Сочетание алгоритмов предобработки данных, машинного обучения и отбора переменных для выбора оптимального признакового пространства в классификации биообразцов методом хроматомасс-спектрометрии

И.В. Плющенко*, И.А. Родин

Московский государственный университет имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д.1, стр. 3.

plyush1993@bk.ru

В последнее десятилетие активно развиваются подходы к выявлению признаков для классификации многокомпонентных смесей. Такие технологии принято называть «омиксными». Типичным примером может служить метаболомика, в рамках которой выявляются потенциальные биомаркеры заболеваний и/или изучается обмен веществ. Наиболее распространенной аналитической платформой следует считать сочетание хроматографических методов разделения с масс-спектрометрическим детектированием. Подобные гибридные техники порождают огромные массивы информации с сотнями тысяч и миллионами единиц данных и тысячами и десятками тысяч признаков. Выбор наиболее информативных переменных (аналитических сигналов), позволяющих полностью охарактеризовать систему для однозначной и правильной классификации является нетривиальной задачей.

Предложен эвристический подход, основанный на последовательном применении нескольких этапов обработки данных. На первой стадии проводят фильтрацию признаков с наименьшей ожидаемой информативностью посредством дисперсионного анализа. Затем проводится коррекция интенсивности на основе сочетания линейной регрессии с сингулярным векторным разложением остатков. Для дальнейшего отбора переменных проводили настройку моделей машинного обучения, оставляли только важные переменные и определяли оптимальный набор наиболее информативных признаков методом рекурсивного отбора. Качество набора переменных проверяли проекционными и кластерными методами. В ряде случаев удается снизить признаковое пространство на несколько порядков. Были проанализированы несколько наборов данных, в том числе из открытых репозиториев.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-33-90071 Аспиранты