

Внедрение алгоритмов машинного обучения в анализ геномных и феномных данных в области психиатрии

Нургалиев Т.И.

Лаборатория Нейробиологии и Молекулярной Фармакологии, Институт Трансляционной Биомедицины

Резюме. В данной статье кратко описаны современные методы анализа данных в области психиатрии с использованием машинного обучения, а также обозначены возможные перспективы и существующие ограничения этого подхода.

Ключевые слова: машинное обучение, обучение с учителем, обучение без учителя, анализ геномных данных.

Machine learning methods for analyzing genomic and phenomic data in psychiatry

Nurgaliev T.I.

Laboratory of Neuroscience and Molecular Pharmacology, Institute of Translational Biomedicine, Saint-Petersburg State University

Summary. This review briefly describes modern approaches of data analysis in psychiatry using machine learning and gives possible prospects and common obstacles of this approach.

Keywords: machine learning, supervised learning, unsupervised learning, genomics data analysis

Геномные и феномные данные, полученные в результате масштабных исследований психических заболеваний, представляют собой гетерогенные многомерные данные, которые сложно анализировать, используя лишь подходы и методы классической статистики. Наука о данных в настоящее время включает также альтернативный подход к исследованию механизмов сложных комплексных заболеваний — использование алгоритмов машинного обучения, которое уже продемонстрировало высокую эффективность в исследованиях в области психиатрии и потенциально способно обеспечить переход к персонализированной медицине в клинической практике.

Обучение с учителем (деревья принятия решений, K-ближайших соседей, метод опорных векторов, некоторые виды нейронных сетей и др.) — это один из широко распространенных типов задач машинного обучения, в ходе которого алгоритм обучается на парах «стимул-реакция». Результатом имплементации алгоритма служит способность прогнозировать «реакцию» по «стимулу» для незнакомого объекта в результате восстановления искомой зависимости. Подобные алгоритмы могут быть использованы для предсказания наиболее эффективного фармакологического агента и его дозировки для лечения конкретного пациента на основе его геномной информации и/или совокупности фенотипов (сканов МРТ и измерений по клиническим опросникам). Обучение с учителем также продемонстрировало высокую чувствительность в качестве диагностиче-

ского инструмента, в частности для ранней диагностики психических расстройств. Алгоритмы этого типа также способны предсказывать траекторию развития заболевания: переход в хроническую форму, быстрый тип развития или быструю ремиссию, вероятность рецидивов, тяжесть протекания заболевания и качество жизни пациента. Особого внимания также заслуживают регрессионные модели машинного обучения, в частности лассо-регрессия, наиболее точно оценивающая полигенный показатель риска из всех существующих в настоящее время методов.

Среди задач обучения с учителем наибольшего внимания в контексте психиатрии заслуживает задача кластеризации (метод ближайших соседей, алгоритмы DBSCAN и Affinity Propagation, часть нейронных сетей). Задача кластеризации сводится к разбиению выборки на кластеры — подмножества, состоящие из сходных объектов. Способность алгоритмов этого типа к агностическому разбиению выборки пациентов на подгруппы, используя скрытые паттерны данных, делает их перспективным инструментом для пересмотра классификации болезней на основе сочетания молекулярных механизмов заболеваний с их клиническими фенотипами.

Следует также отметить ограничения на применение методов машинного обучения в психиатрических исследованиях. Наиболее важным из них является размер выборки, который наиболее критичен в случае глубокого обучения. Другим ограничивающим моментом также является разреженность данных — большинство алгоритмов не способны работать с данными, в которых есть какие-либо пропуски.

Сведения об авторах

Нургалиев Тимур Ильверович — лаборант-исследователь, лаборатория Нейробиологии и Молекулярной Фармакологии, Институт Трансляционной Биомедицины, Санкт-Петербургский Государственный Университет, магистрант программы «Биоинформатика», Санкт-Петербургский Государственный Университет. E-mail: hmtimurhm@gmail.com