Выявление закономерностей в клинических данных: исследование роли больших языковых моделей и алгоритмов кластеризации

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1737530

NLP | precision medicine | анализ данных | диагностика | клинические данные

Цель

Большие языковые модели (LLM — Large Language Models) продемонстрировали исключительные показатели в обработке естественного языка, однако их полезность в анализе структурированных клинических данных остается относительно малоизученной. Данное пилотное исследование исследует, могут ли эмбеддинги, сгенерированные LLM, сохранять структурную целостность клинических наборов данных и улучшать предиктивное моделирование, особенно в условиях ограниченных ресурсов.

Методы

Мы применили методы снижения размерности, такие как метод главных компонент (PCA — Principal Component Analysis), t-распределенное стохастическое вложение соседей (t-SNE — t-distributed Stochastic Neighbor Embedding) и кластеризация k-средних (k-means), для сравнения исходных структур данных со структурами, полученными из эмбеддингов LLM. Метрики оценки включали косинусное сходство, площадь под кривой (AUC — area under the curve) и R^2, примененные к 100 синтетическим наборам данных и

двум реальным клиническим наборам данных: базе медицинских данных UCI и записям пациентов с эндокардитом. Мы оценили несколько архитектур LLM, включая BERT, RoBERTa, Llama 2 и E5-small, сосредоточившись на предиктивной точности и вычислительной эффективности.

Результаты

Эмбеддинги LLM тесно воспроизводили исходные структуры данных: BERT достиг косинусного сходства 0,95 на линейных наборах данных, а Llama 2 (30B) — 0,85 на квадратичных наборах данных, хотя и с более высокими вычислительными затратами. Предиктивная производительность значительно улучшилась во всех случаях с увеличением отношения переменных субъекта (SVR — subject variable ratio). Были выявлены три группы с аналогичной производительностью, которые лучше помогали и значительно лучше помогали. Эти группы различались в зависимости от уравнения, использованного для генерации синтетических данных.

Обсуждение

Эти результаты подчеркивают потенциал LLM для улучшения анализа структурированных данных путем выявления оптимальных условий, таких как пороги SVR, для их практического использования. Также подчеркивается компромисс между вычислительной стоимостью и производительностью в различных архитектурах LLM, что указывает на необходимость выбора модели, специфичной для контекста.

Заключение

LLM могут быть эффективно использованы для повторного использования существующих клинических наборов данных для индивидуальных клинических вопросов, таких как оптимизация времени хирургического вмешательства для пациентов с инфекционным эндокардитом и эмболическим инсультом. Этот подход продвигает персонализированную медицину и поддерживает принятие клинических решений на основе данных.