УДК 004.852

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ КОДИРОВОК КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ ДЛЯ СОВРЕМЕННЫХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ТАБЛИЧНЫХ ДАННЫХ

Чуйко А. И.[1], Ивасенко И. Д.[1]
Научный руководитель – канд. физико-математических наук, доцент Деева И. Ю.[1]
[1]Университет ИТМО
ideeva@itmo.ru

Работа выполнена в рамках темы НИР №625134 «Исследование и разработка фронтирных методов искусственного интеллекта и их приложений».

Введение

Табличные данные широко используются в прикладных задачах и, как правило, содержат значимую долю категориальных и дискретных признаков. Генеративные модели табличных данных формируют непрерывные внутренние представления, и выбранная кодировка категорий существенно влияет на качество синтетики: неудачная параметризация может увеличивать размерность, разреженность и искажать распределения. Для предсказательных моделей выполнялись масштабные бенчмарки кодировок [1], однако для генеративных моделей такие рекомендации остаются ограниченными. Цель работы – систематически оценить влияние методов кодирования категориальных признаков на качество синтетических данных, генерируемых пятью моделями (CTGAN, TabDDPM, TVAE, TabPFN, WGAN), и сформировать рекомендации по их выбору для различных характеристик данных (размер выборки, кардинальность и дисбаланс категорий) [1–3].

Основная часть

Исследование проводится на 26 табличных наборах данных, различающихся долей категориальных, дискретных и непрерывных признаков, а также кардинальностью, степенью дисбаланса категорий и объёмом выборки. Сравниваются 12 способов представления категориальных признаков: one-hot, ordinal, frequency, sum, polynomial, helmert, backward difference, binary, hash encoding, similarity encoding, Gumbel–Softmax relaxation и GEL embedding. В качестве генеративных моделей выбраны CTGAN, TabDDPM, TVAE, TabPFN и WGAN. Эксперимент выполняется по единому протоколу:

- Числовые признаки нормализуются с помощью StandardScaler, все преобразования обучаются только на тренировочной части и затем применяются к тестовой.
- Для каждой модели на референсном разбиении (80/20) с помощью Optuna подбираются гиперпараметры, минимизирующие расстояние Вассерштейна между реальными и синтетическими данными.
- Финальная оценка проводится в цикле 5-кратной кросс валидации с фиксированными оптимальными гиперпараметрами.

Качество синтетических данных оценивается по трём группам метрик:

- Сходство распределений: расстояние Вассерштейна, средняя маргинальная KL-дивергенция, максимальное среднее расхождение (MMD).
- Сохранение взаимосвязей: норма Фробениуса разности корреляционных матриц, вычисленных в исходном пространстве признаков.

– Практическая полезность: сценарий TSTR (Train on Synthetic, Test on Real) с использованием CatBoostRegressor; вычисляется относительное отклонение коэффициента детерминации синтетических данных от эталонного значения на реальных данных.

Анализ направлен на выявление зависимостей эффективности кодировок от кардинальности, степени дисбаланса категорий и размера выборки для каждого семейства генеративных моделей.

Выводы

В работе выполняется систематическое сравнение методов кодирования категориальных признаков для пяти современных генеративных моделей табличных данных на 26 наборах данных. Оценка производится по комплексу метрик, характеризующих fidelity, структуру и utility синтетических выборок. По итогам экспериментов планируется построить «карту режимов» — рекомендации по выбору оптимального способа кодирования в зависимости от свойств данных (кардинальность, дисбаланс, размер выборки) и типа генеративной модели. Окончательные выводы и количественные результаты будут представлены после завершения полного цикла экспериментов.

Литература

1. Clerici F., Nobani N. Categorical variable encoding methods for tabular data: a benchmarking study // Int. J. Data Sci. Anal. – 2026. – doi:10.1007/s41060-025-00886-w.
2. Kotelnikov A., Baranchuk D., Rubachev I., Babenko A. TabDDPM: Modelling Tabular Data with Diffusion Models // arXiv. – 2022. – arXiv:2209.15421.
3. Xu L., Skoularidou M., Cuesta-Infante A., Veeramachaneni K. Modeling Tabular Data using Conditional GAN // NeurIPS. – 2019. – arXiv:1907.00503.