SHAPLEY-ЗНАЧЕНИЕ КАК МЕРА КАЧЕСТВА МАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ

Заворотнюк Д., Сорокин А., Пеков С.[1], Елиферов В., Бочаров К.[2], Николаев Е.[1], Попов И.

Московский физико-технический институт, Институтский переулок, д.9, Долгопрудный, 141701, Россия, +7 (495) 408 45 54, E-mail: info@mipt.ru
[1]Сколковский институт науки и технологий, Большой бульвар д.30, стр.1, Москва, 121205, Россия, +7 (495) 280 14 81, E-mail: inbox@skoltech.ru
[2]Институт энергетических проблем химической физики им. В.Л. Тальрозе при Федеральном исследовательском центре химической физики им. Н.Н. Семенова, Ленинский проспект, д. 38, к. 2, Москва, 119334, Россия, +7 (499) 137 82 58

Наличие в выборке низкокачественных данных и данных, соответствующих смешанным случаям, когда содержат линейные комбинации объектов из разных классов, приводит к резкому ухудшению качества полученных предсказаний. Поэтому существует необходимость построения алгоритмов автоматического контроля качества входных данных.

Для решения задачи было предложено использовать алгоритм Shapley [1]. Мы реализовали этот алгоритм на примере масс-спектрометрических данных, полученных с образцов тканей головного мозга человека пациентов с диагнозами глиобластома и патология не опухолевой природы. Для масс-спектрометрии с прямой ионизацией характерны такие особенности, как слабая воспроизводимость молекулярного профиля, нестабильный ионный ток, особенно, когда в ходе проведения эксперимента происходит переключение режима сбора ионов. Это приводит к тому, что даже соседние сканы могут сильно различаться между собой, и сильно влияет на анализ экспериментальных данных.

Для небольшого набора масс-спектрометрических сканов были рассчитаны Shapley-значения, которые в дальнейшем были использованы для построения регрессионной модели и предсказания Shapley-значений для остальных сканов. Результаты показывают, что исключение масс-спектрометрических сканов с низкими Shapley-значениями из анализа повышает точность классификационных моделей и уменьшает время, необходимое для построения моделей.

Работа выполнена в рамках Государственного задания Министерства науки и высшего образования (соглашение No 075-00337-20-02, проект No 0714-2020-0006).

Литература

1. Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning // 36th International Conference on Machine Learning, ICML 2019, 2019-June, 4053–4065.