Секция 4. ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

УДК 57.087

Применение ансамбля методов проекции на латентные структуры в задаче анализа пептидных микрочипов

Д.С. Анисимов, М.А. Рязанов, А.И. Шаповал
АлтГУ, Барнаул

В работе рассматривается применение мета-алгоритма машинного обучения в задаче классификации участников медицинского эксперимента. Теоретическое описание и исходные данные эксперимента рассмотрены в работах [1–3]. К результатам эксперимента были применены методы уменьшения размерности [4]. Однако одним из существенных недостатков вышеприведённых методов является маленькая обучающая выборка. Для решения данной проблемы предлагается использовать мета-алгоритм bagging, в основе которого лежит создание ансамбля моделей, каждая из которых обучается на своём «bootstrap» [5].

Рассмотрим мета-алгоритм bagging. Пусть n – размер исходной выборки, m – количество подвыборок, которое нужно получить. Каждая из m подвыборок также имеет размер n и строится на основании исходной методом изъятия с возвращением, то есть каждый объект исходной выборки имеет вероятность p=1/n быть добавленным в подвыборку на i-й итерации её построения независимо от того, был ли он добавлен туда на предыдущих итерациях. Вероятность того, что объект не попадёт в некоторую подвыборку определим так:

$$p_{oob} = (1-\frac{1}{n})^n \xrightarrow[n\to\infty]{} \frac{1}{e} \approx 0,368,$$

то есть, при больших n примерно 37% объектов не попадут в bootstrap и их можно использовать для проверки модели. Данный способ проверки известен как out-of-bag estimation [6]. Далее на каждой из m подвыборок обучается модель, в данном случае регрессия на латентные структуры. Затем для тестового объекта находим отклик в каждой модели, после чего производим агрегирование откликов усреднением (в случае регрессии) или голосованием (в случае классификации).

Для экспериментов использовались данные пептидных микрочипов (330K) имеющих на своей поверхности 330 тысяч пептидов. В качестве предварительной обработки использовались исследованные ранее методы [2, 3]. В частности, данные предварительно логарифмирова-

лись по основанию 2, затем подвергались медианной нормализации для подавления отклонений фонового свечения различных чипов.

Результаты перекрёстной проверки, в ходе которой на каждой итерации в качестве тестовой выборки использовались все технические повторы одного из доноров, а все остальные данные образовывали обучающую выборку, оценивались на основании кривой мощности критерия (ROC-кривой), чувствительности и специфичности.

По результатам тестов ошибка классификации с использованием ансамбля из 100 моделей равнялась 30% (Se≈73%, Sp≈61%) при семи латентных структурах. Более полные результаты будут представлены в докладе.

Библиографический список

1. Подлесных С.В., Колосова Е.А., Щербаков Д.Н., Шайдуров А.А., Анисимов Д.С., Рязанов М.А., Джонстон С.А., Шойхет Я.Н., Петрова В.Д., Лазарев А.Ф., Шаповал А.И. Взаимодействие антител сыворотки крови пациентов при раке молочной железы с синтетическими пептидами // Бюлл. Эксп. Биол. Мед. – 2015.

2. Анисимов Д.С., Рязанов М.А. Шаповал А.И. Подход к обработке многомерных данных пептидных микрочипов // Известия АлтГУ. – 2015. – №1/2(85). С. 77-80.

3. Анисимов Д.С., Рязанов М.А., Шаповал А.И. Применение метода проекции на латентные структуры в задачах классификации на примере данных пептидных микрочипов // МАК-2016: сборник трудов всероссийской конференции по математике, Барнаул, 29 июня – 1 июля 2016 г. – Барнаул: Изд-во АлтГУ, 2016. – С. 92

4. Эсбенсен К. Анализ многомерных данных. Избранные главы / пер. с англ. С.В. Кучерявского; под ред. О.Е. Родионовой. – Барнаул: Изд-во Алт. ун-та, 2003. – 157 с.

5. Breiman L. Bagging predictors // Machine Learning. – 1996. – 24 (2): P. 123–140. doi: 10.1007/BF00058655

6. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. – New York: Springer. 2013. – 426 p.