ИССЛЕДОВАНИЕ МЕТОДОВ ГЕНЕРАЦИИ ДАННЫХ ДЛЯ МЕДИЦИНСКИХ ИССЛЕДОВАНИИЙ

Губин Е.И.[1], Котов А.О.[2]
[1] Томский политехнический университет, к.ф.-м.н., доцент ОИТ ИШИТР
[2] Томский политехнический университет, ИШИТР, магистр группы 8ПМ31
e-mail: аок37@tpu.ru

Аннотация

Для решения проблемы недостаточности данных в медицинских исследованиях применяются методы синтезирования данных такие как GAN, TVAE, Gaussian Copula. Затем производиться оценка генерации в сравнении с оригинальной выборкой. Важным методом оценки является логистическая регрессия. Перед синтезированием производиться очистка данных от аномальных.

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

Введение

Целью данной работы является дальнейшее развитие методов генерации данных для медицинских исследований, начатых авторами в [1], и продолженных с использованием современных прогнозных моделей и сравнивая прогнозные точности исходных и размноженных данных [2].

Основная часть

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

Эксперимент проводился следующим образом: тестовый датасет случайным образом делился на группы с размерами 16, 35, 60, 120, 240 и 480 строк, используя генератор случайных чисел. Эти группы стали основой для поэтапного анализа на платформах SDV (Synthetic Data Vault), таких как TVAE, CTGAN и GaussianCopula. 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, GaussianCopula) обучалась на подготовленных данных с одинаковыми настройками гиперпараметров для обеспечения сопоставимости результатов.

Метод синтезирования данных CTGAN - 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 году 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".

1. 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.

2. 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 (Тензорный Вариационный Автоэнкодер) – это модель, расширяющая вариационные автоэнкодеры (VAE) для работы с многомерными данными (тензорами), такими как изображения и временные ряды.

Основные компоненты:

1. Кодировщик: преобразует входные данные в латентное пространство, генерируя параметры распределения.

2. Декодировщик: восстанавливает оригинальные данные из латентного представления.

3. 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:

1. Копула: Функция, которая связывает многомерное распределение с его маргинальными распределениями. Гауссовская копула использует многомерное нормальное распределение для описания зависимости между переменными.

2. Маргинальные распределения: это индивидуальные распределения каждой переменной, которые могут быть различными (например, нормальными, экспоненциальными и т. д.).

Процесс синтезирования данных:

1. Определение маргинальных распределений: выбираются распределения для каждой переменной.

2. Параметризация копулы: определяются параметры зависимости (например, корреляционная матрица для гауссовской копулы).

3. 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 используя методы оценки бибилотеки SDV.

Библиотека SDV (Synthetic Data Vault) использует несколько методов для оценки качества синтетических данных:

1. Сравнение распределений: Анализ распределений оригинальных и синтетических данных с помощью визуализаций (гистограммы) и статистических тестов (например, тест Колмогорова-Смирнова).

2. Метрики: Оценка точности с использованием метрик, таких как средняя абсолютная ошибка (MAE) и среднеквадратичная ошибка (RMSE).

Результаты эксперимента

В ходе проведенного численного эксперимента, получены результаты влияния исходных размеров подвыборок и параметра оценки встроенного метода оценки, основанного на среднем показатели качества генерации на качество прогнозной модели. Результаты проведенного исследования показаны в Таблице 1.

Таблица 1. Средняя оценка качества генерации данных

Исходные размеры групп(строк)	**Наименование метода**		
	TVAE	**CTGAN**	**GaussianCopula**
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 GaussianCopula, 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 регресии, модели были обучены на синтетических данных. Первая модель была оучена 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 AUC, обе модели продемонстрировали высокую предсказательную способность и схожую стабильность.

График ROC кривых оригинального датасета представлен на рис 1.

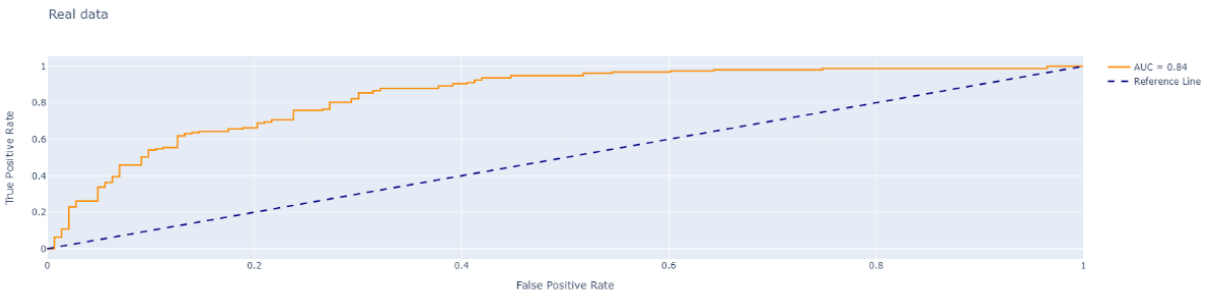

Рис. 1. ROC-кривые оригинльного датасета

График ROC кривых синтетического датасета представлен на рис. 2

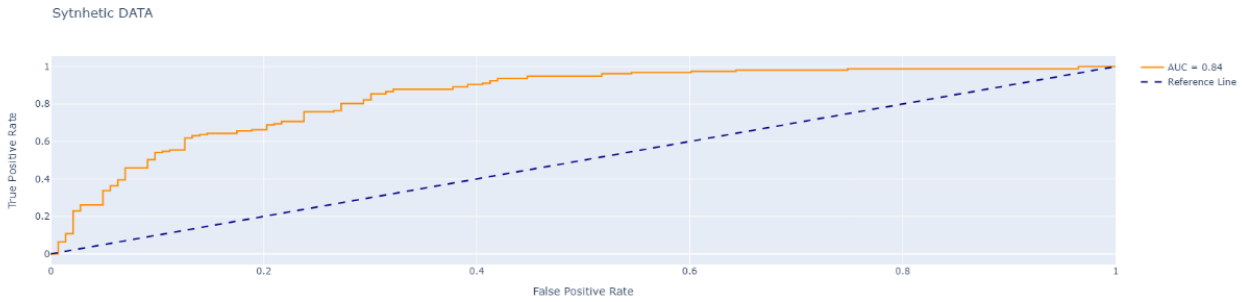

Рис. 2. ROC-кривые синтетического датасета

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 кросс-валидации:

- Для модели, построенной на синтетических данных результат:
- 0.74 accuracy with a standard deviation of 0.03
- Для модели, а оригинальных данных
- 0.85 accuracy with a standard deviation of 0.05

На основании представленных данных можно сделать вывод, что показатели являются приемлемыми. Точность модели на оригинальных данных (0.85) значительно превышает уровень 0.74 на синтетических данных, что указывает на более высокое качество и 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.

В заключение, 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. Авторы уверены, что дальнейшие исследования в этой области приведут к значительным прорывам и улучшениям в медицинской практике и научных исследованиях.

Список использованных источников

1. Губин Е.И., Котов А.О. Использование современных технологий синтезирования данных для медицинских исследований // Перспективы науки. – 2025. – № 3. – С. 10-12.

2. SDV Documentation. – [Электронный ресурс]. – URL: sdv.dev (дата обращения: 15.01.2025).

3. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными – М: 2016. – 2017. – С. 50-120, 135-196.

4. Документация Scikit-learn. – [Электронный ресурс]. – URL: scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата обращения: 15.01.2025).