УДК 004.85

Генерация табличных данных на основе мостов Шрёдингера с использованием GBDT-аппроксиматоров дрейфа

Карташов И. О. (ИТМО), Лопатин И. А. (ИТМО)
Научный руководитель – кандидат физико-математических наук, доцент Деева И. Ю.
(ИТМО)
igor.kartash2019@mail.ru

Введение

Синтез высококачественных табличных данных остается актуальной задачей ввиду гетерогенности признаков и сложных внутренних зависимостей. Существующие подходы на основе GAN и стандартной диффузии (TabDDPM) часто сталкиваются с проблемой компромисса между точностью воспроизведения распределений (fidelity) и прикладной полезностью (utility). В данной работе рассматривается аппарат мостов Шрёдингера (Schrödinger Bridges, SB), который формулирует генерацию данных как задачу энтропийно-регуляризованного оптимального транспорта. Согласно результатам последних исследований, SB-решатели демонстрируют высокую стабильность обучения и точность в сохранении глобальной геометрии данных, однако их эффективность во многом зависит от качества аппроксимации функций дрейфа.

Основная часть

Методологической основой предлагаемого решения является пересмотр архитектурного подхода к аппроксимации оператора дрейфа в рамках итерационного алгоритма IPF (Iterative Proportional Fitting). В классических реализациях мостов Шрёдингера функции дрейфа параметризуются непрерывными нейросетевыми моделями (преимущественно MLP). Однако такие модели обладают индуктивным смещением (inductive bias) в сторону гладкости, что препятствует адекватному моделированию табличных данных, характеризующихся резкими изменениями плотности распределения и категориальной разнородностью.

Ключевая инновация состоит в интеграции ансамблевых методов на основе решающих деревьев в качестве непараметрических аппроксиматоров дрейфа. В отличие от полносвязных нейронных сетей, алгоритмы градиентного бустинга (GBDT) более устойчивы к мультиколлинеарности признаков и эффективно выявляют локальные нелинейные паттерны в условиях ограниченных выборок. Реализация метода предполагает обучение дискретного набора бустинговых регрессоров для различных временных срезов диффузионного процесса. Это позволяет моделировать перенос вероятностной массы через кусочно-постоянную аппроксимацию векторного поля дрейфа, что значительно точнее восстанавливает разрывы в данных.

Использование GBDT позволяет отказаться от сложной нормализации и квантования признаков, сохраняя при этом структурную целостность таблицы. Экспериментальная проверка подтвердила, что предложенная модификация не только снижает расстояние Вассерштейна между реальным и синтетическим распределениями, но и обеспечивает устойчивость к переобучению. Таким образом, гибридизация теории оптимального транспорта и градиентного бустинга обеспечивает синергетический эффект, выраженный в повышении точности моделирования межатрибутных зависимостей.

Выводы

Результаты работы раскрывают значительный научный потенциал интеграции градиентного бустинга в математический аппарат мостов Шрёдингера, закладывая фундамент для создания нового класса гибридных генеративных моделей. Предложенный переход к

GBDT-аппроксиматорам не только обеспечивает прецизионное сохранение межатрибутных зависимостей в режиме TSTR (Train on Synthetic - Test on Real), но и демонстрирует исключительную устойчивость алгоритма при моделировании распределений со сложной нелинейной геометрией.

Практическая значимость и гибкость метода открывают широкие горизонты для дальнейших исследований. Представляется крайне перспективным развитие данного направления в контексте разработки алгоритмов «управляемого дрейфа» для генерации данных с заданными свойствами, а также адаптация транспортных методов к задачам синтеза в условиях экстремального дефицита обучающих выборок, где классические нейросетевые подходы достигают предела своей эффективности.

Литература

1. Диффузионный мост Шрёдингера с приложениями к генеративному моделированию на основе скоринга: Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling / Де Бортоли В., Торнтон Д., Хенг Д., Дусе А. // arXiv. – Текст: электронный. – 2021. – URL: https://doi.org/10.48550/arXiv.2106.01357 (дата обращения 18.02.2026).

2. Сопоставление диффузионных мостов Шрёдингера: Diffusion Schrödinger Bridge Matching / Ши Ю., Де Бортоли В., Кэмпбелл Э., Дусе А. // arXiv. – Текст: электронный. – 2023. – URL: https://doi.org/10.48550/arXiv.2303.16852 (дата обращения 18.02.2026).

3. TabDDPM: моделирование табличных данных с использованием диффузионных моделей / TabDDPM: Modelling Tabular Data with Diffusion Models / Котельников А., Баранчук Д., Рубачев И., Бабенко А. // arXiv. – Текст: электронный. – 2023. – URL: https://doi.org/10.48550/arXiv.2209.15421 (дата обращения 18.02.2026).

4. Моделирование табличных данных с использованием условных GAN: Modeling Tabular Data using Conditional GAN / Сюй Л., Скулариду М., Куэста-Инфанте А., Вирамачанени К. // arXiv. – Текст: электронный. – 2019. – URL: https://doi.org/10.48550/arXiv.1907.00503 (дата обращения 18.02.2026).

Карташов И. О. _____
Деева И. Ю. _____