

Предсказание вероятности наличия колоректального рака с использованием LLM в условиях недостатка данных.

Курочкина Д.Н. (МФТИ).

Научный руководитель — Егоров К.С. (руководитель направления по анализу данных, Лаборатория ИИ, Сбер)

Введение. Своевременная диагностика онкологических заболеваний может значительно улучшить прогноз и шансы на выживание, поскольку может способствовать своевременному лечению пациентов за счет соблюдения сроков лечения. В современном мире способы предсказания вероятности развития рака являются предметом многочисленных исследований [1]. Последние годы набирает популярность подобный анализ с помощью языковых моделей [2].

Целью данной работы является использование больших языковых моделей для предсказания вероятности наличия колоректального рака у пациента на основании записей электронных медицинских карт из датасета MIMIC-IV в условиях недостатка данных.

Основная часть.

В качестве источника данных использовался открытый датасет MIMIC-IV v. 2.2. Для работы с датасетом были написаны SQL-скрипты, с помощью которых выбраны ключевые параметры для анализа. В качестве ключевых параметров на основании литературных данных были выбраны уровень железа, ферритина, гемоглобина, тромбоцитов и жалобы пациента [3]. Из датасета выбраны все пациенты с диагнозом колоректальный рак — 1176 пациентов, а так же пациенты без диагноза колоректальный рак общим числом до 4000 пациентов. В качестве языковой модели выбраны модели с открытым исходным кодом 'llama3.3:70b' и 'qwen2.5:72b'. Для обработки данных был написан промпт, записывающий предсказание в отдельный столбец. Далее рассчитана точность (accuracy), которая для llama составила 0.64, а для qwen — 0.72. F1 score как для llama так и для qwen составил 0.51. ROC AUC Score для llama составил 0.64, для qwen - 0.66. Примечательно, что в исходном датасете отсутствовали данные в части записей, что позволило оценить точность в зависимости от количества отсутствующих записей (ключевых параметров).

Выводы.

По всему датасету точность (accuracy) для llama составила 0.64, а для qwen — 0.72. Наибольшая точность для llama — 0.68, а для qwen — 0.79 была зафиксирована в случаях, когда присутствовали данные о 2 записях из 5. Таких записей было больше всего — 2711. Вопреки ожиданиям, при наличии всех записей у пациента точность предсказания llama составила всего 0.44, что не является максимальным значением, это может быть связано с тем, что в датасете было всего 32 записи у которых присутствовало все 5 рассматриваемых для предсказания параметров. Минимальное значение точности для qwen составило 0.51, в случае отсутствия 4 значений из 5 (588 записей) и в случае отсутствия 5 значений из 5 (37 записей).

При сравнении F1 score в условиях дефицита данных, наибольшее значение продемонстрировала qwen — 0.59 в случае наличия 2 параметров из 5, в аналогичных условиях для llama значение F1 score составило 0.56. Наибольшее значение F1 score — 0.59 - llama показала при наличии 3 значений из 5. В аналогичных условиях для qwen значение F1 score составило 0.55. Ожидаемо минимальное значение F1 score как для llama так и для qwen равно 0 при отсутствии 5 из 5 значений, тк в таких условиях предсказание силы не имеет.

При сравнении метрики ROC-AUC в условиях дефицита данных, как qwen так и llama показали результат 0.73 при наличии 2 значений из 5, а при отсутствии 5 значений из 5 метрика составила 0.5, что не имеет предсказательной силы.

В целом, исходя из полученных данных, языковая модель 'qwen2.5:72b' демонстрирует более корректные предсказания нежели 'llama3.3:70b'.

Список использованных источников:

1. Zhang B., Shi H., Wang H. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach // Journal of Multidisciplinary Healthcare. 2023. (16). С. 1779–1791.
2. Wang W., Jin X. Prostate cancer prediction model: A retrospective analysis based on machine learning using the MIMIC-IV database // Intelligent Pharmacy. 2023. № 4 (1). С. 268–273.
3. Hamilton W. [и др.]. The importance of anaemia in diagnosing colorectal cancer: a case–control study using electronic primary care records // British Journal of Cancer. 2008. № 2 (98). С. 323–327.