

# О ЗАДАЧЕ КЛАССИФИКАЦИИ РАКОВЫХ ЗАБОЛЕВАНИЙ НА ОСНОВАНИИ ДАННЫХ О МЕТИЛИРОВАНИИ ДНК

*И.Б. Крылов<sup>1</sup>, М.В. Иванченко<sup>1</sup>, А.А. Заикин<sup>1,2</sup>*

<sup>1</sup>*Нижегородский государственный университет им. Н.И. Лобачевского*

<sup>2</sup>*Университетский колледж Лондона*

Кратко описаны подходы к изучению данных о метилировании ДНК и его связи с развитием онкологических заболеваний. Также изложена суть поиска наиболее значимых генов, шаблоны метилирования которых в значительной степени меняются при появлении ракового заболевания. Решается задача классификации, полученная точность составляет более 95 процентов.

## **Введение**

Сегодня очень важной задачей является создание новых механизмов ранней диагностики онкологических заболеваний. Ведь именно ранняя диагностика и своевременное начатое лечение заболевания в разы увеличивают шансы на выздоровление. В настоящее время диагноз ставится преимущественно на поздних стадиях, что приводит к плохому исходу.

Выявление новых факторов риска рака для разработки чувствительных диагностических методов невозможно без фундаментальных исследований в области системной биологии, биоинформатики и методов анализа данных. Однако по-настоящему комплексных исследований, которые включали бы все эти направления, пока не ведется. Уникальность нашего подхода состоит в разработке методов, основанных на системном исследовании данной внутригеномной и межгеномной структуры метилирования ДНК с наивысшим на сегодняшний день разрешением в 485 000 точек на геном.

## **1. Поиск наиболее значимых генов и мер метилирования**

В имеющихся данных содержится информация по 13 типам рака, данные взяты из базы данных с открытым доступом The Cancer Genome Atlas [1]. Будем рассматривать каждый тип рака в отдельности и решать задачу бинарной классификации, т.е. соотносить набор данных, содержащий значение меры метилирования для каждого гена, с классами «болен» и «здоров».

Решение данной задачи для каждого из 13 типов рака по отдельности является первым этапом исследований. Результатом решения данной задачи будет, во-первых, классификатор, способный давать ответ вопрос: «Болен ли этот человек раком конкретного типа?». Во-вторых, мы можем определить, шаблоны метилирования каких генов значительным образом различны, если пациенты принадлежат к разным группам «болен» или «здоров».

Будем решать задачу, рассматривая каждый ген в отдельности. Рассмотрим путь решения пошагово:

1. Исходные данные делим на обучающую и тестовую выборки.
2. Для каждого гена:

- a. Применяем метод оптимизации для настройки параметров логистической регрессии (стохастический градиентный спуск), максимизируя функцию правдоподобия. Для усадки параметров пользуемся регуляризацией.
  - b. Строим ROC-кривую (англ. receiver operating characteristic, операционная характеристика приёмника) [2].
  - c. Высчитываем AUC (англ. area under ROC curve, площадь под ROC-кривой) [3].
3. Сортируем список генов по убыванию их AUC.
  4. Устанавливаем минимальный проходной порог для генов  $AUC_{min} = 0.90$  и исключаем из списка те гены, которые имеют AUC меньше, чем  $AUC_{min}$ .

В результате выполнения такой последовательности действий для каждого типа рака, мы получаем индивидуальный список генов для каждого вида онкологии, шаблоны метилирования которых значительным образом различны (в соответствии с мерой метилирования) в случаях наличия и отсутствия заболевания.

## **2. Решение задачи классификации**

Более сложной задачей является задача отнесения конкретного набора данных об уровне внутригенного метилирования к классу «здоров» или «тип рака 1...13», в данном случае нас интересует, какое конкретное заболевание из списка имеет пациент. Для решения данной задачи используем модель «софтмакс»-регрессии [4], позволяющий оценить вероятность наличия того или иного вида заболевания. Для параметрической оптимизации используется метод стохастического градиентного спуска.

### **Заключение**

В результате работы был выявлен ряд наиболее эффективных мер из рассматриваемого набора и список генов, на метилирование которых стоило бы обратить особое внимание при решении задачи классификации. Был построен классификатор, который не только может говорить о наличии ракового заболевания, но и идентифицировать его тип с высокой точностью, более чем 95%.

### **Литература**

1. The Cancer Genome Atlas homepage. NCI and the NHGRI. Retrieved 2009-04-28.
2. Signal detection theory and psychophysics. – New York, NY: John Wiley and Sons Inc., 1966.
3. Lobo, Jorge M.; Jiménez-Valverde, Alberto; and Real, Raimundo (2008), AUC: a misleading measure of the performance of predictive distribution models // *Global Ecology and Biogeography*. 17: 145–151
4. Greene, William H., *Econometric Analysis*. Fifth edition. Prentice Hall, 1993. P. 720-723.