УДК 004.021

ГЕНЕРАЦИЯ АВТОМАТИЧЕСКОГО БЕНЧМАРКА НА ОСНОВЕ ОШИБОК МОДЕЛИ

Захаров К.А. (ИТМО)
Научный руководитель – доктор технических наук, профессор Бухановский А.В. (ИТМО)

Введение. Растущее число бенчмарков (эталонных наборов данных для оценки качества моделей) в области машинного обучения, ориентированных на конкретную предметную область, привело к методологическому прогрессу [1], однако для внедрения в реальных условиях требуется иной подход к оценке. Во многих реальных внедрениях специалисты-практики уже используют высококачественные модели, которые в среднем работают хорошо [2], но дают сбой в небольшой группе случаев, критически важных для приложений. При оценке моделей машинного обучения, основным требованием является не только повышение общей точности, но и, что более важно, надежное улучшение именно в тех редких случаях, когда происходят сбои. Обычные наборы эталонных данных, предназначенные для определения новых предлагаемых моделей с использованием большого количества размеченных примеров, плохо подходят для этого сценария, поскольку количество задокументированных случаев сбоев для хорошо настроенной модели, по сути, невелико. Поэтому получение достаточного количества высококачественных данных для статистического сравнения качества модели остается сложной задачей. Для решения этой проблемы предлагаются синтетические тесты, основанные на моделях, предназначенные для выявления сбоев в существующих моделях.

Основная часть. В данной работе были решены следующие задачи: 1) формализована задача генерации автоматического бенчмарка для моделей классического машинного обучения; 2) предложен двухэтапный подход для генерации автоматического бенчмарка, основываясь на ответах модели и ее ошибок; 3) проведено экспериментальное исследование применимости автоматического бенчмарка для оценки качества моделей машинного обучения [3].

Алгоритм начинает свою работу с определения «плохих» точек данных, то есть таких точек, где модель ошибается больше всего. При этом, в нашем сценарии, для достаточно хорошей модели, «плохих» точек данных не может быть велико. Эти точки выделяются при помощи оператора отсечения, например, для регрессионной модели таким оператором может быть отсечение по порогу квадрата разности между предсказанием модели и тестовыми точками данных.

После определения «плохих» точек данных запускается алгоритм аугментации данных. Это делается для того, чтобы на следующем этапе, алгоритм аппроксимации плотности этих точек, смог обучиться качественно. Для этого и необходимо сначала обогатить обучающую выборку для генератора. В нашем методе это делается при помощи генетического алгоритма. Генетический алгоритм может принимать как категориальные признаки, так и непрерывные. Операторы скрещивания и мутации вносят случайные изменения в индивидах (признаках) и далее функция оптимизации отбирает такие индивиды, которые имеют наибольшую ошибку прогноза. При этом в функцию ошибки добавляется и регуляризация для того, чтобы избежать критических отклонений в индивидах.

После аугментации получается датасет для обучения генеративной модели, которая выступает аппроксиматором плотности распределения «плохих» точек данных. В качестве генеративной модели используется вариационный автокодировщик, который модифицирован для работы с категориальными признаками. Также, генератор является условным, то есть он может принимать на вход метки классов или зависимую переменную, в случае регрессии. При этом для задач, где нет выделенного таргета, генератор может быть безусловным.

В экспериментальном исследовании были использованы несколько наборов данных и

обучены на них регрессионные модели и модели классификации. В качестве метрик оценки использовались SMAPE, F-мера, ROCAUC и расстояние Вассерштейна. Эксперименты показали, что разработанный алгоритм может качественно дополнять данные, где исходная модель ошибается больше всего, тем самым можно проверять истинное качество модели на редких примерах, где она ошибается.

Выводы. В рамках работы формализована задача генерации автоматического бенчмарка для оценки моделей классического машинного обучения. Был построен двухэтапный конвейер для аугментации и генерации эталонного набора данных по выделенным «плохим» точкам данных. При этом «плохие» точки подбираются также автоматически исходя из предсказаний модели. Экспериментальное исследование разработанного алгоритма показало состоятельность и преимущества такого подхода в оценки качества моделей. Стоит также отметить, что при модификации операторов мутации и скрещивания в генетическом алгоритме, можно адаптировать наш подход для создания бенчмарков в области больших языковых моделей (LLM).

Список использованных источников:

1. Maheshwari, G., Ivanov, D. and Haddad, K.E., 2024. Efficacy of synthetic data as a benchmark. arXiv preprint arXiv:2409.11968.
2. Shi, Q., Tang, M., Narasimhan, K. and Yao, S., 2024. Can Language Models Solve Olympiad Programming?. arXiv preprint arXiv:2404.10952.
3. Zakharov, K. and Boukhanovsky, A., 2025. Model-Aware Automatic Benchmark Generation with Self-Error Instructions for Data-Driven Models. Machine Learning and Knowledge Extraction, 7(4), p.148.