2Н

УДК 004.8

Н.А. Новоселова, И.Э. Том
Объединенный институт проблем информатики НАН Беларуси, г. Минск
Беларусь, 220012, г. Минск, ул. Сурганова, 6, *novosel@newman.bas-net.by*

Алгоритм ранжирования признаков для обнаружения биомаркеров в данных генной экспрессии

N. Novoselova, I. Tom
United Institute of Informatics Problems National Academy of Sciences of Belarus, Minsk
Belarus, 220012, Minsk, Surganova 6, novosel@newman.bas-net.by

Algorithm of Feature Ranking for Biomarker Discovery in Gene Expression Data

Н.А. Новоселова, І.Е. Том
Об'єднаний інститут проблем інформатики НАН Білорусі, м. Мінськ
Білорусь, 220012, м. Мінськ, вул. Сурганова, 6, *novosel@newman.bas-net.by*

Алгоритм ранжирування атрибутів для виявлення біомаркерів у даних генної експресії

В статье рассматривается алгоритм ранжирования генов, полученных с использованием технологии микрочипов. Вектор рангов рассчитывается путем проведения классификаций случайных выборок из анализируемого набора данных. На каждой последующей итерации алгоритма ранг генов, участвующих в успешной классификации, повышается. В отличие от ранее используемых подходов, алгоритм позволяет повысить обобщающие свойства классификационных моделей за счет построения сбалансированных обучающих выборок, а также учесть информативность комбинации генов путем оценки их подмножеств.
Ключевые слова: ранжирование признаков, биомаркер, классификация, генная экспрессия.

The article considers the gene ranking algorithm for the microarray data. The rank vector is estimated by classifications of the random data samples. At each iteration the ranks of genes participating in the successful classification become higher. Unlike other methods of feature selection the proposed algorithm allows to increase the generality of the classification models by the construction of the balanced training samples and to take into account the descriptiveness of the gene combinations by the subsets estimation.
Ключевые слова: feature ranking, biomarker, classification, gene expression.

У статті розглядається алгоритм ранжирування генів, отриманих з використанням технології мікрочіпів. Вектор рангів розраховується шляхом проведення класифікацій випадкових вибірок з аналізованого набору даних. На кожній подальшій ітерації алгоритму ранг генів, що беруть участь в успішній класифікації, підвищується. На відміну від раніше використовуваних підходів алгоритм дозволяє підвищити узагальнювальні властивості класифікаційних моделей за рахунок побудови збалансованих навчальних вибірок, а також врахувати інформативність комбінації генів шляхом оцінки їх підмножин.
Ключові слова: ранжирування атрибутів, біомаркер, класифікація, експресія генів.

Введение

Улучшение диагностики онкологических заболеваний является одной из первостепенных задач, решаемых в области здравоохранения. Поставленный диагноз яв-

ляется основой для проведения и разработки курса терапии с целью достижения максимальной эффективности лечения с одновременной минимизацией токсичности. В последнее время большое внимание уделяется исследованиям генетической информации раковых клеток для уточнения существующих и выявления новых подтипов (классов) онкологических заболеваний на молекулярном уровне, что будет способствовать эффективности последующего лечения. Технология микрочипов представляет собой новый класс биотехнологий, она позволяет измерять одновременно уровни экспрессии тысяч генов, что дает богатый материал для более полного понимания различий между подтипами опухолей на молекулярном уровне и, следовательно, для осуществления более достоверной их классификации. Данные, полученные с помощью технологии микрочипов, используются для определения генов с повышенной или пониженной экспрессией подтипов онкологических заболеваний, а также с целью выявления новых биомаркеров для клинической диагностики и прогноза [1], [2]. К сожалению, извлечение зависимостей из такого рода данных затрудняется наличием несоответствия между количеством случаев (обычно менее ста) и количеством генов (обычно десятки тысяч), что приводит к переобучению, т.е. построению классификационной модели с низкими обобщающими свойствами (низкая точность на независимой тестовой выборке) [3]. Кроме того, полученные данные, как правило, содержат как ошибки измерений, так и систематические ошибки, что может значительно повлиять на точность классификации. Поэтому классификации, так же как и кластеризации, должна предшествовать предобработка признакового пространства, а именно осуществление отбора признаков (генов) путем ранжирования или сокращения их количества [4].

Методы отбора признаков для задачи диагностики онкологических заболеваний, как и в общем для всех задач классификации, можно разделить, в первую очередь, на две группы: фильтровочные и упаковочные методы [5]. Фильтровочные методы позволяют удалять неинформативные признаки согласно общим характеристикам данных. Они более просты в вычислительном плане и не привязаны к конкретному алгоритму классификации. Во встроенных методах выбранный классификатор участвует в построении подмножества признаков, тем самым, позволяя повысить точность классификации. Недостатком встроенных методов является высокая вычислительная сложность в связи с необходимостью на каждой итерации отбора применять классификационный алгоритм. Для подмножества информативных генов, отбираемых из данных генной экспрессии, применяются как фильтровочные методы, например, стандартные параметрические тесты (t-тест [6]), непараметрические тесты (знаково-ранговый тест Вилкоксона [7]), так и упаковочные [8], [9]. Однако, зачастую, у разных авторов подмножества отбираемых генов различаются, что может быть результатом как несоответствия размерностей анализируемых данных, так и игнорирования зависимостей между генами. Таким образом, на сегодняшний день не существует однозначных рекомендаций по выбору методов отбора подмножества информативных генов [10].

Цель данной статьи – разработка нового алгоритма отбора признаков, который позволяет преодолеть недостатки как фильтровочного, так и упаковочного подходов, и основан на ранжировании генов путем классификаций повторных случайных выборок из исходного набора данных. Предлагаемый алгоритм обладает следующими преимуществами: позволяет избежать переобучения за счет формирования обучающих матриц меньшей размерности, с преобладанием количества случаев над количеством генов; позволяет учитывать информативность комбинации генов путем оценки подмножества признаков классификационным алгоритмом.

В статье представлены результаты тестирования предложенного алгоритма на наборе данных по лейкемии, оценена биологическая значимость подмножества из 20-и генов с наибольшим рангом и проведен сравнительный анализ с результатами, полученными другими авторами.

Описание предложенного алгоритма

Пусть $X_{M\times N}$ представляет собой матрицу исходных данных экспрессии генов, где M – количество признаков (генов), а N – количество случаев или объектов данных, вектор $x_i = (x_i^1, x_i^2, \ldots, x_i^N), i = 1, M$ обозначает значения гена для N случаев или генный профиль. Каждый объект данных имеет метку класса, и в нашем исследовании соответствует одному из двух классов. Алгоритм автоматически масштабируется на обработку данных с большим количеством классов. Обозначим через $N_i, i = 1, 2$ количество объектов i-го класса, тогда $N = N_1 + N_2$.

Алгоритм осуществляет ранжирование признаков путем классификации повторных случайных выборок из матрицы исходных данных экспрессии генов $X_{M\times N}$. При этом выполняется процесс двойного отбора, т.е. случайным образом выбирается $m < M$ генов и $n < N$ объектов данных. Количество отбираемых объектов данных определяется параметром β, а количество генов такое, что $m \le n/2$. Такая схема отбора позволяет на k-й итерации формировать сбалансированную матрицу $Y_{m\times n}^k$, которую, согласно исследованиям авторов [3], менее вероятно переобучить. Количество итераций K выбирается таким образом, чтобы каждый ген анализируемого набора данных отбирался определенное количество раз, достаточное для адек-ватной оценки его ранга. Дополнительным критерием останова итерационного процесса, кроме количества итераций, является значение коэффициента корреляции вектора рангов на текущей и предыдущей итерации.

На каждом последующем шаге итерационного процесса, т.е. через определенное заданное количество итераций k_{iter}, формируется выходная матрица $E_{M\times 4}^k$, с количеством строк, равным исходному числу генов ($i = \overline{1, M}$), и четырьмя столбцами. Значение первого столбца T_i^k ($i = \overline{1, M}$) соответствует общему количеству раз, когда ген i включен в выборку для анализа после очередных k_{iter} итераций. Значение второго столбца S_i^k ($i = \overline{1, M}$) соответствует общему количеству раз, когда ген i отобран в состав успешной модели классификации. Под успешной моделью подразумевается модель, имеющая ошибку классификации меньше предварительно заданного значения параметра α. Значение третьего столбца $P_i^k = S_i^k / T_i^k$ ($i = \overline{1, M}$) соответствует предсказательной способности i-го гена и в четвертом столбце R_i^k ($i = \overline{1, M}$) содержится значение ранга i-го гена, рассчитанное на основе P_i^k после k_{iter} итераций. Гены с бóльшим значением P_i^k являются более информативными и, следовательно, имеют более высокий ранг.

Каждая итерация алгоритма $k = \overline{1, K}$ состоит из выполнения следующих шагов:

1. Для построения матрицы для анализа $Y_{m\times n}^k$ на k-й итерации применяется процедура двойного отбора строк и столбцов матрицы из $X_{M\times N}$. Матрица $Y_{m\times n}^k$ состоит

из m случайным образом отобранных генов из всего исходного множества M, n_1 случаев из множества N_1 исходных случаев первого класса и n_2 случаев из множества N_2 исходных случаев второго класса, таких что $n_1 / N_1 = n_2 / N_2 = \beta$, $n_1 + n_2 = n$ и $m < n$. Таким образом формируется обучающее множество, состоящее из n случаев, и тестовое множество, состоящее соответственно из $N - n$ случаев.

2. К обучающей выборке, заданной матрицей $Y^k_{m\times n}$, применяется классификационный алгоритм. В качестве классификационного алгоритма используется метод сжимающихся центроидов [11]. Процедура перекрестной проверки используется для поиска оптимального классификатора, который имеет минимальную ошибку на обучающей выборке $Y^k_{m\times n}$. Использование процедуры позволяет определить пороговое значение Δ^k, соответствующее наименьшей ошибке классификатора, полученной с использованием наименьшего подмножества генов $x_1, x_2, \ldots, x_l$, где $l \le m$. Полученная классификационная модель $C^k = f(x_1, x_2, \ldots, x_l)$ верифицируется на тестовом множестве из $N - n$ случаев. Ошибка классификации e рассчитывается следующим образом

$$e^k = \frac{FP + FN}{N - n},$$

где FP – ошибка первого рода, FN – ошибка второго рода.

В случае, если для всех возможных пороговых значений Δ^k и подмножеств генов $x_1, x_2, \ldots, x_l$, $l \le m$ минимальная ошибка классификатора, построенного на обучающей выборке, превосходит значение α, то верификация на тестовом множестве не выполняется и осуществляется переход к шагу 4.

3. Если на тестовом множестве ошибка классификационной модели на тестовом множестве $e^k \le \alpha$, т.е. меньше, чем предварительно заданный порог, то модель принимается как успешная и выполняется модификация выходной матрицы $E_{M\times 4}$. Для каждого x_i гена анализируемой на k-й итерации матрицы $Y^k_{m\times n}$ определяется соответствующий ген матрицы $E_{M\times 4}$ и значения столбцов T_i^k, S_i^k и P_i^k рассчитываются следующим образом:

$$T_i^k = \begin{cases} T_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_m) \\ T_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_m) \end{cases} \quad S_i^k = \begin{cases} S_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_l) \\ S_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_l) \end{cases}. \qquad P_i^k = S_i^k / T_i^k$$

4. Если на тестовом множестве ошибка классификационной модели $e^k > \alpha$, классификационная модель считается неуспешной и отобранные на k-й итерации гены являются неинформативными. Модель в этом случае можно определить как переобученную на обучающей выборке. Значения столбцов T_i^k, S_i^k и P_i^k выходной матрицы $E_{M\times 4}$ для каждого гена из отобранного подмножества $(x_1, x_2, \ldots, x_m)$ рассчитываются как

$$T_i^k = \begin{cases} T_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_m) \\ T_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_m) \end{cases} \quad S_i^k = S_i^{k-1} \quad P_i^k = S_i^k / T_i^k.$$

Значения рангов определяются путем сортировки прогностических значений генов P_i^k, $i = \overline{1, M}$ в порядке убывания и заносятся в столбец R_i^k, $i = \overline{1, M}$ матрицы $E_{M\times 4}$.

Каждые k_{iter} итераций оцениваются изменение в ранжировании генов путем расчета коэффициента ранговой корреляции Спирмена между векторами R^k и R^{k*} на k-й и $k*$-й итерациях, где $k* = k - k_{iter}$:

$$T_i^k = \begin{cases} T_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_m) \\ T_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_m) \end{cases} \quad S_i^k = S_i^{k-1} \quad P_i^k = S_i^k / T_i^k .$$

Значения рангов определяются путем сортировки прогностических значений генов P_i^k, $i = \overline{1, M}$ в порядке убывания и заносятся в столбец R_i^k, $i = \overline{1, M}$ матрицы $E_{M\times 4}$.

Каждые k_{iter} итераций оцениваются изменения в ранжировании генов путем расчета коэффициента корреляции Спирмена между векторами рангов R^k и R^{k*} на k-й и $k*$-й итерациях, где $k* = k - k_{iter}$:

$$r = 1 - 6\sum_{i=1}^{M} \frac{\left(R_i^k - R_i^{k*}\right)^2}{M\left(M^2 - 1\right)}.$$

Алгоритм прекращает свою работу в случае, если коэффициент корреляции $r > \gamma$ (например $\gamma = 0.99$), т.е. когда ранговая последовательность стабилизируется.

Результаты экспериментов

Для тестирования предложенного алгоритма ранжирования признаков и проведения сравнительного анализа результатов в нашем исследовании использовался набор данных пациентов с лейкемией [12]. Набор данных включает два типа острой лейкемии: острая лимфобластная форма (ALL) и острая миелоидная лейкемия (AML). Набор состоит из 47 случаев лейкемии ALL (38 случаев B-cell лейкемии и 9 случаев T-cell лейкемии) и 25 случаев AML, которые характеризуются экспрессией 7129 генов. Согласно протоколу, описанному в [13], были выполнены следующие шаги предобработки данных: пороговая обработка с наименьшим значением экспрессии, равным 100 и наибольшим – 1600; фильтрация с исключением генов со значениями экспрессии, удовлетворяющими одному из следующих условий: $\max/\min < 5$ или $(\max - \min) \le 500$, где max – максимальное значение и min – минимальное значение экспрессии гена; логарифмическая трансформация значений экспрессии генов. После выполнения предобработки было отобрано 3571 генов. Набор данных можно скачать по ссылке [14].

Для проведения эксперимента были выбраны следующие параметры: количество итераций $K = 300000$, пороговое значение ошибки классификации $\alpha = 0.2$, значение доли случаев, случайным образом отбираемых из исходного набора на каждой итерации $\beta = 0.7$, количество итераций между выполнением расчета критерия останова k_{iter}=20000 путем оценки коэффициента корреляции с пороговым значением $r = 0.09$. Каждые k_{iter} оценивалась выходная матрица $E_{M\times 4}$. В табл. 1 представлены 20 генов с наибольшим рангом, полученные в результате эксперимента.

Для каждого из 20 подмножеств, начиная с одноэлементного, состоящего из гена с наивысшим рангом, и далее путем добавления последующего гена в ранжированном списке, была построена классификационная модель и оценена ошибка классификации всего исходного набора данных. В качестве наиболее информативного подмножества, которое может рассматриваться как набор потенциальных биомаркеров заболевания, было отобрано подмножество с наименьшей ошибкой

классификации. На рис. 1 представлены значения ошибки классификации, соответствующие количеству отобранных генов от 2 до 20. Согласно полученным значениям ошибки, рассчитанным с использованием процедуры перекрестной проверки на всем исходном наборе данных, в качестве биомаркеров заболевания можно выбрать подмножество, состоящее из четырех генов (Cd33, Zyxin, APLP2, CST3), с точностью классификации 98,6%. Необходимо отметить, что подмножество из 10 генов дает несколько меньшую ошибку классификации, однако требует определения значений экспрессии большего количества генов.

Таблица 1 – Описание генов с наибольшим рангом

№	Обозначение гена	Описание	№	Обозначение гена	Описание
1	Cd33	Рекомбинантный белок человека	11	DNTT	Терминальная диоксинуклеотид-трансфераза
2	Zyxin	Циксин	12	MARCKSL1	MARCKS-подобный протеин
3	APLP2	Предшественник бета-амилоида	13	CD79A	Полипептид Ig-альфа (иммуно-глобулин ассоциированная альфа)
4	CST3	Цистатин С	14	CTSA	Катепсин А
5	MGST1	глутатион-S-трансферазы	15	CSTA	Цистатин А
6	CTSD	Катепсин D	16	SERPINB1	Серпин В1
7	CFD	Адипсин	17	FAH	Фумарилацетоацетат гидролаза
8	CCND3	Циклин D3	18	CFP	Пропердин – глобулярный белок сыворотки крови человека и млекопитающих животных.
9	CD63	Мембранный белок, гликопротеин из семейства тетраспанинов	19	VPREB1	Легкие цепи иммуноглобулинов. Ткане-специфичный ген, который неактивен в ES (эмбриональная стволовая клетка), но который активируется во время ранних стадий развития B-lymphocyte.
10	TCF3	Фактор транскрипции	20	SPTAN1	Спектрин альфа-цепь
			21	MPO	Миелопероксидаза

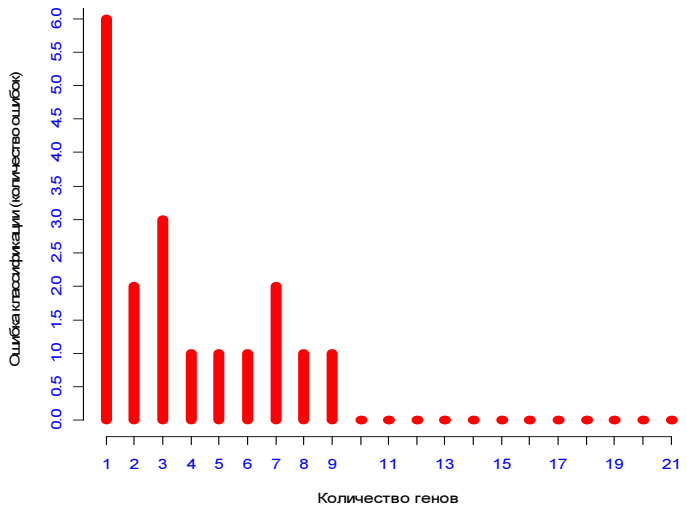

Рисунок 1 – Ошибка классификации для количества генов от 2 до 20

Сравнительный анализ результатов, полученных с использованием предложенного алгоритма, показал его эффективность при определении наиболее информативных генов, т.к. большее их количество присутствует в списках информативных генов других авторов [15-18], а сравнимая точность классификации (с использованием процедуры перекрестной проверки) достигается с использованием меньшего их количества (табл. 2).

Таблица 2 – Результаты сравнительного анализа различных методов отбора признаков

	Точность классификации	Количество генов
HBE метод [15]	97.2%	4 гена
Метод [16]	98.6%	132 гена
Метод [17]	98.6%	3 кластера генов (размер одного кластера от 1 до 23 генов)
Предложенный алгоритм	98.6%	4 гена

Анализ биологической значимости отобранных генов

Для того чтобы оценить биологическую значимость полученных результатов ранжированное подмножество из 20 наиболее информативных генов было проанализировано с использованием инструмента анализа биологических процессов InnateDB (http://www.innatedb.ca). Была выявлена их потенциальная функциональная связь с фенотипом заболевания и биологические пути, в которых они участвуют. Значимость участия комбинации генов в определенном биологическом процессе оценивалась путем расчета соответствующего *p*-уровня. Согласно полученным результатам анализа два (адипсин CFD и комплементарный фактор пропердин CFP) из 20 генов участвуют в альтернативном пути системы комплемента, подавлении инфекционных агентов (*p*-уровень = 0.00006). В работе [18] адипсин был также выбран в качестве биомаркера подтипа лейкемии, и была описана его роль в дифференциации миелоидных клеток. Два гена (CD33 и DNTT) из 20 связаны с процессом дифференцировки гемопоэтических клеток (*p*-уровень = 0.00429). CD33 кодирует мембранный протеин клетки и сильно экспрессирован на поверхности лейкемических бластов, применяется для дифференциации миелоидных и лимфоидных клеток [19]. DNTT (деоксинуклеотидтрансфераза) экспрессируется в популяциях нормальных и злокачественных пре-В и пре-Т лимфоцитах во время ранних стадий дифференцировки.

Два гена (циклин D3 и CD79A) участвуют в биологическом процессе клеточного цикла и сигнального пути В-клеточого рецептора (p-уровень = 0.01489). Также два гена циклин D3 и циксин участвуют в процессе фокусной адгезии (*p*-уровень = =0.0225). В работе [20] циксин был отобран как один из биомаркеров фенотипов лейкемии. Уровень его экспрессии играет важную роль в дифференцировке подтипов лейкемии, однако его прямая связь в гемопоэзе не выявлена. Согласно последним исследованиям циксин-кодируемые протеины могут регулировать транскрипцию гена путем взаимодействия с фактором транскрипции. Таким образом, наиболее информативные 20 генов с наибольшим рангом, полученные с использованием предложенного алгоритма, функционально значимы как для процесса дифференцировки гемопоэтических клеток, так и патогенеза онкологических заболеваний. Многие из отобранных нами генов отобраны также в качестве биомаркеров для распознавания подтипа лейкемии в ряде других исследований [18-20].

Оценка сходимости алгоритма

Для определения оптимальных прогностических значений каждого гена необходимо оценить все возможные выборки m случаев из исходной матрицы данных $X_{M\times N}$, а именно $C_M^m = (M-1)!/(m-1)!(M-m)!$ различных комбинаций, что является сложным в вычислительном плане с количеством вычислений, пропорциональным $O(M!)$. Используемый нами подход является приближенным, эвристическим и получает оценки прогностических значений генов путем повторного анализа случайных выборок m случаев из исходной матрицы данных $X_{M\times N}$. Следовательно, необходимо проанализировать сходимость предложенного алгоритма и способность получать достоверные оценки ранжированных прогностических значений. Для этого мы проанализируем вектор рангов для различного количества итераций алгоритма. Как показано на рис. 2 для 20 генов, расположенных вверху ранжированного списка, значения рангов на 20 000 и 40 000 итерациях достаточно сильно отличаются.

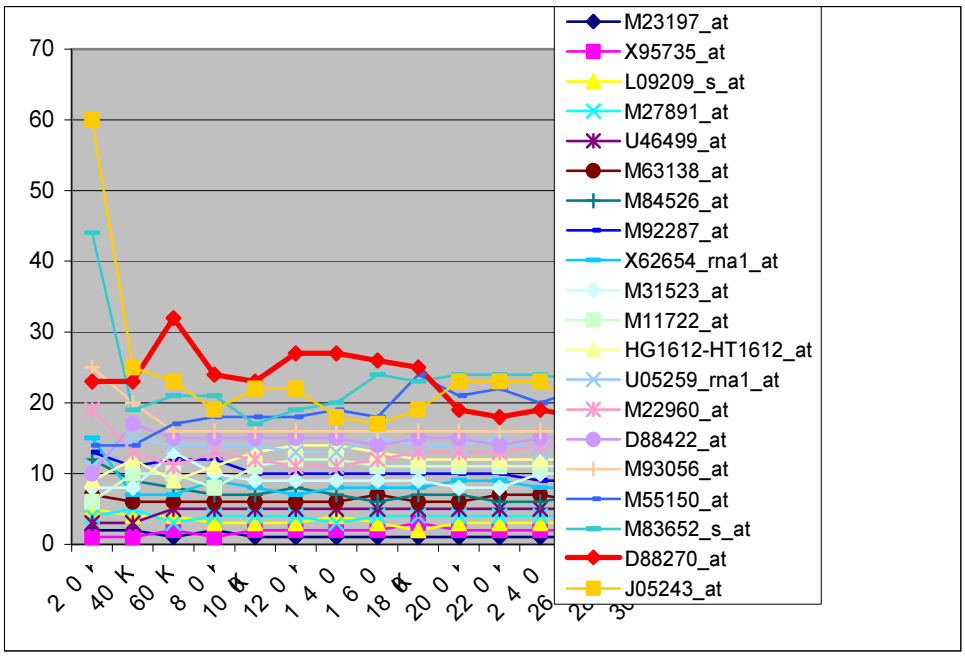

Рисунок 2 – Изменение рангов 20 генов в зависимости от количества итераций

Однако, после приблизительно 200 000 итераций значения рангов стабилизируются. Таким образом, предложенный алгоритм сходится к близкому к оптимальному решению, и может достаточно эффективно осуществлять ранжирование признаков, используя сравнительно небольшое количество итераций.

Выводы

Предложенный алгоритм ранжирования признаков позволяет выделять наиболее информативные гены в данных генной экспрессии, причем стабильность значений рангов отдельных генов обусловлена многократной оценкой выборок из исходной матрицы данных, что помогает избежать переобучения, и способствует получению несмещенных оценок вектора рангов. Согласно алгоритму на каждой итерации выполняется классификация случайным образом сформированной обучающей выборки с

верификацией классификационной модели на тестовой выборке. Точность классификации является показателем прогностической способности отобранных на очередной итерации комбинации генов. Выходная матрица количественно фиксирует как прогностическую способность, так и ранг отдельных генов, которые модифицируются на каждой последующей итерации, приближаясь к оптимальному ранжированию. Критерием оптимальности вектора рангов является стабилизация его значений для отдельных генов, что оценивается с использованием критерия ранговой корреляции Спирмена.

Предложенный алгоритм протестирован на наборе данных по лейкемии и проведена оценка сходимости алгоритма на примере 20 генов, расположенных вверху ранжированного списка. Оценена биологическая значимость исследуемого подмножества отобранных генов, которая позволяет сделать вывод об их тесной связи с процессами, происходящими в лейкемических клетках. Таким образом, отобранные гены с наивысшим рангом не являются результатом случайного отбора. Сравнительный анализ с работами других авторов по исследуемому набору данных показал преимущество предложенного нами алгоритма, а именно в качестве результата отобрано наименьшее подмножество из четырех биомаркеров, обеспечивающих такую же или лучшую точность классификации.

Литература

1. Liu X. An entropy-based gene selection method for cancer classification using microarray data / X. Liu, A. Krishnan, A. Mondry // BMC Bioinformatics. – 2005. – Vol. 6, № 76.
2. Новоселова Н.А. Методы анализа данных генной экспрессии. Обзор и перспективные направления развития (Novoselova, N.A. Methods for gene expression analysis. Survey and perspective directions) / Н.А. Новоселова, И.Э. Том. – LAMBERT Academic Publishing GmbH&Co. – 2012. – 68 p. – ISBN 978-3-659-16145-2.
3. Dougherty E.R. Performance of feature selection methods / E.R. Dougherty, J. Hua, C. Sima // Curr Genomics. – 2009. – Vol. 10. – P. 365-374.
4. Wang Y. Gene selection from microarray data for cancer classification a machine learning approach / Y. Wang, I.V. Tetko, M.A. Hall // Comp Biol Chem. – 2005. – Vol. 29. – P. 37-46.
5. Kohavi R. Wrapper for feature subset selection / R. Kohavi, G. John // Artificial Intelligence. – 1997. – Vol. 97, № 1-2. – P. 273-324.
6. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles / [Thomas J.G., Olson J.M., Tapscott S.J., Zhao L.P.] // Genome Res. -2001. – Vol. 11. – P. 1227-1236.
7. Antoniadis A. Effective dimension reduction methods for tumor classification using gene expression data / A. Antoniadis, S. Lambert-Lacroix, F. Leblanc // Bioinformatics. – 2003. – Vol. 19. – P. 563-570.
8. Filter versus wrapper gene selection approaches in DNA microarray domains / [Inza I., Larranaga P., Blanco R., Cerrolaza, A.] // Artif. Intell. Med. – 2004. – Vol. 31, № 2. – P. 91-103.
9. Xiong M. Biomarker identification by feature wrappers / M. Xiong, Z. Fang, J. Zhao // Genome Research. -2001. – Vol. 11. – P. 1878-1887.
10. Saeys Y. A review of feature selection techniques in bioinformatics / Y. Saeys, I. Inza, P. Larranaga // Bioinformatics. – 2007. – Vol. 23. – P. 2507-2517.
11. Diagnosis of multiple cancer types by shrunken centroids of gene expression / [Tibshirani R., Hastie T., Narasimhan B., Chu G.] // Proc Natl Acad Sci U S A. – 2002. – Vol. 99. – P. 6567-6572.
12. Molecular classification of Cancer:class discovery and class prediction by gene expression monitoring / [Golub T.R., Slonim D.K., Tamayo P. et al.] // Nature. – 1999. – Vol. 286. – P. 531-537.
13. Dudoit S. Comparison of discrimination methods for the classification of tumors using gene expression data / S. Dudoit, J. Fridlyand, T. Speed // J Am Stat Assoc. - 2002. – Vol. 97. – P. 77-87.
14. Whitehead Institute Center for Genomic Research: cancer genomics [Электронный ресурс]. – Режим доступа : http://www-genome.wi.mit.edu/cancer
15. Optimization Based Tumor Classification from Microarray Gene Expression Data / [Dagliyan O., Uney-Yuksektepe F., Kavakli I.H., Turkay M.] ; [Электронный ресурс] // PLoS ONE. – 2011. – № 6(2). – Режим доступа : e14579. doi:10.1371/journal.pone.0014579.

16. Optimization models for cancer classification extracting gene interaction information from microarray expression data / [Antonov A., Tetko I.V., Mader M.T. et al.] // Bioinformatics. – 2004. – Vol. 20. – P. 644-652.
17. Dettling M. Supervised clustering of genes / M. Dettling, P. Buhlmann [Электронный ресурс] // Genome Biol. – 2002. – Vol. 3. – Режим доступа : research0069.1–0069.15.
18. Biomarker discovery in microarray gene expression data with gaussian processes / [Chu W., Ghahramani Z., Falciani F., Wild D.L.] // Bioinformatics. – 2005. – Vol. 21. – P. 3385-3393.
19. Yang A.J. Bayesian variable selection for disease classification using gene expression data / A.J. Yang, X.Y. Song // Bioinformatics. – 2010. – Vol. 26. – P. 215-222.
20. Gene selection from microarray data for cancer classification – a machine learning approach / [Y. Wang et al.] // Comput. Biol. Chem. – 2005. – Vol. 29, № 1. – P. 37-46.

Literatura

1. Liu X. An entropy-based gene selection method for cancer classification using microarray data / X. Liu, A. Krishnan, A. Mondry // BMC Bioinformatics. – 2005. – Vol. 6, № 76.
2. Novoselov NA Methods for analysis of gene expression data. Overview and prospects for development (Novoselova, NA Methods for gene expression analysis. Survey and perspective directions) / N. Novoselov, IE Tom. - LAMBERT Academic Publishing GmbH & Co. - 2012. - 68 p. - ISBN 978-3-659-16145-2.
3. Dougherty E.R. Performance of feature selection methods / E.R. Dougherty, J. Hua, C. Sima // Curr Genomics. – 2009. – Vol. 10. – P. 365-374.
4. Wang Y. Gene selection from microarray data for cancer classification a machine learning approach / Y. Wang, I.V. Tetko, M.A. Hall // Comp Biol Chem. – 2005. – Vol. 29. – P. 37-46.
5. Kohavi R. Wrapper for feature subset selection / R. Kohavi, G. John // Artificial Intelligence. – 1997. – Vol. 97, № 1-2. – P. 273-324.
6. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles / [Thomas J.G., Olson J.M., Tapscott S.J., Zhao L.P.] // Genome Res. -2001. – Vol. 11. – P. 1227-1236.
7. Antoniadis A. Effective dimension reduction methods for tumor classification using gene expression data / A. Antoniadis, S. Lambert-Lacroix, F. Leblanc // Bioinformatics. – 2003. – Vol. 19. – P. 563-570.
8. Filter versus wrapper gene selection approaches in DNA microarray domains / [Inza I., Larranaga P., Blanco R., Cerrolaza, A.] // Artif. Intell. Med. – 2004. – Vol. 31, № 2. – P. 91-103.
9. Xiong M. Biomarker identification by feature wrappers / M. Xiong, Z. Fang, J. Zhao // Genome Research. -2001. – Vol. 11. – P. 1878-1887.
10. Saeys Y. A review of feature selection techniques in bioinformatics / Y. Saeys, I. Inza, P. Larranaga // Bioinformatics. – 2007. – Vol. 23. – P. 2507-2517.
11. Diagnosis of multiple cancer types by shrunken centroids of gene expression / [Tibshirani R., Hastie T., Narasimhan B., Chu G.] // Proc Natl Acad Sci U S A. – 2002. – Vol. 99. – P. 6567-6572.
12. Molecular classification of Cancer:class discovery and class prediction by gene expression monitoring / [Golub T.R., Slonim D.K., Tamayo P. et al.] // Nature. – 1999. – Vol. 286. – P. 531-537.
13. Dudoit S. Comparison of discrimination methods for the classification of tumors using gene expression data / S. Dudoit, J. Fridlyand, T. Speed // J Am Stat Assoc. - 2002. – Vol. 97. – P. 77-87.
14. Whitehead Institute Center for Genomic Research: cancer genomics [Электронный ресурс]. – Режим доступа : http://www-genome.wi.mit.edu/cancer
15. Optimization Based Tumor Classification from Microarray Gene Expression Data / [Dagliyan O., Uney-Yuksektepe F., Kavakli I.H., Turkay M.] ; [Электронный ресурс] // PLoS ONE. – 2011. – № 6(2). – Режим доступа : e14579. doi:10.1371/journal.pone.0014579.
16. Optimization models for cancer classification extracting gene interaction information from microarray expression data / [Antonov A., Tetko I.V., Mader M.T. et al.] // Bioinformatics. – 2004. – Vol. 20. – P. 644-652.
17. Dettling M. Supervised clustering of genes / M. Dettling, P. Buhlmann [Электронный ресурс] // Genome Biol. – 2002. – Vol. 3. – Режим доступа : research0069.1–0069.15.
18. Biomarker discovery in microarray gene expression data with gaussian processes / [Chu W., Ghahramani Z., Falciani F., Wild D.L.] // Bioinformatics. – 2005. – Vol. 21. – P. 3385-3393.
19. Yang A.J. Bayesian variable selection for disease classification using gene expression data / A.J. Yang, X.Y. Song // Bioinformatics. – 2010. – Vol. 26. – P. 215-222.
20. Gene selection from microarray data for cancer classification – a machine learning approach / [Y. Wang et al.] // Comput. Biol. Chem. – 2005. – Vol. 29, № 1. – P. 37-46.

RESUME

N. Novoselova, I. Tom

Algorithm of Feature Ranking for Biomarker Discovery in Gene Expression Data

In the paper the ranking algorithm of gene expression data obtained from microarray measurements is considered. The proposed algorithm allows to select the most informative genes by ranking, where the stability of the ranks of individual genes is ensured by estimation of multiple samples from initial data matrix. Such an approach helps to avoid overfitting and enables the unbiased estimate of the vector of ranks. At each iteration the classification model constructed on the randomly generated training sample is verified on the test sample. The classification accuracy is the indicator of the prognostic ability of the individual genes. After the successful classification the ranks of the participating genes become higher, meanwhile the search for the optimal ranking is performed not for each individual gene, but the whole combination. The output matrix is modified after pre-determined number of iterations, registering both the prognostic ability and the ranks of the individual genes. The stability of the rank vector serves as an optimality criterion and is estimated by computing Spearman correlation coefficient between the current and previous rank order.

The proposed algorithm has been tested on the leukemia dataset and its convergence is analyzed, considering the twenty top-ranked genes. The analysis of the biological significance of the investigated gene subset allows to confirm its obvious functional relevance to the phenotype it predicts and the processes, taking place in the leukemic cells. It assures that the top-ranked genes are highly unlikely to be selected by chance. The comparative analysis of the developed algorithm on the leukemia dataset shows its advantage over analogs, notably the selected set of biomarkers is smaller, consisting of four genes, which provide similar or higher classification accuracy and preserve the high classification accuracy.

Статья поступила в редакцию 02.04.2013.