Секция 4. ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

УДК 57.087

Применение метода проекции на латентные структуры в задачах классификации на примере данных пептидных микрочипов

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

В работе рассматривается применение метода избавления от мультиколлинеарности многомерных данных путём проецирования в пространство меньшей размерности. Классическим и математически обоснованным инструментом уменьшения размерности многомерных данных является метод главных компонент (МГК), позволяющий отделить информационную составляющую данных от шума. Но нами был выбран метод проекции на латентные структуры (ПЛС), который в отличие от МГК, при разложении матрицы данных использует информацию о классах объектов [1].

Для экспериментов использовались данные пептидных микрочипов двух поколений (10K и 330K) имеющих на своей поверхности 10 тысяч и 330 тысяч пептидов, соответственно. Микрочипы 10K использовались в задаче классификации доноров с диагнозом рак молочной железы и контрольных доноров [2].

В качестве предварительной обработки использовались исследованные ранее методы [2, 3]. В частности, данные предварительно логарифмировались по основанию 2, затем подвергались медианной нормализации для подавления различий фонового свечения различных чипов. Дальнейший анализ с использованием ПЛС проходил в двух режимах: с предварительным отбором относительно небольшого числа (1-10%) информативных переменных (на основании критерия Уилкоксона-Манна-Уитни [4]), и без отбора информативных переменных. Такой подход позволил впоследствии оценить степень влияние шумов при избытке переменных и недостаток информации при малом количестве переменных.

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). Для 10K-чипов наименьшая ошибка классификации (6%) достигалась при использовании 6-и латентных структур построенных по всему множеству переменных. При этом чувствительность (Se) и специфичность (Sp) равнялись 92.5% и 94% соответственно. Для 330K-чипов на всём множестве переменных ошибка классификации равнялась 22% (Se=70%, Sp=87.8%) при 3-х латентных структурах, но в случае выделения 30000 информативных переменных (примерно 10%), минимальная ошибка увеличивалась до 38% (Se=55%, Sp=69.5%) и достигалась на 5-и латентных структурах. Более полные результаты будут представлены в докладе.

Использованный проекционный метод сокращения размерности позволил уменьшить размерность данных до 2-6 переменных, которые в отличие от МГК упорядочены по ковариации с метками классов объектов [1]. Дальнейшие исследования будут направлены на анализ возможностей совместного использования ПЛС и более мощных классификаторов, таких как SVM и нейронные сети.

Библиографический список

1. 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.

2. Анисимов Д.С. О некоторых алгоритмах обработки пептидных микрочипов // Сборник научных статей международной конференции «Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования», Барнаул, 20-24 октября, 2015. – Барнаул: Изд-во Алт. ун-та, 2015. – С. 619–624.

3. 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.

4. Mann H. B., Whitney D. R. On a test of whether one of two random variables is stochastically larger than the other // Annals of Mathematical Statistics. – 1947. – № 18. – P. 50–60.