УДК 517

Подход к обработке многомерных данных пептидных микрочипов

Д.С. Анисимов[1], М.А. Рязанов[1], А.И. Шаповал[1,2]

[1] Алтайский государственный университет (Барнаул, Россия)
[2] Университет штата Аризона (Темпи, США)

Approaches for Large Volume Data Analysis Obtained Utilizing Peptide Microarrays

D.S. Anisimov[1], M.A. Ryazanov[1], A.I. Chapoval[1,2]

[1] Altai State University (Barnaul, Russia)
[2] Arizona State University (Tempe, USA)

Рассматривается подход к обработке многомерных данных пептидных микрочипов. Основными этапами используемой технологии являются предобработка с целью уменьшения аппаратных ошибок измерений, снижения размерности для выделения переменных, наилучшим образом описывающих исходные данные, и классификация тестовых данных, результатом которой является определение класса нового объекта, с использованием множества объектов-образцов. Ввиду малого количества тестовых данных (в работе использовались пробы 25 доноров, из которых 15 — условно здоровые и 10 — с диагнозом рака молочной железы) работа нацелена на апробацию различных алгоритмов обработки, анализ применимости и выявление путей их дальнейшего развития с целью улучшения качества и повышения устойчивости результатов при применении этих алгоритмов. Результатом работы является технология обработки данных пептидных микрочипов, и при дальнейшей доработке возможно ее применение на реальных данных с использованием большего количества образцов и большего количества классов.

Ключевые слова: пептидный микрочип, обработка многомерных данных, метод наименьших квадратов, метод опорных векторов, наивный байесовский классификатор, k-ближайших соседей.

This paper discusses strategies for high density peptide microarray data analysis. Main steps of the analysis include preprocessing, normalization (the reduction of hardware measurement errors), data volume reduction (the selection of statistically significant variables which describe the original data the best) and data classification (class determination among multiple samples). The study was designed for different algorithms optimization and testing to analyze large volume data. The main objectives of the study were improving existing algorithms and enhancing data sustainability for these algorithms utilization to analyze data obtained with peptide microarray. The work was performed on a relatively small sample set (25 donors, 15 — healthy and 10 — with breast cancer). The result of this study is a developed technology for peptide microarray data analysis, which may be used to evaluate a large number of samples and a greater number of classes.

Keywords: peptide microchip, processing of multidimensional data, ordinary least squares, support vector machines, naive Bayes classifier, k-nearest neighbors.

DOI 10.14258/izvasu(2015)1.2-13

Ранняя диагностика онкологических заболеваний является основным направлением научных исследований Российско-американского противоракового центра, созданного в Алтайском государственном университете. Данные исследования проводятся совместно с центром инновационной медицины Аризонского государственного университета и Алтайским краевым онкологическим диспансером.

Основой исследований является обработка данных пептидных микрочипов, разработанных в институте биодизайна университета штата Аризона [1]. В общем виде пептидный микрочип представляет собой подложку из нейтрального материала (стекла или полимерных материалов), на которую нанесены искусственно созданные последовательности аминокислот (пептиды).

В данной статье рассматривается проблема обработки данных, полученных с микрочипов, состоящих из четырех независимых блоков, содержащих по 10368 пептидов. Из них 80 «EMPTY» — пептидов, не содержащих аминокислот, 112 «FIDUCIAL» — пептидов, предназначенных для контроля, и 10176 уникальных пептидов, имеющих информационный характер.

По окончании биологической части эксперимента по каждому блоку микрочипа получается цифровая характеристика светимости пептидов, включая служебную информацию об условиях проведения эксперимента, минимальная математическая обработка. Общая структура получаемого файла описана в [2].

Предварительным этапом применения математических методов предобработки полученных многомерных данных является процесс их нормализации. Наиболее перспективными методами нормализации светимости пептидов являются методы, изложенные в работах [1, 3, 4]. Нами исследованы следующие алгоритмы нормализации:

— нормализация относительно фона;
— нормализация приведением к медианной светимости;
— квантильная нормализация;
— нормализация по методике, изложенной в работе М. Кретич и М. Чиари [4].

Приведем пример нормализации полученных данных на основе методики нормализации относительно фона. Суть данной методики заключается в том, что имея медиану светимости пептида i F_i и медиану светимости фона данного пептида B_i находим нормализованное значение N_i путем вычитания или деления светимости (логарифма светимости) пептида на светимость (логарифм светимости) фона $(i = 1,...,n)$:

$$N_i = F_i - B_i; \quad (1)$$

$$N_i = F_i / B_i; \quad (2)$$

$$N_i = \log(F_i + 1) - \log(B_i + 1); \quad (3)$$

$$N_i = \log(F_i + 1) / \log(B_i + 2). \quad (4)$$

На рисунке 1 показаны визуальные различия методов нормализации (1)–(4), а также ненормализованных данных и фонового свечения.

На рисунке 1 видно, что некоторые царапины и артефакты, хорошо заметные при отсутствии предобработки (рис. 1а) и на фоновом свечении (рис. 1б), после предобработки почти неразличимы.

Рассматривая дальнейшие пути анализа полученных данных, мы применили классические методы статистического анализа и классификации.

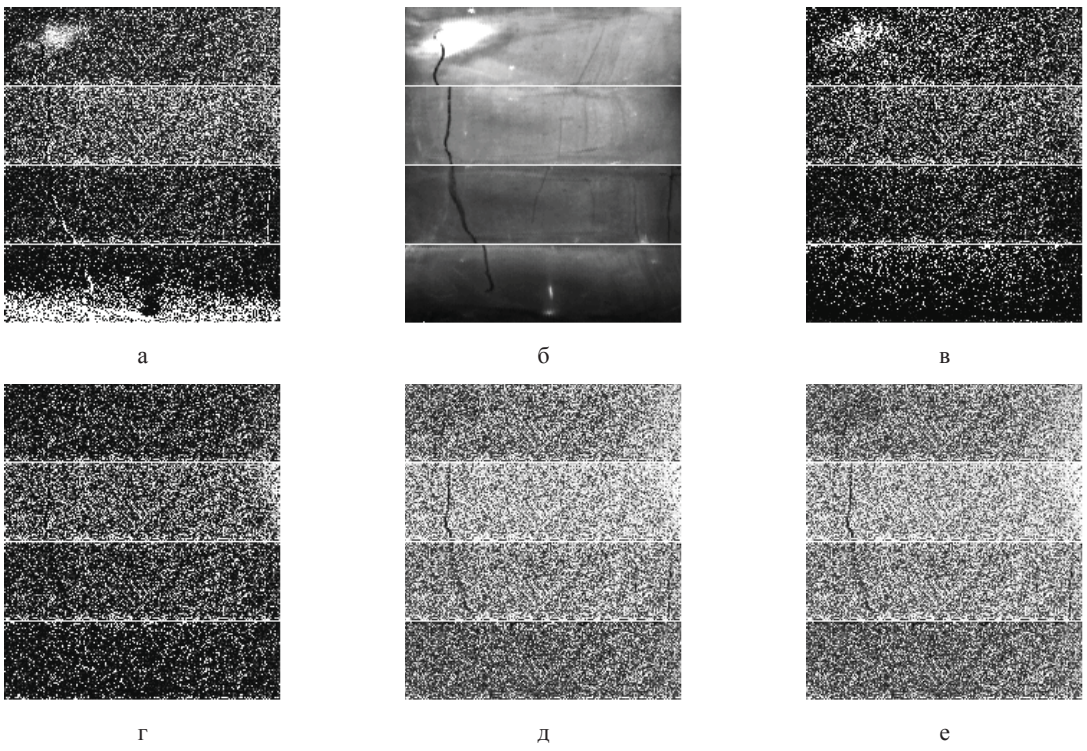

Рис. 1. Визуализация данных одного микрочипа: а — без предобработки; б — фоновое свечение; в–е — с использованием предобработки по формулам (1)–(4) соответственно

Результаты классификации тестовых данных при различных методах предобработки

Предобработка	Классификатор	Чувствительность	Специфичность	Точность
$F_i - B_i$	МНК	0.40	0.53	0.48
	SVM	0.90	0.73	0.80
	НБК	0.80	0.93	0.88
	KNN	0.60	0.60	0.60
F_i / B_i	МНК	0.30	0.87	0.64
	SVM	0.80	0.87	0.84
	НБК	0.60	0.93	0.80
	KNN	0.80	0.53	0.64
$\log(F_i + 1) - \log(B_i + 1)$	МНК	0.40	0.87	0.68
	SVM	0.90	0.87	0.88
	НБК	0.70	0.93	0.84
	KNN	0.80	0.67	0.72
$\log(F_i + 1) / \log(B_i + 2)$	МНК	0.90	0.93	0.92
	SVM	0.80	0.93	0.88
	НБК	0.50	0.87	0.72
	KNN	0.60	0.67	0.64

Примечание. Методы предобработки с использованием различных классификаторов: МНК — метод наименьших квадратов; SVM — метод опорных векторов; НБК — наивный байесовский классификатор; KNN — k-ближайших соседей

В данной работе представлены результаты исследования следующих методов: метод наименьших квадратов; метод опорных векторов; наивный байесовский классификатор; метод k-ближайших соседей. Вычислительные эксперименты по проведению анализа данных вышеуказанными методами проводились на основе перекрестной проверки «leave-one-out», при которой каждый из образцов по очереди используется для проверки, в то время как остальные используются для обучения классификаторов. В итоге получили усредненные оценки точности, чувствительности и специфичности, которые представлены в таблице. В наших расчетах средняя точность составила 0,75, чувствительность — 0,68, а специфичность — 0,8.

Представленная сводная таблица полученных результатов показывает, что применение данных методов возможно, но авторы считают, что каждый метод имеет ряд недостатков, которые могут поставить под сомнение возможность их использования в случае реального применения в условиях медицинской диагностики онкозаболеваний.

Рассмотрим эти недостатки более подробно. Метод наименьших квадратов: размерность данных значительно превышает количество образцов, что в свою очередь влечет недоопределенность информационной матрицы и невозможность ее обращения. Наивный байесовский классификатор: основной недостаток — малое количество повторных наблюдений в сравнении с размерностью данных, которое приводит к недообучению классификатора, как следует из результатов в приведенной таблице. Метод опорных векторов и метод k-ближайших соседей лишены вышеизложенных недостатков, однако авторы предполагают, что высокая зашумленность данных ставит под сомнение достоверность полученных результатов.

Следующий этап обработки — уменьшение размерности исходных данных. На данном этапе были рассмотрены два подхода: исключение пептидов, не несущих существенной информации, и выбор информативных пептидов.

Информативными назовем такие пептиды, которые имеют некоторое закономерное различие между классами исследуемых объектов.

Для исключения неинформативных пептидов использовался метод, описанный в [4]. Его суть состоит в отбрасывании пептидов, светимость которых ниже определенного порога, рассчитанного по следующей формуле:

$$t = \mathrm{M}(\mathrm{EMPTY}) + 2 * \mathrm{SD}(\mathrm{EMPTY}), \qquad (5)$$

где t — порог светимости; EMPTY — вектор пептидов, заведомо не несущих никакой информации; M — операция вычисления среднего значения; SD — стандартное отклонение.

Данным методом удалось уменьшить размерность с 10368 пептидов до 10026, т. е. примерно на 3,3 %, что не является существенным.

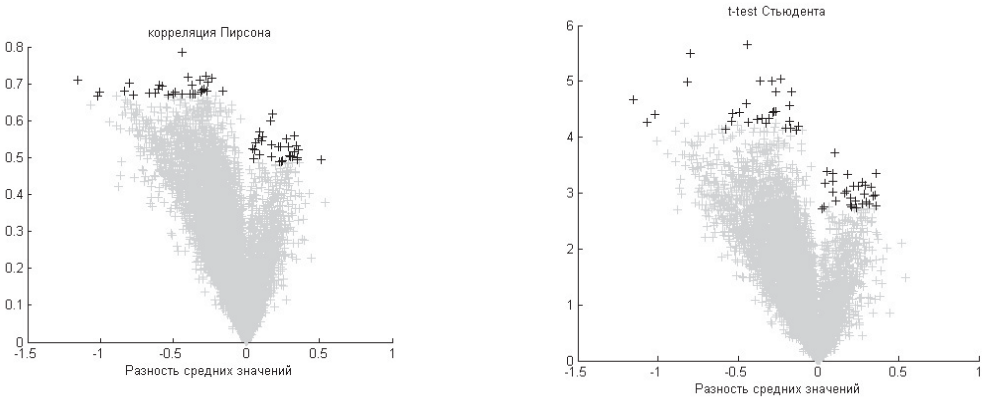

Рис. 2. Упорядочивание пептидов по значению критериев Пирсона и Стьюдента

Следующий подход — выбор информативных пептидов с применением критериев проверки гипотез. Были выбраны t-критерий Стьюдента и критерий корреляции Пирсона. Предполагая, что информативность пептида напрямую зависит от значений критериев, мы упорядочили все множество пептидов (см. рис. 2). В итоге имеем возможность выбрать m пептидов, которые в соответствии с принятым критерием наилучшим образом описывают различие классов объектов.

Для обоснования выбора количества информативных пептидов m была проведена классификация тестовых данных описанными выше способами. При этом m изменялось в пределах от 1 до 200 с шагом 1. Количество пептидов, при которых классификатор i работает наилучшим образом, выбирается исходя из решения следующей задачи оптимизации:

$$m_i = \arg\min_m \left(\sum_j E_{ij}(m) \right), \qquad (6)$$

где m_i — оптимальное количество пептидов для i-го классификатора; $E_{ij}(m)$ — ошибка распознавания классификатором i объекта j при использовании m информативных пептидов.

Таким образом, в статье представлены результаты вычислительных экспериментов, направленных на исследование методов обработки многомерных данных пептидных микрочипов. Одним из основных является следующий результат: каждому из рассмотренных классификаторов для получения оптимальных оценок нужен свой алгоритм нормализации. Так, с точки зрения чувствительности, специфичности и точности наилучшим методом предварительной обработки имевшихся многомерных данных оказалась нормализация по формуле (4) с последующей их классификацией методом наименьших квадратов.

Изложенный в данной статье подход к обработке данных пептидных микрочипов является основой для проведения дальнейших исследований. Одним из возможных путей развития технологии является применение нестатистических методов анализа и методов интервального анализа, а также методов, изложенных в работах [5, 6].

Библиографический список

1. Stafford P., Cichacz Z., Woodbury N., Johnston S.A. Immunosignature System for Diagnosis of Cancer // PNAS. — 2014. DOI:10.1073/pnas.1409432111.

2. Форматы файлов GenePix (GenePix® File Formats) [Электронный ресурс]. — URL: http://mdc.custhelp.com/app/answers/detail/a_id/18883/~/genepix%C2%AE-file-formats.

3. Sykes K., Legutki J.B., Stratford P. Immunosignaturing: a critical review // Cell Press. — 2012. DOI:10.1016/j.tibtech.2012.10.012.

4. Cretich M., Chiari M. Peptide Mycroarrays. Methods and Protocols // Humana Press. — 2009. DOI:10.1007/978-1-60327-394-7.

5. Максимов А.В., Оскорбин Н.М. Многопользовательские информационные системы: основы теории и методы исследования. — Барнаул, 2013.

6. Оскорбин Н.М. Математические модели систем с латентными переменными // Известия Алт. гос. ун-та. — 2012. — №1/2 (73).