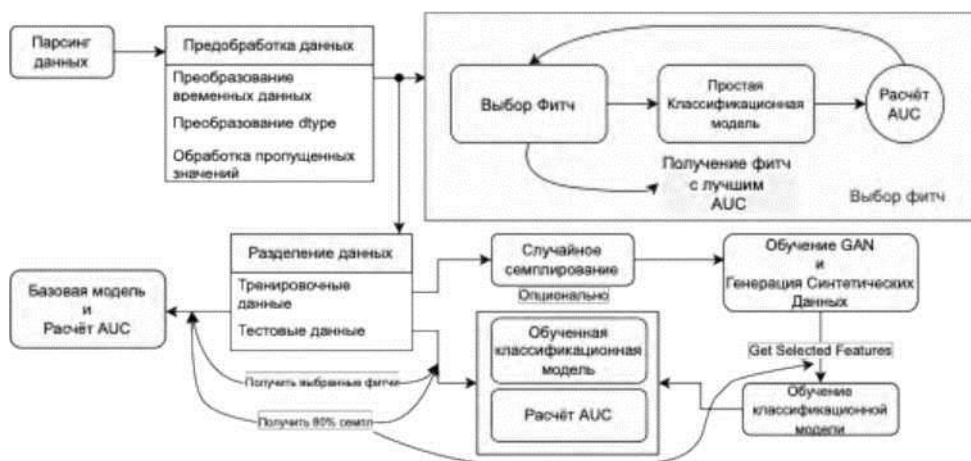
Формула изобретения

Способ генерация реалистических табличных данных с выбросами, включающий использование архитектуры, позволяющей производить условную генерацию для фокусирования модели на специфических подмножествах данных, сохраняя редкие категории и сложные зависимости между признаками, инициализирование генератора случайным шумом для формирования синтетических данных, обучение генератора на повышение его способности формировать реалистические синтетические данные с применением метода обратного распространения ошибки, обучение дискриминатора на различении синтетических данных, поступающих от генератора и реальных данных с целью повышения реалистичности синтетических данных, отличающийся тем, что для генерации синтетических данных применяют архитектуру генеративно-сопоставительной сети (zGAN), содержащая генератор, дискриминатор, модуль генерации ковариацией (covGEN), хешированный фильтр правдоподобия, условный вариационный автоэнкодер (cVAE), а также блоки позволяющие генерировать синтетические целевые переменные, причем, применение архитектуры (zGAN) обеспечивает генерацию синтетических табличных данных с целенаправленно

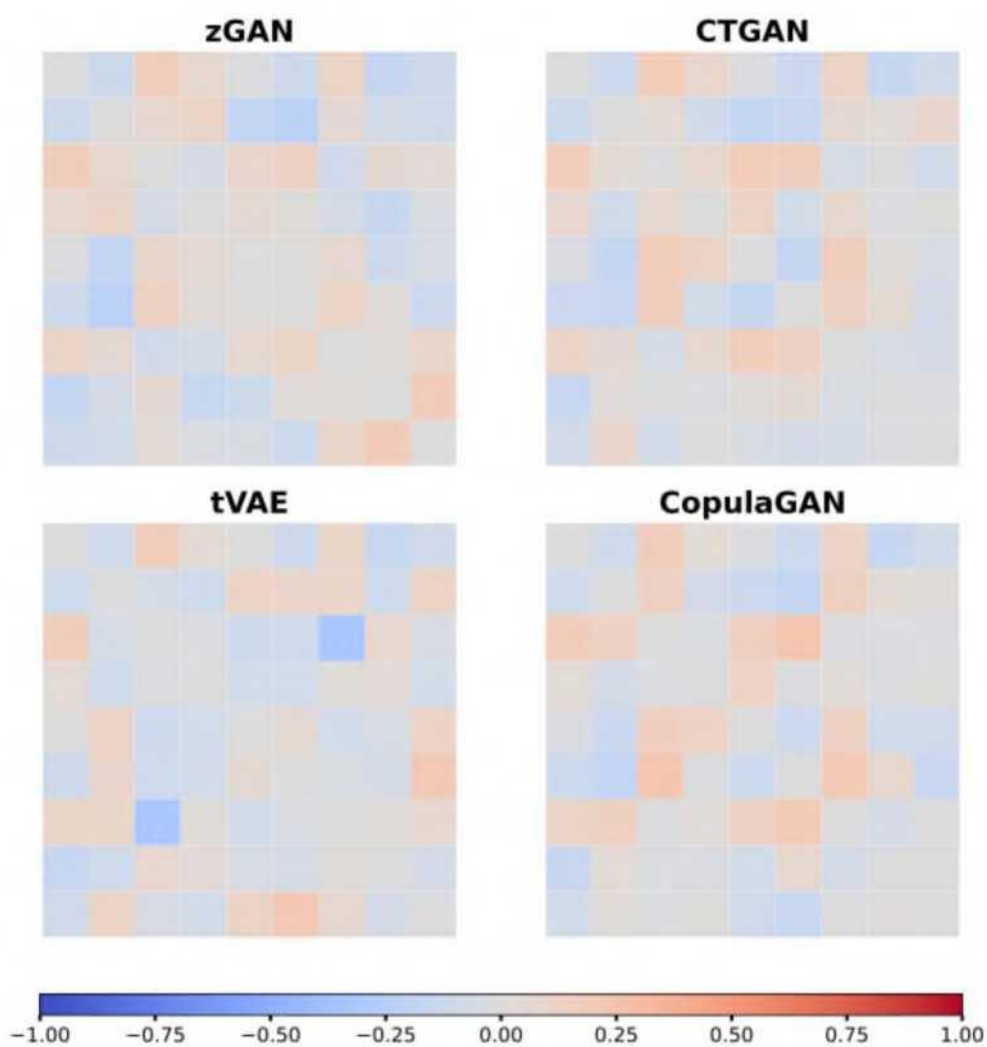
включенными выбросами, распределительными в соответствии нормальным, Лапласовским, Вейбулловским, Гумбелевским и Леви распределениями, используют условным вариационным автоэнкодер (cVAE) для генерации синтетических ковариационных матриц с целью последующей генерации выбросов, вычисляют ковариационные матрицы на основе реальных данных с использованием модуля covGEN для создания выбросов, обеспечивают выбор между применением синтетических ковариационных матриц, получаемых от модуля cVAE и ковариационных матриц, полученных на основе реальных данных от модуля covGEN для последующей целенаправленной генерации синтетических выбросов, применяют фильтр правдоподобия для повышения приватности сгенерированных данных путем использования хеш-кодов с исключением сохранения клиентских данных, а также проводят вычисление показателей качества, таких как AUC, для оценки стабильности и предсказательной способности модели на основе сгенерированных данных и применяют синтетических данных для обучения моделей машинного обучения и анализа влияния выбросов на показатель качества классификационной модели.



Фиг. 1



Фиг. 2



Фиг. 3

Компьютерный набор: Махмудов Ф.