УДК 004.852

КАЛИБРОВКА СИНТЕТИЧЕСКИХ ТАБЛИЧНЫХ ДАННЫХ ДЛЯ КОРРЕКТНОГО ВЫБОРА МОДЕЛЕЙ

Филатов Г. В.[1]
Научный руководитель – канд. физ.-мат. наук, доцент Деева И. Ю.[1]
[1]Университет ИТМО
filatovgwork@gmail.com

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

Введение

По сей день табличные данные остаются одним из ключевых форматов в анализе научных данных, машинном обучении и исследовательских задачах в области медицины [2, 4], экономики и социальных наук [3], однако большинство современных подходов к синтетической валидации разрабатываются преимущественно для задач компьютерного зрения и не адаптированы к табличному домену [4]. При этом классическая схема разделения ограниченного набора реальных данных на обучающую и валидационную выборки приводит к потере ценной информации для обучения моделей и особенно критична при сценариях, когда сбор большого числа данных ограничен экономическими и социальными факторам. Одновременно бурное развитие генеративных моделей делает возможной генерацию практически неограниченного количества синтетических образцов[5], что порождает идею использовать синтетическую выборку как замену реальной валидации при выборе моделей в табличном домене.

Ряд работ по использованию синтетических данных для оценки моделей показывает, что при достаточном качестве генераторов ранжирование по ошибкам на синтетическом множестве может частично сохранять порядок моделей по ошибкам на реальном тесте, однако существующие результаты получены в основном для изображений (CIFAR-10, ImageNet) [1] и не дают готового решения для табличного домена. Более того, без специальной калибровки синтетический домен зачастую смещён относительно реального, что приводит к систематическому искажению оценок качества и потере согласованности ранжирования, особенно при использовании простых генераторов.

Основная часть

В работе предлагается алгоритм калибровки синтетической табличной валидации, позволяющий проводить выбор моделей без выделения отдельной реальной валидационной выборки. На всём доступном реальном тренировочном наборе обучаются предсказательные модели, а роль валидации берёт на себя синтетический датасет, сгенерированный с помощью табличных генераторов (CTGAN, TVAE, TabDDPM, TabPFN, Gaussian Copula). Наиболее важным инструментом подхода является калибровка синтетических наборов с использованием потерь отдельных образцов для согласования ранжирования синтетических и реальных данных.

Для набора моделей, использующихся для калибровки составляется матрица потерь: для каждого синтетического образца рассчитывается величина потерь (логарифмическая потеря, ошибка классификации, среднеквадратичная или средняя абсолютная ошибка) в зависимости от решаемой задачи. Далее решается задача минимизации расхождения между вектором средних реальных потерь и взвешенной комбинации синтетических данных с добавлением L2-регуляризации. Векторы весов синтетических образцов находятся численными методами при ограничении значений интервалом от 0 до 1 для сохранения интерпретируемости и в дальнейшем используется

для перевзвешивания значений потерь новых моделей по их результатам на синтетическом наборе.

Представленная постановка основывается на потерях моделей и не привязана к конкретному типу задачи, что позволяет единообразно обрабатывать и задачи классификации, и регрессии на основе выбранной функции потерь.

Разработанный метод протестирован на пяти табличных классификационных датасетах с портфелем из 15 калибровочных и 29 тестовых моделей, а также на пяти регрессионных датасетах с набором 15 калибровочных и 24 тестовых моделей. В результате после калибровки доля корректно отранжированных моделей для большинства комбинаций генераторов и датасетов возрастает до 50-70%. Кроме того, корреляции Спирмена между ошибками на синтетических и реальных данных, используемый для оценки качества ранжирования потерь, после калибровки для лучших генераторов превышает 0.9, что свидетельствует о высоком согласовании рангов.

Выводы

Был разработан и реализован экспериментальный фреймворк, основанный на методе калибровки синтетических табличных данных по потерям отдельных образцов, позволяющий использовать синтетическую выборку, как надёжную замену реальной валидации при выборе моделей. Предлагаемый подход не требует изменения архитектур моделей и опирается исключительно на оценку качества моделей через функции потерь, благодаря чему легко интегрируется в существующие пайплайны и приводит к существенному росту согласованности ранжирования и корреляции потерь по сравнению с некалиброванными синтетическими данными. Полученные результаты показывают, что при корректной калибровке синтетическая валидация может эффективно заменить реальную и компенсировать дефицит данных при обучении моделей в табличном домене.

Литература

1. Shoshan A., Bhonker N., Kviatkovsky I., Fintz M., Medioni G. Synthetic data for model selection // Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, 2023. P. 31633–31656.
2. Jadon A., Kumar S. Leveraging generative AI models for synthetic data generation in healthcare: Balancing research and privacy // Proc. 2023 International Conference on Smart Applications, Communications and Networking (SmartNets). 2023. P. 1–4.
3. Caliskan H., Yayla O. F., Genç Y. A comparative analysis of synthetic data generation with VAE and CTGAN models on financial credit loan offer data // Proc. 2023 8th International Conference on Computer Science and Engineering (UBMK). 2023. P. 212–217.
4. Hu Q., Yuille A., Zhou Z. Synthetic Data as Validation // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 8726–8737.
5. Xu L., Veeramachaneni K. Synthesizing Tabular Data using Generative Adversarial Networks // arXiv preprint arXiv:1811.11264. 2018.