Секция 3. СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ В ПРИКЛАДНЫХ ЗАДАЧАХ

УДК 519.23

Применение ледж-коэффициента в задаче бинарной классификации биомедицинских данных с ДНК-микрочипов

И.Ю. Бойко
АлтГУ, г. Барнаул

В связи с развитием технологий, используемых в современных биомедицинских исследованиях, происходит увеличение объема информации, подлежащей анализу. Одним из актуальных примеров является класс задач бинарной классификации многомерных данных, полученных с ДНК-микрочипов [1, 2]. Такая информация представлена значениями числовых признаков, количество которых измеряется тысячами, что значительно увеличивает время анализа данных. Для качественного решения рассматриваемых задач классификации широко используются алгоритмы фильтрации, поскольку среди методов отбора признаков они являются наиболее вычислительно эффективными. Идея этого подхода состоит в выборе подмножества признаков, упорядоченных согласно некоторой заданной мере [3]. Однако, распространенные в настоящее время алгоритмы фильтрации не вполне сосредоточены на выявлении связи между числовым и бинарным признаками, свойственной рассматриваемым задачам. В связи с этим для ее оценивания в работе [4] был введен ледж-коэффициент корреляции, в статье [5] предложены алгоритмы по его вычислению, в работе [6] описан алгоритм фильтрации, основанный на применении ледж-коэффициента.

Для исследования был использован набор данных с ДНК-микрочипов, подготовленный группой ученых во главе с Э. Гравье, содержащий сведения о 168 пациентах, у которых был диагностирован рак молочной железы. Данные о каждом объекте представлены значениями 2905 числовых признаков. По результатам 5 лет наблюдений после диагностики объекты данных были размечены следующим образом. Класс A сопоставили 111 пациентам (за время наблюдений не произошло появление метастазов). Класс B был

сопоставлен остальным 57 пациентам (произошло появление метастазов) [7].

Для отбора признаков мы использовали ледж-критерий, а также алгоритмы фильтрации на основе t-критерия Стьюдента и U-критерия Манна-Уитни, применение которых распространено в рассматриваемом классе задач.

Отбор признаков выполнялся с уровнями значимости 0.05 и 0.1. Затем применялся, либо не применялся метод проекции на латентные структуры (PLS), использование которого в исследуемых задачах обсуждалось в работе [8].

Отбор значимых признаков и последующая классификация были выполнены с использованием перекрестной проверки типа «один против всех» (Leave One Out). В качестве классификатора использовался метод опорных векторов с ядром в виде радиально-базисных функций (rbf-SVM). Для расчетов использовался язык программирования Python 3.7, модули NumPy, scikit-learn. Далее представлены основные результаты бинарной классификации, выполненной по вышеописанной методике.

Таблица 1 – Результаты классификации при использовании различных методов отбора признаков с уровнем значимости 0.05

№	Снижение размерности (статистическое)	Снижение размерности (проекционное)	Точность	AUC
1	-	-	0.744	0.825
2	-	PLS	0.700	0.758
3	t-критерий	-	0.789	0.864
4	t-критерий	PLS	0.756	0.822
5	U-критерий	-	0.733	0.845
6	U-критерий	PLS	0.622	0.653
7	Ледж-критерий	-	0.722	0.769
8	Ледж-критерий	PLS	0.644	0.606

Таблица 2 – Результаты классификации при использовании различных методов отбора признаков с уровнем значимости 0.1

№	Снижение размерности (статистическое)	Снижение размерности (проекционное)	Точность	AUC
1	-	-	0.744	0.825
2	-	PLS	0.700	0.756
3	t-критерий	-	0.778	0.859
4	t-критерий	PLS	0.722	0.781
5	U-критерий	-	0.756	0.848
6	U-критерий	PLS	0.711	0.761
7	Ледж-критерий	-	0.722	0.805
8	Ледж-критерий	PLS	0.678	0.810

Из представленных выше расчётов видно, что применение ледж-критерия для отбора значимых признаков даёт результаты сравнимые с теми, которые получены распространенными алгоритмами.

Качество классификации снижается с применением PLS, при котором используется линейная разделимость классов, что может говорить о наличии полезной нелинейной информации в данных.

Библиографический список

1. Onskog J., Freyhult E., Landfors M., Ryden P., Hvidsten T.R. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning // BMC Bioinformatics. – 2011. – V.12. – P. 77-89.

2. Mohammed A., Biegert G., Adamec J., Helikar T. CancerDiscover: An integrative pipeline for cancer biomarker and cancer class prediction from high-throughput sequencing data // Oncotarget. – 2018. – V. 9(2). – P. 2565–2573.

3. Hira Z., Gillies D. A review of feature selection and feature extraction methods applied on microarray data // Advances in Bioinformatics. – 2015. – V. 2015. – P. 1-13.

4. Дронов С.В., Петухова Р.В. Один вид связи между номинальной и бинарной переменными // Известия АлтГУ. – 2010. – №1/2 (65). – С. 34–36.

5. Дронов С.В., Бойко И.Ю. Метод оценки степени связи бинарного и номинального показателей // ПДМ. – 2015. – №4(30). – С. 109-119.

6. Бойко И.Ю., Дронов С.В. Критические точки распределения ледж-коэффициента // Сборник трудов Всероссийской конференции по математике «МАК-2016», Барнаул, 29 июня - 1 июля 2016 г. – Барнаул: Изд-во АлтГУ, 2016. – С. 13–15.

7. Gravier E. A prognostic DNA signature for T1T2 node-negative breast cancer patients // Genes, Chromosomes and Cancer. – 2010. – V.49(12). – P. 1125–1134.

8. Анисимов Д.С., Подлесных С.В., Колосова Е.А., Щербаков Д.Н., Петрова В.Д., Джонстон С.А., Лазарев А.Ф., Оскорбин Н.М., Шаповал А.И., Рязанов М.А. Анализ многомерных данных пептидных микрочипов с использованием метода проекции на латентные структуры // Математическая биология и информатика. – 2017. – №2(25). – С. 435-445.