УДК 004.8

ПРИМЕНЕНИЕ МЕТОДОВ ГРАДИЕНТНОГО БУСТИНГА В ФРЕЙМВОРКАХ АВТОМАТИЧЕСКОГО МАШИННОГО ОБУЧЕНИЯ

Харьковской Р.Р. (Университет ИТМО),
Стебеньков А.С. (Университет ИТМО)
Научный руководитель - к.т.н., доцент Никитин Н.О. (Университет ИТМО),

Введение. Автоматическое машинное обучение (AutoML) – инструмент, который позволяет полностью или частично автоматизировать применение методов машинного обучения к реальным задачам. Во многих популярных AutoML фреймворках уже реализовано использование градиентного бустинга для построения пайплайнов ML. Градиентные бустинговые методы над решающими деревьями – высокоэффективный и широко используемый метод машинного обучения. Актуальность этих алгоритмов подчеркивается их способностью строить нелинейные модели машинного обучения, объединяя слабые обучающие алгоритмы в последовательную композицию. Примерами успешных реализаций алгоритмов градиентного бустинга являются XGBoost, LightGBM и CatBoost, которые стали стандартом во многих задачах анализа данных. Их популярность объясняется высокой точностью прогнозов, возможностью работы с различными необработанными табличными данными, содержащими как числовые, так и категориальные признаки[1].

В данном тезисе рассмотрено использование методов градиентного бустинга в фреймворках автоматического машинного обучения. Данные системы наиболее полезны для тех пользователей, у которых нет большого опыта в области машинного обучения, однако перед ними стоит задача построения высококачественной модели.

Основная часть

В AutoML фреймворке FEDOT требуется реализация самых успешных алгоритмов градиентного бустинга, таких как XGBoost, LightGBM и CatBoost. На данный момент они реализованы на первой итерации, поэтому требуются дальнейшие улучшения в работе данных моделей. В результате составления пайплайна фреймворк определит наиболее подходящий алгоритм, основываясь на следующих характеристиках каждой из моделей:

1. XGBoost – фреймворк, реализованный как исследовательский open-source проект. Преимуществами данного фреймворка является возможность работы с необработанными данными, производительность, эффективная оптимизация вычислений, наличие нескольких стратегий построений деревьев и настраиваемая регуляризация для избежания переобучения. Однако данный фреймворк не имеет встроенного метода для кодирования категориальных признаков.
2. LightGBM – фреймворк, реализованный компанией Microsoft. Преимуществами алгоритма является скорость работы, эффективное использование памяти (самый легковесный из тройки фреймворков). Эффективность достигает за счет внедрения методов GOSS[2] и EFB[2]. Также LightGBM хорошо работает при обучении на больших наборах данных, однако может легко переобучиться на небольших датасетах.
3. CatBoost – фреймворк, реализованный компанией Яндекс. Заметным улучшением CatBoost является его способность выполнять несмещенную

оценку градиента, которая уменьшает переобучение. Особенностью CatBoost является автоматическое преобразование категориальных признаков в числовые [1]. Однако данный фреймворк обладает медленной реализацией на CPU, а также большими затратами оперативной памяти.

Описанные ранее фреймворки затрагивают многие возможные проблемы (дисбаланс классов, наличие категориальных признаков, переобучение), поэтому достаточно часто именно они используются для построения высококачественной модели, созданной с помощью автоматического машинного обучения.

В контексте сравнительного анализа, проведенного с фреймворками AutoGluon и LightAutoML, были выявлены определенные различия в метриках бустинговых моделей. Анализ результатов показал, что некоторые показатели эффективности данных моделей в FEDOT оказались ниже по сравнению с конкурирующими фреймворками.

Эти наблюдения могут быть обусловлены различиями в подходах к реализации бустинговых алгоритмов, а также особенностях внутренних механизмов работы фреймворков. Понимание и анализ этих различий представляют важный аспект для дальнейшего улучшения и развития FEDOT с целью достижения более высоких результатов.

Выводы

В ходе исследования применения бустинговых методов на деревьях, таких как XGBoost, LightGBM и CatBoost, в AutoML фреймворках, были выявлены ключевые характеристики и преимущества каждого из этих алгоритмов. В ходе исследования были выявлены различия в метриках при сравнении FEDOT с AutoGluon и LightAutoML, поэтому дальнейшая работа была направлена на оптимизацию бустинговых моделей внутри FEDOT. Это включает в себя проведение дополнительных исследований, улучшение использования алгоритмов и улучшение гиперпараметров с целью обеспечения более конкурентоспособных результатов в сравнении с аналогичными инструментами автоматического машинного обучения.

Список использованных источников:

1. Mienye I. D and Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects, in IEEE Access, 2022, vol. 10, pp. 99129-99149, doi: 10.1109/ACCESS.2022.3207287.
2. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, pages 3149–3157. Curran Associates Inc., 2017

Харьковской Р.Р. (автор) Подпись

Никитин. Н.О. (научный руководитель) Подпись