УДК 519.251
ББК 22.172

ИЗУЧЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ ТКАНИ ПО ЭКСПРЕССИИ ГЕНОВ НА МИКРОЧИПАХ

Беляев А. О.[1]
(*Учреждение Российской академии наук
Институт проблем управления РАН, Москва*)

Исследуется связь числа генов, использованных для классификации контрольной и подверженной онкологическому заболеванию ткани по экспрессии генов на микрочипах, с качеством классификации. Предложен непараметрический робастный метод классификации, основанный на расстоянии Хэмминга до медиан в двух классах. Показано, что исключение из процедуры обучения генов, мало информативных по симметризованному информационному расстоянию Кульбака-Лейблера, повышает качество классификации, оцениваемое по экзаменационной выборке.

Ключевые слова: экспрессия генов на микрочипах, информативные гены, робастное правило классификации, расстояние Хэмминга, доля верных ответов на экзаменационной выборке.

1. Введение

Активное развитие иструментальных методов анализа биомедицинской информации и, в частности, завершение международной программы "Геном человека", ставит новые задачи как в области биологии, медицины, здравоохранения, так и области математической обработки данных. Широкое внедрение результатов генетического анализа в научные исследования позволяет

[1] *Александр Олегович Беляев, студент (punk2-k@bk.ru)*

выявлять факторы риска развития тяжёлых наследственных заболеваний, прогнозировать "генетическую программу" здоровья человека и даже ставить задачу поиска индивидуальных рецептов "долголетия" [1].

Широкое внедрение результатов генетического анализа в практику требует разработки формальных методов, позволяющих отсеивать мало информативные и искажённые наблюдения, строить процедуры принятия решений, не требующие привлечения чрезмерных априорных предположений о структуре и характере распределений данных, а также позволяющие проводить надёжную верификацию и оценку достоверности полученных результатов. Сложность разработки подобных процедур и методов связана с тем, что современные данные генетического анализа, например, с использованием микрочипов, имеют структуру, противоречащую требованиям классической математической статистики. В таких данных число наблюдений - количество исследуемых образцов, много меньше числа изучаемых факторов - числа генов. Результаты анализа с использованием современных микрочипов оперируют экспрессией десятков тысяч генов, полученной лишь от сотен образцов тканей. Это связано, во-первых, с дороговизной генетического анализа, которая в скором времени будет преодолена, и, во-вторых, с неоднородностью как изучаемых групп пациентов и здоровых людей, так и с многообразием проявлений самого заболевания.

Методы решения задач анализа таких данных развиваются по направлению снижения размерности задачи, то есть уменьшения числа изучаемых факторов, и по направлению разработки робастных, устойчивых методов принятия решений. В статье предложен робастный метод дифференциальной классификации образцов нормальной и подверженной онкологическому заболеванию ткани по экспрессии различного числа генов на микрочипах. Исследуется влияние числа наиболее информативных по симметризованному расстоянию Кульбака-Лейблера [3] генов на точность классификации, оцениваемую по экзаменационной выборке.

2. Классификация по медиане

Широко распространённые методы классификации опираются на ряд априорных предположений. В методе наивного Байеса предполагается независимость признаков в рамках изучаемых классов, при построении дискриминантных функций Фишера считается, что в каждом классе признаки имеют нормальное распределение [4]. Часто такие предположения приводят к приемлемому результату, но, как оказывается, при малом числе наблюдений работают плохо и часто более простые, но робастные по отношению к априорным предположениям правила, на практике дают лучшие результаты [5].

Для построения классификатора, не зависящего от априорных предположений о распределении признаков в классах, в статье предлагается использовать медиану, которая является устойчивой характеристикой. Для описания предлагаемого метода вычислим по обучающей выборке вектор значений медианы признаков без разделения на классы и преобразуем элементы обучающей выборки в бинарные значения 0 и 1 по правилу: если признак меньше соответствующего признака медианы, то он кодируется значением 0, иначе он кодируется 1. Кодировка нового вектора z, предъявляемого для опознавания, осуществляется аналогично. Для закодированных векторов вычисляются расстояния Хэмминга вектора z от закодированной обучающей выборки первого и второго классов по формуле

$$s_k = \frac{1}{L_k}\sum_{i=1}^{n}\sum_{j=1}^{L_k}\left|x_i^{jk} - z_i\right|,$$

здесь k=1,2 - индекс класса, L_1 и L_2 - число векторов в обучающей выборки первого и второго класса соответственно, z_i - значение i-той координаты закодированного вектора z, n- число признаков. Для принятия решения о принадлежности вектора z к первому или ко второму классу вычисляется индекс $S=(s_1+s_2)/2$, который сравнивается с пороговым значением g. Если $S>g$, то вектор z относится к классу 2. При $S<g$ вектор z относится к классу 1. При $S=g$ вектор z равновероятно относится

к классу 1 или 2. Величина порога g выбирается по кривой ошибок, которая строится по результатам классификации обучающей выборки.

3. Результаты классификации реальных данных

При анализе использовались полученные на микрочипах данные по экспрессии генов в метастазе, отнесенные к экспрессии генов в нормальной ткани и по экспрессии генов в опухоли, отнесенные к экспрессии генов в метастазе при раке печени. Такая предобработка проводилась для устранения систематических инструментальных погрешностей. Были сформированы две таблицы по 127 столбцов, что соответствует 127 пациентам, и по 7581 строк, что соответствует данным экспрессии 7581 гена. Для построения и проверки правила классификации пациенты были разделены случайным образом на две группы: обучение (63 человека) и экзамен (64 человека). Все элементы обучающей выборки и классифицируемые векторы z кодировались согласно описанному выше принципу по формулам

$$x_i^{jk} = \begin{cases} 0, & x_i^{jk} < M \\ 1 & x_i^{jk} \geq M \end{cases}, \qquad z_i = \begin{cases} 0, & z_i < M \\ 1 & z_i \geq M \end{cases}.$$

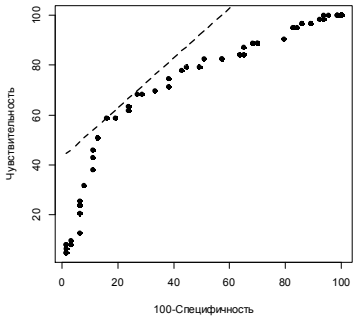

Рис. 1. Кривая ошибок классификации по медиане при использовании всех генов.

Результат классификации обучающей выборки при использовании всех генов при различных значениях порога g представлен на рис. 1 в виде традиционной кривой ошибок (ROC кривой). По оси абсцисс отложена вероятность ошибочного принятия класса 2 за класс 1, а по оси ординат вероятность верного опознания класса 1. Обе вероятности выражены в процентах. Пунктирная линия касается кривой в точке, которая соответствует порогу при минимальной суммарной ошибке классификации, которая в данном случае оказывается равной 57,2% и не может считаться удовлетворительной.

Гораздо лучший результат получается, если проводить классификацию по 50% наиболее информативным по симметризованному расстоянию Кульбака-Лейблера генам, отобранным согласно процедуре, описанной в [2]. Соответствующая кривая ошибок приведена на рис. 2. Минимальная ошибка классификации при этом равна 17,5%.

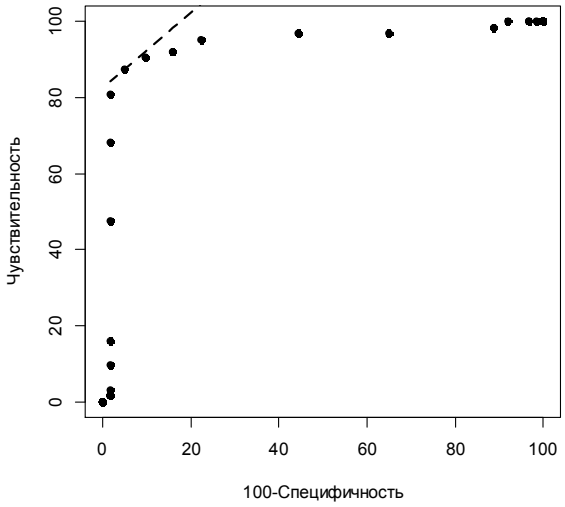

Рис. 2. Кривая ошибок классификации по медиане при использовании 50% наиболее информативных генов.

При классификации по 10% наиболее информативных генов минимальная ошибка классификации, вычисленная на обучающей выборке оказывается равной 3,2%.

В таблице 1 приведены минимальные ошибки классификации, вычисленные по обучающей выборке с помощью кривой ошибок, и ошибки классификации на экзаменационной выборке при найденных по обучающей выборке значениях порога g, в зависимости от доли наиболее информативных генов. Из таблицы видно, что удаление числа мало информативных генов уменьшает ошибку классификации как на обучающей, так и на экзаменационной выборке. При этом ошибка классификации на экзаменационной выборке соизмерима с ошибкой классификации на обучающей выборке.

Таблица.1. Ошибки классификации при удалении мало информативных генов

Доля удалённых генов (%)	Число оставшихся генов	Ошибка на обучении (%)	Ошибка на экзамене (%)
50	3790	17,5	20,3
60	3032	14,3	28,1
70	2274	11,1	23,4
80	1516	9,5	6,2
90	758	3,2	4,7

4. Заключение

В статье проводится экспериментальное исследование влияния числа используемых генов на качество классификации типа ткани. Для классификации применяется правило, основанное на расстоянии Хэмминга между обучающими и распознаваемым векторами, закодированными относительно медианы обучающей выборки. Проведённое исследование демонстрирует, что отбор генов по их информативности, оцениваемой через расстояние Кульбака-Лейблера между распределениями генов в

двух классах, повышает качество классификации как на обучающей, так и на контрольной выборках. Низкое качество распознавания при использовании всего набора генов объясняется сложностью решаемой задачи и малым размером обучающей выборки. В настоящем исследовании вместо усложнения решающего правила был выбран путь отбраковывания генов, имеющих близкие распределения экспрессии в двух классах, и использования генов, распределения экспрессии которых наиболее различимы в рассматриваемых тканях. Классификация на основании таких генов позволила повысить результат классификации, используя при этом очень простое, но робастное правило классификации.

Литература

1. БАРАНОВ В.С. *Программа "Геном человека" и научная основа профилактической медицины* // Вестник РАМН. − 2000. − №10. − С. 27-36.
2. ВОРОНЦОВ А.С., МИХАЛЬСКИЙ А.И. *Снижение размерности в задаче анализа данных экспрессии генов на микрочипах* // Вестник МГТУ МИРЭА. − 2014. − **№1(2)** . − С. 141-146.
3. КУЛЬБАК С. *Теория информации и статистика.* − М.: Наука, 1967. −408 с.
4. Мерков А.Б. *Распознавание образов: Введение в методы статистического обучения*. − М.: Едиториал УРСС, 2011. −256 с.
5. Раудис Ш. *Влияние объёма выборки на качество классификации* / Ш. Раудис // Статистические проблемы управления. − Вильнюс. Институт математики и кибернетики АН Литовской ССР, 1984. − Вып. 66. − С. 9-42.

INVESTIGATION OF TISSUE CLASSIFICATION PERFORMANCE ON MICROCHIP GENE EXPRESSION DATA

Alexander Beliaev, Institute of Control Sciences of RAS, Moscow, student (punk2-k@bk.ru).

Abstract: A classifier performance is evaluated in dependence to number of genes, used to classify the control and cancer tissues, based on the microchip gene expression data. A robust classification procedure, based on Hamming distance to inner class medians, is proposed. It is demonstrated that elimination of genes, which are less informative in terms of Kullback–Leibler divergence, from the learning procedure increases the resulting classification performance, estimated using examine sample.

Keywords: microchip gene expression data, informative genes, robust classification, Hamming distance, classifier performance.