УДК 004.05
DOI: 10.18799/29495407/2025/2/92
Шифр специальности ВАК: 2.3.1

Методы генерации данных для медицинских исследований

Е.И. Губин✉, А.О. Котов

Томский политехнический университет, Россия, г. Томск

✉gubine@tpu.ru

Аннотация. Представлено краткое описание современных технологий генерации данных и их возможное дальнейшее использование в медицинских исследованиях с применением методов синтезирования данных и выбор наиболее точных из библиотеки SDV (Synthetic Data Vault).

Ключевые слова: GAN, TVAE, Gaussian Copula, данные, логистическая регрессия, ROC-кривые

Для цитирования: Губин Е.И., Котов А.О. Методы генерации данных для медицинских исследований // Известия Томского политехнического университета. Промышленная кибернетика. – 2025. – Т. 3. – № 2. – С. 46–50. DOI: 10.18799/29495407/2025/2/92

UDC 004.05
DOI: 10.18799/29495407/2025/2/92

Data generation methods for medical research

E.I. Gubin✉, A.O. Kotov

National Research Tomsk Polytechnic University, Tomsk, Russian Federation

✉gubine@tpu.ru

Abstract. The article presents a brief description of modern data generation technologies and their possible further use in medical research using data synthesis methods and selecting the most accurate ones from the SDV (Synthetic Data Vault) library.

Keywords: GAN, TVAE, Gaussian Copula, data, logistic regression, ROC curves

For citation: Gubin E.I., Kotov A.O. Data generation methods for medical research. *Bulletin of the Tomsk Polytechnic University. Industrial cybernetics*, 2025, vol. 3, no. 2, pp. 46–50. DOI: 10.18799/29495407/2025/2/92

Введение

Данная статья является дальнейшим развитием методов генерации данных для медицинских исследований, начатых авторами в работе [1] и продолженных с использованием современных прогнозных моделей, при сравнении прогнозных точностей исходных и размноженных данных в работе [2].

Основные методы синтезирования исходных данных

В рамках численного эксперимента был использован тестовый датасет, содержащий 1500 записей с различными атрибутами, такими как возраст, доход, уровень образования и профессиональные навыки и т. д. В общей сложности в датасете присутствует 15 признаков, как числовых, так и категориальных, что делает его подходящим для проверки возможностей изучаемых методов.

Эксперимент проводился в следующей последовательности: тестовый датасет случайным образом делился на группы с размерами 16, 35, 60, 120, 240 строк с использованием генератора случайных чисел. Эти группы стали основой для поэтапного анализа на платформах SDV (Synthetic Data Vault),

таких как TVAE, CTGAN и Gaussian Copula. Для каждой группы проводились тестовые расчеты синтетических и реальных данных для всех трех платформ. Сравнение результатов осуществлялось с использованием встроенного метода оценки качества синтетических и реальных данных из библиотеки SDV. Каждая из моделей (TVAE, CTGAN, Gaussian Copula) обучалась на подготовленных данных с одинаковыми настройками гиперпараметров для обеспечения сопоставимости результатов.

Метод синтезирования данных CTGAN [3–6] – модель, основанная на генеративно-состязательных сетях, специально адаптированная для табличных данных с поддержкой категориальных переменных и управления условными зависимостями.

Метод GAN

Generative Adversarial Networks (GAN) были впервые предложены в 2014 г. исследователем Иэном Гудфеллоу и его коллегами в статье "Generative Adversarial Nets".

Принцип работы GAN

GAN состоит из двух нейронных сетей: генератора и дискриминатора, которые обучаются одновременно в процессе, называемом "соперничеством".

- Генератор: эта сеть принимает случайный шум (обычно из нормального распределения) в качестве входных данных и генерирует новые образцы данных, которые должны быть похожи на реальные данные из обучающего набора. Цель генератора – создавать такие данные, которые будут трудно отличимы от реальных.
- Дискриминатор: эта сеть принимает на вход как реальные данные, так и данные, сгенерированные генератором. Она должна определить, являются ли данные реальными или сгенерированными. Дискриминатор выдает вероятность того, что входные данные реальные.

Процесс обучения:

- Соперничество: генератор и дискриминатор обучаются одновременно. Генератор пытается улучшить свои способности к созданию реалистичных данных, в то время как дискриминатор улучшает свои навыки по различению реальных и синтетических данных.
- Обновление весов: в процессе обучения генератор получает обратную связь от дискриминатора, что позволяет ему улучшать свои выходные данные. Дискриминатор, в свою очередь, также обновляет свои веса на основе ошибок, связанных с неправильной классификацией данных.
- Цель: обе сети стремятся к оптимизации своих функций потерь. Генератор пытается минимизировать вероятность того, что дискриминатор правильно классифицирует его выходные данные как поддельные, а дискриминатор пытается максимизировать свою точность в классификации.

В результате этого процесса обе сети становятся все более совершенными, что приводит к созданию высококачественных синтетических данных. GAN нашли широкое применение в различных областях, таких как создание изображений, видео, музыки и даже текстов.

Метод TVAE

TVAE (Triplet based Variational Autoencoder, Тензорный Вариационный Автоэнкодер) – это модель, расширяющая вариационные автоэнкодеры (*Variational Autoencoder* – VAE) для работы с многомерными данными (тензорами), такими как изображения и временные ряды.

Основные компоненты:

- кодировщик: преобразует входные данные в латентное пространство, генерируя параметры распределения;
- декодировщик: восстанавливает оригинальные данные из латентного представления;
- вариационное предположение: оценивает вероятностное распределение латентных переменных для учета неопределенности.

Процесс обучения:

- функция потерь: состоит из двух частей – реконструкции (качество восстановления данных) и регуляризации (KL-дивергенция);
- обучение: использует градиентный спуск для минимизации функции потерь.

Гауссовская копула – это математический инструмент, который позволяет моделировать зависимость между многими случайными переменными, сохраняя при этом их маргинальные распределения. Она используется в синтезировании данных для создания многомерных выборок с заданными зависимостями.

Основные компоненты:

- копула: функция, которая связывает многомерное распределение с его маргинальными распределениями. Гауссовская копула использует многомерное нормальное распределение для описания зависимости между переменным;
- маргинальные распределения: это индивидуальные распределения каждой переменной, которые могут быть различными (например, нормальными, экспоненциальными и т. д.).

Процесс синтезирования данных:

- определение маргинальных распределений: выбираются распределения для каждой переменной;
- параметризация копулы: определяются параметры зависимости (например, корреляционная матрица для гауссовской копулы);

- генерация зависимых данных: сначала генерируются независимые нормально распределенные случайные величины, которые затем преобразуются с помощью обратного преобразования маргинальных распределений.

Первым экспериментом было прямое сравнение методов генерации и оценка их с использованием методов оценки библиотеки SDV.

Библиотека SDV использует следующие методы для оценки качества синтетических данных:

- сравнение распределений: анализ распределений оригинальных и синтетических данных с помощью визуализаций (гистограммы) и статистических тестов (например, тест Колмогорова–Смирнова);
- метрики: оценка точности с использованием метрик, таких как средняя абсолютная ошибка и среднеквадратичная ошибка.

Результаты эксперимента

В ходе проведенного численного эксперимента получены результаты влияния исходных размеров подвыборок и параметра оценки встроенного метода оценки, основанного на среднем показателе качества генерации, на качество прогнозной модели. Результаты проведенного исследования показаны в таблице.

Таблица. *Средняя оценка качества генерации данных*
Table. *Average rating of data generation quality*

Исходные размеры групп (строк) Initial sizes of groups (lines)	Наименование метода Method name		
	TVAE	CTGAN	Gaussian/Copula
16	0,72	0,74	0,86
30	0,79	0,78	0,73
60	0,79	0,78	0,61
120	0,68	0,81	0,57
240	0,73	0,84	0,57

На основании полученных результатов можно сделать вывод, что выбор метода генерации зависит от специфики задачи: для данных с преобладанием числовых признаков и линейных зависимостей лучше всего подходит Gaussian Copula, для категориальных данных и задач с несбалансированными классами оптимальным выбором является CTGAN, для сложных задач, требующих моделирования многомерных зависимостей, наиболее эффективным является TVAE.

Следующим этапом эксперимента было сравнение показателей оценки логистической регрессии, модели были обучены на синтетических данных. Первая модель была обучена на оригинальной выборке, вторая – на синтетических данных, полученных с использованием CTGAN.

Сравнительный анализ моделей, обученных на оригинальных и синтетических данных, показал, что синтетические данные могут быть использованы в качестве эффективной альтернативы при отсутствии доступа к реальным данным. Несмотря на небольшую разницу в статистиках ROC, обе модели продемонстрировали высокую предсказательную способность и схожую стабильность.

На рис. 1, 2 представлены ROC-кривые для оригинального датасета и синтезированного.

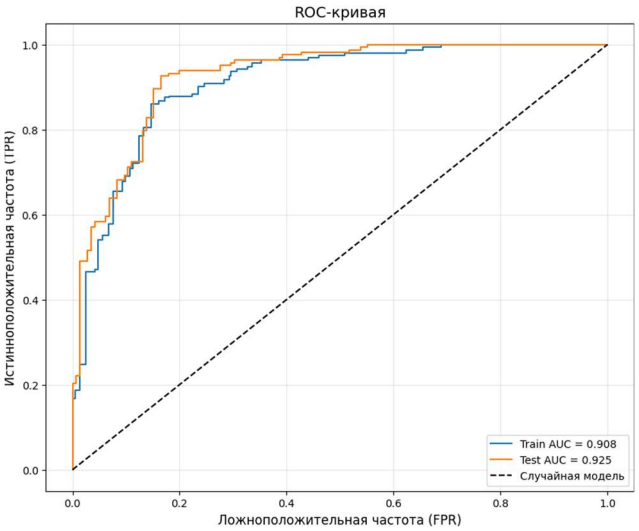

Рис. 1. *ROC-кривые оригинального датасета*
Fig. 1. *ROC curves of the original dataset*

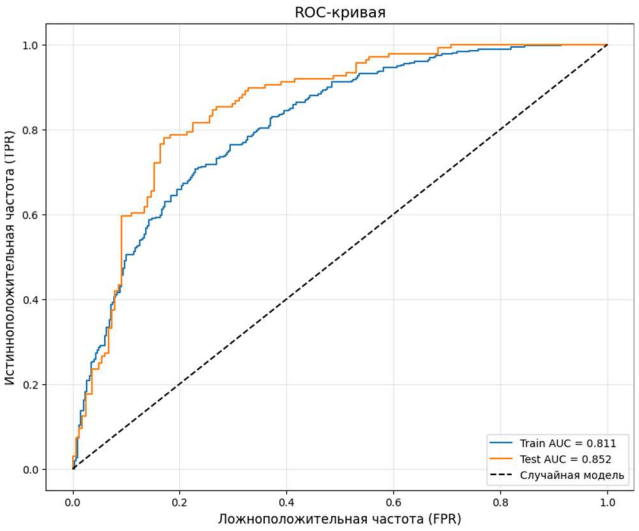

Рис. 2. *ROC-кривые синтетического датасета*
Fig. 2. *ROC curves of the synthetic dataset*

Эти результаты подчеркивают важность дальнейшего исследования возможностей применения синтетических данных в различных областях, включая медицину, где конфиденциальность и безопасность данных являются приоритетами. Использование CTGAN для генерации синтетических

данных открывает новые горизонты для разработки и тестирования моделей машинного обучения без риска нарушения прав пациентов или утечки конфиденциальной информации.

Результаты статистики кросс-валидации представлены ниже.

- Для модели, построенной на синтетических данных: accuracy=0,74 со стандартным отклонением 0,03.
- Для модели, на оригинальных данных: accuracy=0,85 со стандартным отклонением 0,05.

На основании полученных данных можно сделать вывод, что статистические показатели оценки точности обеих выборок являются приемлемыми. Точность модели на оригинальных данных (0,85) и на синтетических данных (0,74) указывает на удовлетворительную надежность модели. Стандартные отклонения также находятся в пределах допустимых значений, что свидетельствует о стабильности результатов.

Заключение

Результаты работы подтвердили высокую эффективность методов генерации синтетических данных для обучения моделей машинного обучения, что открывает новые возможности для решения задач, связанных с недостатком реальных данных. Применение синтетических данных позволяет не только обойти ограничения, связанные с конфиденциальностью и доступом к данным, но и создать более разнообразные выборки, что может повысить обобщающую способность моделей.

Можно отметить, что использование синтетических данных открывает новые горизонты в области анализа больших данных и машинного обучения. Это позволяет исследователям и практикам разрабатывать более точные и надежные модели для решения сложных задач в медицине и других сферах. Синтетические данные могут стать ключевым инструментом в борьбе с недостатком информации и помогут создать более безопасные и эффективные системы принятия решений в здравоохранении.

СПИСОК ЛИТЕРАТУРЫ

1. Губин Е.И., Виничук С.В. Влияние размера исходных данных на прогнозный анализ // Перспективы науки. – 2023. – № 10 (169). – С. 10–12. EDN: UFVSIM
2. Губин Е.И., Котов А.О. Использование современных технологий синтезирования данных для медицинских исследований // Перспективы науки. – 2025. – № 3 (186). – С. 19–22.
3. SDV Documentation. URL: https://sdv.dev (дата обращения: 22.04.2025).
4. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. – М.: Диалектика, 2019. – 480 р.
5. Документация Scikit-learn. URL: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.ht ml (дата обращения: 22.04.2025).
6. El Emam Kh., Mosquera L., Hoptroff R. Practical synthetic data generation. – Sebastopol, CA: O'Reilly Media, Inc., 2020. – 241 p.

Сведения об авторах

Евгений Иванович Губин, кандидат физико-математических наук, доцент отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета, Россия, 634050, г. Томск, пр. Ленина, 30. gubine@tpu.ru

Андрей Олегович Котов, магистрант отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета, Россия, 634050, г. Томск, пр. Ленина, 30. aok37@tpu.ru

Поступила: 25.04.2025
После рецензирования: 15.06.2025
Опубликована: 30.06.2026

REFERENCES

1. Gubin E.I., Vinichuk S.V. Impact of input data size on predictive analysis. *Science prospects*. 2023, no. 10 (169), pp. 10–12. (In Russ.) EDN: UFVSIM
2. Gubin E.I., Kotov A.O. Use of modern data synthesis technologies for medical research. *Science prospects*, 2025, no. 3 (186), pp. 19–22. (In Russ.)
3. *SDV Documentation*. Available at: https://sdv.dev (accessed: 22 April 2025).
4. Müller A., Guido S. *Introduction to Machine Learning with Python. A Guide for Data Scientists*. Moscow, Dialectika Publ., 2019. 480 p. (In Russ.)
5. *Scikit-learn documentation*. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForest Classifier.ht ml (accessed: 22 April 2025).
6. El Emam Kh., Mosquera L., Hoptroff R. *Practical synthetic data generation*. Sebastopol, CA, O'Reilly Media, Inc., 2020. 241 p.

About the authors
Evgeny I. Gubin, Cand. Sc., Associate Professor, National Research Tomsk Polytechnic University, 30, Lenin avenue, Tomsk, 634050, Russian Federation. gubine@tpu.ru
Andrey O. Kotov, Master Student, National Research Tomsk Polytechnic University, 30, Lenin avenue, Tomsk, 634050, Russian Federation. aok37@tpu.ru

Received: 25.04.2025
Revised: 15.06.2025
Accepted: 30.06.2026