УДК 004.89

ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ ДЛЯ ОБЪЯСНЕНИЯ МОДЕЛЕЙ ЗДРАВООХРАНЕНИЯ

Жукова И.В. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н., доцент Ковальчук С.В.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе изучены существующие модели в области здравоохранения и медицины на основе публикаций за последние 5 лет, рассмотрены возможности и проблемы, связанные с ними. Также в докладе рассматриваются пути преодоления ограничений использования закрытых моделей в практических проблемах медицины и здравоохранения.

Введение. Искусственный интеллект (ИИ) - широко распространенная и популярная технология, которая используется в различных областях нашей жизни. Следовательно, распространение ИИ оказывает значительное влияние на общество. Действительно, ИИ уже стал вездесущим, и мы привыкли к тому, что модели принимают решения за нас в повседневной жизни: предложение друзей в социальных сетях, рекомендации фильмов, реклама на основе предпочтений и интересов пользователя. В этих областях мы можем использовать модели черного ящика, и это не будет проблемой, потому что мы можем оценить результат без объяснения причин.
Однако в таких областях, как медицина и здравоохранение, при принятии решений, например, при диагностике заболевания или назначении лекарств, важно знать причины решения, которое приняла модель.
Проблема заключается в том, что, хотя алгоритмы ИИ кажутся мощными с точки зрения результатов и прогнозов, они страдают от непрозрачности - это затрудняет понимание их внутреннего механизма работы. Поскольку принятие важных решений для системы, которая не может себя объяснить, представляет очевидную опасность, объяснимость моделей ИИ является актуальным и важным направлением исследований.

Основная часть. По результатам обзора источников были выделены существенные проблемы, с которыми сталкивается сектор здравоохранения при применении моделей искусственного интеллекта. В работе рассматриваются следующие проблемы и пути их решения:

- объяснимость;
- интерпретируемость;
- валидируемость;
- качество данных;
- приватность данных;
- изменчивость во времени.

Более подробно в работе была рассмотрена концепция объяснимого искусственного интеллекта. Для решения проблемы интерпретируемости, объяснимый ИИ предлагает набор методов, которые делают модель более прозрачной. Цель объяснимого ИИ - сделать предсказания модели понятными для человека, сохранив при этом высокую производительность системы. Существует два основных подхода к объяснимому ИИ: Ante-hoc и Post-hoc.
Ante-hoc методы подразумевают объяснимость модели с самого начала, система спроектирована таким образом, что объяснимость встроена в модель искусственного

интеллекта. В этих системах могут использоваться модели, разработанные в виде прозрачных ящиков. Типичными примерами являются линейная регрессия или деревья принятия решений. В моделях Post-Hoc объяснимость включается после разработки модели, то есть обучение осуществляется обычным способом, а объяснимость включается только при тестировании. Такая техника приобрела популярность, так как уже существует множество моделей, которые нуждаются в объяснимости.
Также в работе рассматривается концепция причинно-следственной связи, которая утверждает, что объяснимость недостаточна для моделей здравоохранения. Для достижения успеха необходимо выделить причинно-следственные связи, которые включают в себя измерение качества объяснимости. Для достижения этой цели необходимо разделить объясняемую модель и интерфейс объяснения.

Выводы. В работе были рассмотрены существующие проблемы, возникающие при использовании моделей искусственного интеллекта в моделях здравоохранения и медицины, а также пути их преодоления.
Дальнейшая работа будет заключаться в том, чтобы применить концепции объяснимого искусственного интеллекта и причинно-следственных связей для объяснения результатов работы моделей, а также оценки качества предоставляемых объяснений.

Жукова И.В. (автор)

Ковальчук С.В. (научный руководитель)