Библиографический список

1. Попов Ф.А., Груздев Г.П., Филиппов С.А. Технология разработки программного обеспечения ЭВМ М–400 и М–6000 с использованием ЭВМ БЭСМ–6 // Управляющие системы и машины. – 1980. – №1. – С.41–45.

2. Попов Ф.А., Карлов А.А. Диалэм – диалоговая система для разработки математического обеспечения ЭВМ в режиме эмуляции //Материалы третьей Всесоюзной конференции «Диалог Человек-ЭВМ». – Протвино: ИФВЭ, 1983. – С.69.

3. Попов Ф.А., Жарков А.С., Филиппов С.А. Диалоговая система для программирования микропроцессорных управляющих устройств на основе КТС ЛИУС-2 // Передовой производственный опыт. –1986. – № 5. – С. 25.

4. Жарков А.С., Звольский Л.С., Литвинов А.В., Попов Ф.А. Проблемы создания интегрированных АСУ для производств спецхимии и пути их решения. – Бийск: Изд-во Алт. гос. техн. ун-та, 2014. – 188 с.

5. Абрамов Д.Г., Звольский Л.С., Кодолов А.В., Попов Ф.А. Особенности и перспективы создания АСУ технологическими процессами производств спецхимии // Фундаментальные исследования. – 2015. – № 9. – С. 407–413.

6. Абрамов Д.Г., Кодолов А.В., Попов Ф.А. Особенности построения пользовательских интерфейсов для автоматизированных систем управления производствами спецхимии // Автоматизация в промышленности. –2018. –№6. – С. 52–57.

7. Вельбицкий И.В. Новая графическая концепция программирования // Южно-Сибирский научный вестник. – 2018. – №4(24). –С.83–98.

УДК 519.23

Использование ледж-коэффициента в задаче бинарной классификации данных с пептидных микрочипов

И.Ю. Бойко

АлтГУ, г. Барнаул

Многие современные биомедицинские исследования, проводимые с использованием микрочипов, связаны с поиском методов ранней диагностики онкологических заболеваний и направлены на решение проблем бинарной классификации [1, 2].

Существуют различные виды микрочипов. В большинстве публикаций, посвященных рассматриваемому классу задач, авторы работают с

данными ДНК-микроматриц. Анализ информации с пептидных микрочипов имеет свои особенности и сложности в силу природы данных и конструкции микрочипов. Кроме того, эта информация представлена числовыми значениями набора биомедицинских признаков, количество которых может доходить до сотен тысяч. Следовательно, обработка данных с микрочипов требует применения нестандартных методов анализа данных, в том числе алгоритмов отбора признаков, значимых для решения задачи бинарной классификации. Распространенные в настоящее время методы не вполне сосредоточены на выявлении связи между числовым и бинарным признаками, свойственной задачам рассматриваемого типа. В работе [3] был введен ледж-коэффициент корреляции для оценивания силы такой связи, в статье [4] предложены алгоритмы по его вычислению, в работе [5] описан ледж-критерий отбора значимых признаков на основе представленного коэффициента. Данные методы реализованы нами программно на языке Python 3.7 с использованием пакета NumPy.

В нашем исследовании использована информация с пептидных микрочипов Российско-Американского противоракового центра АлтГУ. Приведем краткие сведения об этих данных. На микрочипы специальным образом были нанесены материалы доноров, относящихся к двум группам. Первая группа сформирована из 40 пациентов с диагнозом рака молочной железы (РМЖ), а вторая из 41 здорового донора без признаков РМЖ. Материал каждого пациента был нанесен на два разных микрочипа. Таким образом, после оцифровки информации с микрочипов итоговый набор данных состоит из 162 объектов и 330034 переменных, значения каждой из которых изменяются в диапазоне от 0 до 65535 [6].

Для выделения значимых признаков были применены t-критерий Стьюдента, U-критерий Манна-Уитни, а также предложенный нами ледж-критерий. Отбор признаков выполнялся с фиксированным уровнем значимости 0.05. В нашей модели использован метод проекции на латентные структуры (PLS), эффективность применения которого к исследуемым данным была представлена в работе [6].

Для оценки качества модели применялась перекрёстная проверка типа «один против всех». На каждой ее итерации происходил отбор значимых признаков на основании указанных выше критериев. Затем, на выбранном множестве пептидов выполнялась настройка модели регрессии на латентные структуры (PLS-R) с предварительным применением, либо без применения метода PLS [6]. В таблице 1 представлены основные результаты использования рассмотренных подходов на вышеописанных данных.

Таблица 1 – Качество классификации тестовых образцов при использовании различных методов отбора признаков

№	Снижение размерности (статистическое)	Снижение размерности (проекционное)	Точность	AUC
1	-	-	0.778	0.824
2	-	PLS	0.679	0.728
3	t-критерий	-	0.556	0.604
4	t-критерий	PLS	0.562	0.617
5	U-критерий	-	0.556	0.564
6	U-критерий	PLS	0.667	0.734
7	Ледж-критерий	-	0.611	0.654
8	Ледж-критерий	PLS	0.778	0.866

Таким образом, использование метода проекции на латентные структуры в сочетании с ледж-критерием отбора значимых признаков позволило несколько улучшить качество бинарной классификации. Произошло увеличение значения AUC с 0.824 до 0.866 при точности классификации, оставшейся неизменно равной 0.778. Отметим, что доля значимых признаков, отобранных с использованием ледж-критерия, составила 11.73% от общего числа признаков.

Библиографический список

1. Alanni R., Hou J., Azzawi H., Xiang Y. A novel gene selection algorithm for cancer classification using microarray datasets // BMC Med Genomics. – 2019. – V. 12. – P. 3441.

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

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

5. 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.

6. 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.