РАЗРАБОТКА АЛОГОРИТМОВ АНАЛИЗА БИОЧИПОВ ДНК С УЧЕТОМ ВЕСОВЫХ ФАКТОРОВ КАЧЕСТВА СПОТОВ

Климук И.В., Свидрицкий А.С., Яцков Н.Н.
Кафедра системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белоруссикй государственный университет
Минск, Республика Беларусь
E-mail: ivanklimuk96@gmail.com

В работе предложены три модификации алгоритма k-ближайших соседей для анализа биочипов ДНК с учетом весовых факторов качества спотов. Представлены результаты сравнения алгоритмов k-ближайших соседей без учета и с учетом параметра качества спотов на примерах смоделированных данных.

Введение

Биочипы или микрочипы ДНК используются для одновременного исследования экспрессии множества генов [1–2]. С помощью технологии биочипов ДНК можно за короткое время обнаружить различные онкологические заболевания, туберкулез и др. [3]. Часто обработка и анализ биочипа затруднены вследствие низкого качества данных и высокого уровня экспериментального шума. Повысить эффективность алгоритмов анализа можно, если учесть параметр качества изображения каждого спота биочипа [4–6]. В работе выбран популярный и универсальный алгоритм классификации k-ближайших соседей.

Цель работы – разработка имитационной модели и алгоритмов k-ближайших соседей для классификации генов с учтом параметра качества спотов на биочипах ДНК.

I. Имитационная модель биочипа

В основе разработанной имитационной модели лежит алгоритм, представленный в [7]. Алгоритм:

Шаг 1. Задание параметров модели: N – число генов, m – число репликантов, p – доля невыраженных генов в выборке.

Шаг 2. Создание и заполнение матрицы M размером $N * m$: первые $N * p$ строк – значением 0, следующие $\frac{N*(1-p)}{2}$ строк – значением -1, последние $\frac{N*(1-p)}{2}$ строк – значением 1.

Шаг 3. Генерация вектора параметров качества спотов Q размером N с использованием бета-распределения с параметрами $a = 2.5$, $b = 3.5$. Данное распределение наиболее близко имитирует значения реальных экспериментов [5].

Шаг 4. Добавление нормального шума к данным:

$$M_i = M_i + rnorm * (1 - Q_i),$$

где $rnorm$ – реализация стандартной нормальной случайной величины.

Строки матрицы M – объекты или гены для классификации, вектор Q – вектор параметров качества каждого объекта.

II. Модификации метода k-ближайших соседей

В ходе анализа биочипов ДНК необходимо классифицировать гены на выраженные, невыраженные и подавленные (значение относительной экспрессии 0, 1, -1 соотвественно)

Разработаны и программно реализованы 3 модификации метода k-ближайших соседей с учетом весовых факторов качества спотов.

1) kNN с учетом параметра качества спота при расчете расстояния между объектами. Расстояния между объектами обучающей и тестируемой выборки:

$$d_{ij}^w = \frac{d_{ij}}{q_j},$$

где q_j – параметр качества j-го объекта обучающей выборки.

2) kNN с учетом параметра качества спота при голосовании. Параметр качества учитывается при назначении метки класса объекту тестируемой выборки. Класс-победитель выбирается по максимальной сумме их параметров качества:

$$Class(n_i) = arg \max_{c \in C} \sum_{j=1}^{l} [c_j = c] q_j,$$

3) kNN с учетом параметра качества спота при расчете расстояний между объектами и при голосовании. Алгоритм включает модификации п. 1) и 2)

III. Результаты

Эффективность работы алгоритмов оценивалась как процент ошибки при классификации всех генов, невыраженных и выраженных генов (представляющих наибольший интерес при анализе биочипов). Выполнено сравнение эффективности алгоритмов в зависимости от различных значений параметров обучающей и тестируемой выборок, а именно:

- N_L – размер обучающей выборки, менялся от 100 до 2000;

- p_L – относительная доля невыраженных генов в обучающей выборке, менялось от 0.05 до 0.95;

- $< Q >$ и $< Q_L >$ – средние значения параметров качества тестируемой и обучающей выборки соответственно, менялись от 0.1 до 0.9.

Зависимость ошибки классификации от параметров модели для разработанных алгоритмов представлены на рисунках 1–4. Модифицированные алгоритмы имеют меньшую ошибку классификации, чем классический алгоритм. На графиках , изображенных на рисунках 1–3, кривая 1 – kNN, 2 – kNN с измененным голосованием, 3 – kNN с пересчетом расстояний и кривая 4 – kNN с измененным голосованием и пересчетом расстояний.

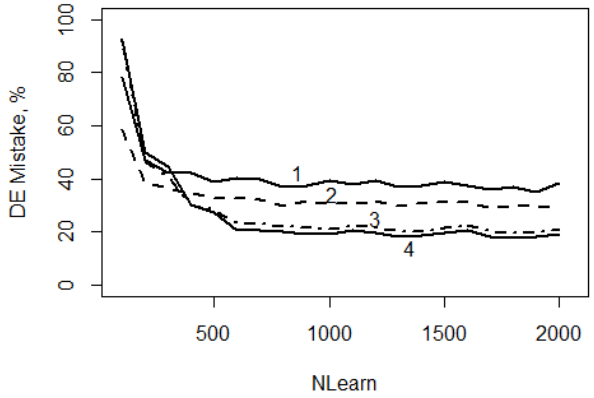

Рис. 1 – Зависимость ошибки классификации выраженных генов от размера обучающей выборки

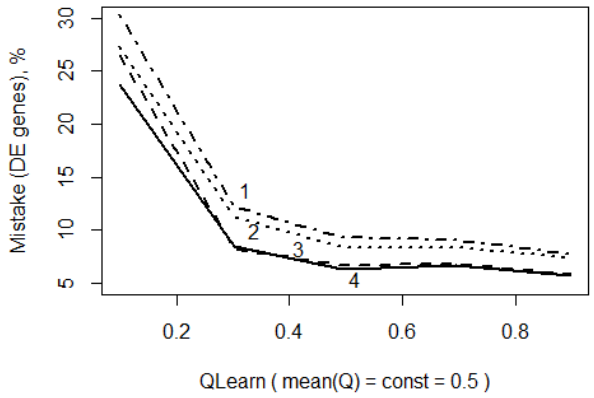

Рис. 2 – Зависимость ошибки классификации выраженных генов от среднего качества тестируемой выборки

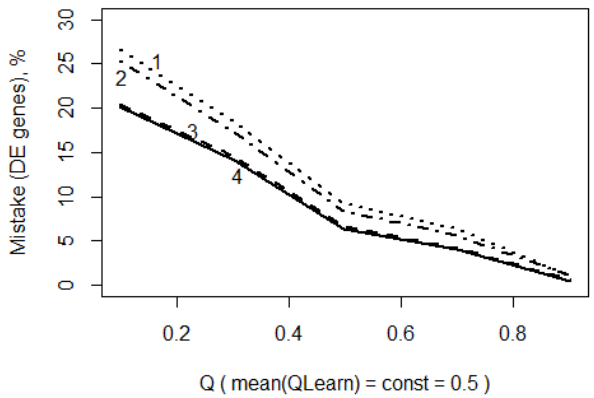

Рис. 3 – Зависимость ошибки классификации выраженных генов от среднего качества обучающей выборки

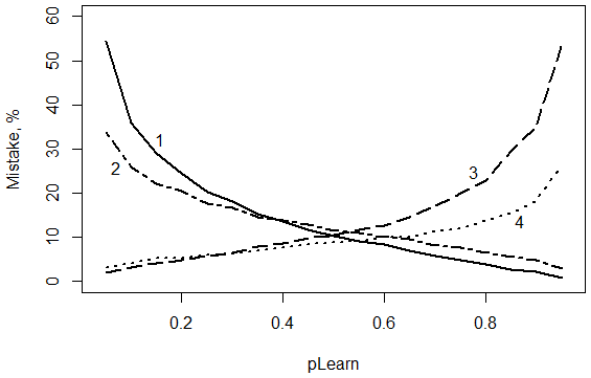

Рис. 4 – Зависимость ошибки классификации от доли невыраженных генов в обучающей выборке: 1 – невыраженных генов для kNN, 2 – невыраженных генов для kNN с измененным голосованием и пересчетом расстояний, 3 – выраженных генов для kNN, 4 – выраженных генов для kNN с измененным голосованием и пересчетом расстояний

IV. Выводы

Разработаны и реализованы метод k-ближайших соседей и три его модификации с учетом параметра качества спотов. Сравнительный анализ эффективности работы алгоритмов на смоделированных данных позволяет сделать выводы:

1) при размере обучающей выборки 500 и более ошибка классификации для всех алгоритмов не меняется;

2) чем хуже качество данных (как обучающей выборки, так и тестируемой), тем лучше с задачей классификации справляются модифицированные алгоритмы;

3) наилучший алгоритм - k-ближайших соседей с учетом параметра качества для расчета расстояний и при голосовании: ошибка классификации на 7% ниже, чем у классического алгоритма.

1. Okuzaki, D. Microarray and whole-exome sequencing analysis of familial Beh¸cet's disease patients/ D. Okuzaki et al – URL://www.ncbi.nlm.nih.gov/pmc/articles/PMC4726226/ (дата обращения: 08.06.2017).
2. Мирзобеков, А. Д. Биочипы в биологии и медицине 21го века //Вестник Российской Академии Наук. 2003. Т. 73. №5. С. 412.
3. Zou, J. Analysis of microarray-identified genes and microRNAs associated with drug resistance in ovarian cancer/ J. Zou, F. Yin, Q. Wang // International Journal of Clinical and Experimental Pathology. 2015.
4. Novikov, E. An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments/ E. Novikov, E. Barillot // BMC Bioinformatics. 2005.
5. Yatskou, M. Advanced spot quality analysis in two-colour microarray experiments/ M. Yatskou, E. Novikov, G. Vetter, A. Muller, E. Barillot,L. Vallar, E. Friederich. //BMC Research Notes (2008)
6. Novikov, E. Software package for automatic microarray image analysis (MAIA)/ E. Novikov, E. Barillot //BMC Bioinformatics. 2007. V. 5. P.639.
7. Dembele D. A Flexible Microarray Data Simulation Model // Microarrays. 2013. 2. P. 115-130.