

УДК 519.63

Анализ и оценка математических моделей синтеза медицинских данных

Сайфетдинов С.Ф.¹, Мамедова Т.Ф.¹

Национальный исследовательский Мордовский государственный университет¹

Аннотация: В исследовании рассматривается проблема дефицита данных в клинической практике и оцениваются методы их искусственного расширения. Сравнение классического алгоритма SMOTE с современными нейросетевыми подходами при работе со сверхмалыми выборками. Тестирование проводилось с использованием строгого протокола валидации (обучение на синтетических данных, тестирование на реальных). Результаты показывают, что табличные автокодировщики наиболее точно воспроизводят внутреннюю структуру медицинских данных, тогда как SMOTE максимизирует точность прогнозов, но может искажать исходное распределение.

Ключевые слова: обогащение данных, генеративно-состязательные сети, малые выборки, машинное обучение, синтез данных.

1. Введение

Применение машинного обучения в медицине часто сталкивается с серьезным ограничением: острой нехваткой данных для обучения моделей. На практике исследователи нередко имеют дело со сверхмалыми выборками, где количество наблюдений может быть менее 50. В таких условиях классические алгоритмы неизбежно переобучаются, запоминая данные конкретных пациентов вместо поиска общих закономерностей.

Для решения этой проблемы применяются методы искусственного увеличения данных. Исторически стандартом здесь выступает алгоритм SMOTE, который создает новые примеры путем линейной интерполяции. Однако в медицинском контексте такой прямолинейный подход может генерировать физиологически невозможные показатели. Логичной альтернативой становятся современные архитектуры глубокого обучения — табличные вариационные автокодировщики и состязательные сети [1, 2], способные улавливать более сложные, скрытые взаимосвязи между признаками.

2. Методология и результаты исследования

С целью определения преимуществ того или иного подхода для анализа медицинских данных, мы провели сравнительный анализ. В качестве тестовой базы использовались клинические показатели пациентов.

Для получения максимально объективных результатов оценка моделей проводилась по строгому протоколу тестирования. Мы применили подход, при котором прогностическая модель обучается исключительно на синтезированных данных, а проверяется на реальных (схема TSTR). Для исключения утечки информации этот процесс был встроен в алгоритм перекрестной проверки с исключением по одному (LOOCV).

Проведенные эксперименты выявили следующую закономерность. Алгоритм SMOTE

продемонстрировал самую высокую предсказательную силу при классификации. С другой стороны, модель TVAE оказалась значительно эффективнее в сохранении оригинальной структуры и связей между клиническими признаками. Что касается CTGAN, то на столь малом объеме данных этот метод оказался наименее стабильным из-за того, что его внутренние механизмы слишком быстро переобучались.

Таким образом, выбор метода зависит от конечной цели. Если критически важно сохранить естественную структуру и взаимосвязи клинических данных, следует отдать предпочтение TVAE. Если же главной задачей является максимальная точность классификации, целесообразно использовать SMOTE, осознавая при этом риски появления аномальных синтетических примеров.

Литература

1. Xu L. et al. Modeling Tabular Data using Conditional GAN // NeurIPS. 2019.
2. Yadav P. et al. Rigorous Experimental Analysis of Tabular Data Generated using TVAE and CTGAN // IJACSA. 2024.

MSC 62R07

Analysis and evaluation of mathematical models for medical data synthesis

S.F. Saifetdinov¹, T.F. Mamedova¹

National Research Mordovia State University¹

Abstract: This study addresses the issue of data scarcity in clinical practice and evaluates methods for their artificial expansion. A comparison of the classic SMOTE algorithm with modern neural network approaches is conducted when working with ultra-small samples. Testing was conducted using a strict validation protocol (training on synthetic data, testing on real data). The results show that tabular autoencoders most accurately reproduce the internal structure of medical data, whereas SMOTE maximizes predictive accuracy but may distort the original distribution.

Keywords: data enrichment, generative adversarial networks, small samples, machine learning, data synthesis.