УДК 519.251
ББК 22.172

ПРИМЕНЕНИЕ ИНФОРМАЦИОННОГО РАССТОЯНИЯ МЕЖДУ ВНУТРИКЛАССОВЫМИ РАСПРЕДЕЛЕНИЯМИ ДЛЯ ВЫБОРА НАБОРА ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ

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

Для оценки информативности генов на основании их экспрессии на микрочипах рассматривается характеристика, близкая к информационному расстоянию Кульбака-Лейблера между распределениями экспрессии гена в образцах контрольной и подверженной онкологическому заболеванию ткани. Описывается методика оценки расстояния между распределениями и результаты расчётов по реальным данным. Показана высокая информативная избыточности использования всех генов и предложено для классификации ткани отбирать гены, информационное расстояние между распределениями экспрессии которых в различных образцах ткани велико.

Ключевые слова: экспрессия генов на микрочипах, расстояние между распределениями, снижение размерности.

1. Введение

Современные тенденции развития медицины связано с совершенствование технологической базы биологических исследований и, в частности, с успехами в области генетических исследований [1]. Доступность большого числа генетических данных и различных биомаркеров позволила приблизиться к понима-

[1] *Алексей Сергеевич Воронцов, студент (lexa_13a@mail.ru)*

нию механизмов возникновения и развития онкологических заболеваний, заболеваний сердечно-сосудистой системы, многих других хронических патологий и открыла перспективы создания новых лекарств и методов лечения, что выражается в тенденции развития персонифицированной медицины [4-7].

В настоящее время промышленно производится большое число разнообразных генных микрочипов, используемых для выяснения предрасположенности человека к различным заболеваниям, оценки реакции организма на различные лекарственные препараты, и многое другое [2]. Генные микрочипы позволяют с высокой точностью диагностировать заболевание не только на ранних стадиях, но и до проявления этого заболевания.

Успешное развитие и внедрение технологии генных микрочипов связано с успехами в области анализа данных, полученных с помощью микрочипов. В то же время, при анализе таких данных возникают и новые задачи. Генные микрочипы позволяют за одно исследование в автоматическом режиме получить информацию об экспрессии тысяч генов. Из-за погрешностей приборов, человеческого фактора, несовершенства методов регистрации сигналов, наличия молчащих генов и многого другого, в выходных данных появляются погрешности и ошибки. Кроме того, для получения статистически надёжного результата при регистрации большого числа генов необходимо иметь достаточное количество повторных независимых наблюдений, то есть проводить генетический анализ большого числа людей, страдающих одинаковой болезнью, что проблематично не только из-за высокой стоимости исследований, но и по причине малой распространённости конкретных патологий в популяции.

Таким образом, возникает задача выбора генов, которые являются наиболее информативными при изучении конкретной патологии. В статье рассматривается метод выделения информативных генов путём сравнения распределений экспрессии генов в образцах контрольной и подверженной заболеванию ткани.

2. Оценка информативности генов

Выделение информативных генов основывается на характеристике расстояния между распределениями экспрессии гена в образцах контрольной и подверженной заболеванию тканей. В качестве такой характеристики в статье рассматривается дивергенция Ку́льбака — Ле́йблера, называемая в теории информации информационной дивергенцией, либо относительной энтропией [2]. Эта характеристика является несимметричной мерой удаленности друг от друга двух вероятностных распределений. Обычно, одно из сравниваемых распределений — это «истинное» распределение, второе — проверяемое, являющееся приближением первого. Для дискретных распределений расстояния Кульбака-Лейблера вычисляется по формуле:

$$D_{KL} = \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

где $p(x)$, $q(x)$ – функции вероятности распределений дискретной случайной величины X при двух рассматриваемых гипотезах.

При рассмотрении задачи анализа экспрессии генов, функции вероятности заменим гистограммами, построенными по наборам значений экспрессий в двух классах. В качестве классов рассмотрим нормальную ткань и подверженную заболеванию. Характеристику, аналогичную расстоянию Кульбака-Лейблера для различия распределений, запишем как

$$D_{KL} = -\frac{1}{2} \sum_x \left(p_1(x) \ln p_2(x) + p_2(x) \ln p_1(x) \right)$$

здесь суммирование проводится по множеству дискретных значений экспрессии гена, использованных для построения гистограмм в двух классах, $p_1(x)$ – значения гистограммы, построенной в первом классе, $p_2(x)$ – значения гистограммы, построенной во втором классе.

Вычислим характеристику D_{KL} для каждого гена и упорядочим гены по мере убывания этой величины. Полученный ряд даёт представление об информативности генов при сравнении двух классов и может использоваться в дальнейшем анализе,

например, для отбора признаков при построении правила классификации.

3. Результаты анализа реальных данных

Методика оценки информативности генов применялась к реальным данным, состоявшим из двух наборов. В первом наборе были представлены данные по экспрессии генов в опухоли при раке груди. Во втором наборе содержались данные по экспрессии генов в нормальной ткани. Данные были предварительно нормированы для устранения возможных инструментальных искажений. Оба набора включали данные по 17682 генам.

На рис. 1 представлена плотность распределения характеристики D_{KL}, вычисленной по всем генам. Эта плотность строилась с помощью процедуры density из статистического пакета R с использованием параметров сглаживания, заданных в процедуре по умолчанию.

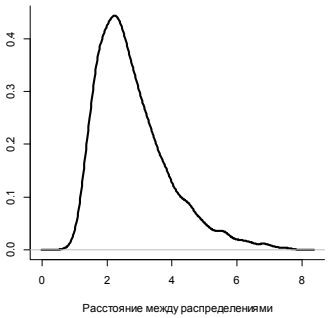

Рис. 1. Плотность распределения значений характеристики D_{KL} для 17682 генов.

На рис. 2 для иллюстрации представлены примеры распределений в двух классах экспрессий генов с маленькой и с большой величиной характеристики D_{KL}.

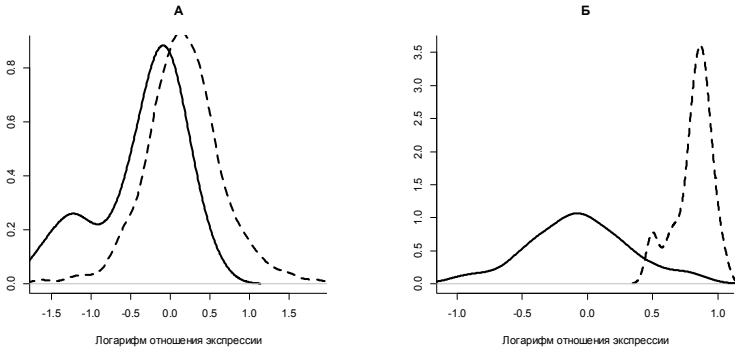

Рис. 2. Плотность распределения экспрессии генов в двух классах. А - *для гена с* $D_{KL} = 2,6$, Б - *для гена с* $D_{KL} = 8,0$. *Сплошная линия - раковая ткань, пунктир - нормальная ткань.*

Из рисунка видно, что маленькая величина характеристики D_{KL} соответствует плохо разделимым распределениям, а большая величина - хорошо разделимым.

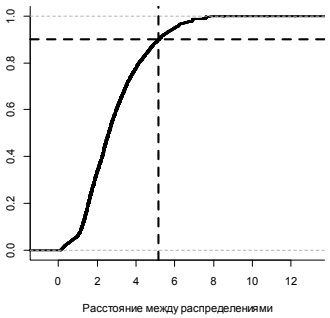

Рис. 3. Кумулятивная функция распределения характеристики D_{KL} *для 17682 генов.*

Чтобы выделить гены, распределение экспрессии которых в двух классах различаются в наибольшей степени, строилась

кумулятивная функция распределения. График кумулятивной функции, построенной с помощью процедуры ecdf из статистического пакета R, представлен на рис. 3. На рисунке вертикальная пунктирная линия указывает для значения характеристики D_{KL} порог, выше которого лишь 10% генов имеют большие значения.

4. Заключение

Проведенный анализ распределений экспрессии генов показал, что с помощью расстояния Кульбака-Лейблера можно произвести отбор генов, для которых характерны существенные различия в распределении их экспрессии в нормальной и в раковой ткани. Такая процедура проводится автоматически и позволяет существенно сократить размерность решаемой задачи, то есть число генов, которые необходимо исследовать на следующих стадиях анализа данных. Отобранные гены могут использоваться при построении решающих правил для классификации типов тканей и для выявления структурных взаимосвязей генов (построения генных сетей). Методика применения расстояния Кульбака-Лейблера для оценки информативности признаков универсальна и может применяться для анализа данных любой природы.

Литература

1. БАРАНОВ В.С. *Программа "Геном человека" и научная основа профилактической медицины //* Вестник РАМН. − 2000. − №10. − С. 27-36.
2. КУЛЬБАК С. *Теория информации и статистика.* − М.: Наука, 1967. −408 с.
3. СВЕШНИКОВА А.Н., ИВАНОВ П.С. *Экспрессия генов и микрочипы: проблемы количественного анализа //* Рос. хим. ж. − 2007. − Т.LI, № 1. − С. 127-135.

4. AU W.W., RUCHIRAWAT M. *Biomarkers in population studies: environmental mutagenesis and risk for cancer* // Rev. Environ. Health. − 2009. − Vol.24, №2. − P. 117-127.
5. CARRARA S., GHOREISHIZADEH S., OLIVO J. *Fully integrated biochip blatforms for advanced healthcare* // Sensors. − 2012. − Vol.12, №8. − P. 11013–11060.
6. GOLDMAN M.A. *Digital drug discovery* // Genome Biology. − 2005. − Vol.6 − P. 348-350.
7. Napoli C., Lerman L.O., Sica V. *Microarray analysis: a novel research tool for cardiovascular scientists and physicians* // Heart. − 2003. − Vol.89 − P. 597–604.

IMPLEMENTATION OF INFORMATION DIVERGENCE BETWEEN INCLASS DISTRIBUTIONS FOR SELECTION OF FEATURE SET FOR CLASSIFICATION

Alexey Vorontsov, Institute of Control Sciences of RAS, Moscow, student (lexa_13a@mail.ru).

Abstract: A problem of informative genes selection is considered. A divergence between gene expression distribution in control and cancer tissues is constructed like the symmetric Kulback-Leibler distance. An estimate of the divergence between gene expression distributions is proposed and the results for data on hepatic cancer are presented. It is shown that no more than 10% of presented genes demonstrate 90% of divergence between gene expression distributions. This fact indicates an informative redundancy of the total set of genes in the presented data.

Keywords: microchip gene expression data; distributions divergence; dimension reduction; informative redundancy.