Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»
Кафедра Системного Программирования

Соковикова Светлана Алексеевна

Создание модуля для эффективной обработки генетических данных с использованием теста хи-квадрат

Курсовая работа

Научный руководитель:
к.ф.-м.н., доцент Малов С.В.

Санкт-Петербург
2016

Оглавление

Оглавление....2
Введение....3
Постановка задачи....4
Критерий хи-квадрат....5
Чтение и запись gds-файлов....7
Алгоритм вычисления....7
Инструменты для вычисления специальных статистических функций....8
Тестирование результата....9
Заключение....9
Список литературы....10

Введение

Тест хи-квадрат – статистический тест, широко используемый в биостатистике, в частности, для полногеномного анализа генетических ассоциаций. Специфика генетики – огромные размеры анализируемых данных (размер не самой большой таблицы - 750*1850000 чисел из диапазона 0:3), что сильно сказывается на времени вычислений. Удобный язык программирования для биостатистических расчетов, на котором реализовано множество статистических инструментов – R. R имеет низкую производительность из-за особенностей реализации транслятора, и, очевидно, не оптимален для работы с большими данными. Отсюда возникает идея реализовать наиболее востребованные статистические методы на более производительном языке, например, C.

Постановка задачи

Моя задача – реализовать программу, которая принимает на вход файл gds и возвращает другой файл gds-формата, содержащий результаты выполнения теста хи-квадрат (вектор pValue). Вычисления должны выполняться в модуле, написанном на языке C и вызываемом из R. Реализация этого модуля – ключевая цель моей работы. Полученная программа должна работать существенно быстрее, чем реализация на R.

Критерий хи-квадрат

Пусть (X,Y) – случайный вектор, $X \in \{1..d_1\}$, $Y \in \{1..d_2\}$.

$p_{ij}=P(X=i, Y=j)$, $p_{ij} \geq 0$, $\sum_{i,j} p_{ij} = 1$.

Условие независимости X и Y: $p_{ij} = p_{i\cdot}p_{\cdot j}$, где $p_{i\cdot} = \sum_j p_{ij}$, $p_{\cdot j} = \sum_i p_{ij}$.

Категориальная гипотеза независимости H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$ $\forall i \in \{1..d_1\}$, $\forall j \in \{1..d_2\}$.

Имеется выборка $(X_1,Y_1),\ldots,(X_n,Y_n)$ из распределения (X,Y), параметр $\theta = \{p_{ij}\}_{ij}$.

Введем частоты $\nu_{ij}=\sum_{ij} 1\{X_k=i, Y_k=j\}$.

Известно, что $\chi^2=\sum_{ij} \frac{(\nu ij - n\widehat{pi.}\widehat{p.j})^2}{n\widehat{pi.}\widehat{p.j}}$ имеет распределение χ^2 при справедливости гипотезы H_0, где $(\widehat{pi.}\,\widehat{p.j})$ – оценки максимального правдоподобия для $(p_{i\cdot},p_{\cdot j})$.

Вычисляем $\widehat{pi.}=\frac{\nu i.}{n}$, $\widehat{p.j}=\frac{\nu .j}{n}$, $\nu_{i\cdot}= \sum_j \nu_{ij}$, $\nu_{\cdot j}= \sum_i \nu_{ij}$.

Тогда $\chi^2=\sum_{ij} \frac{\nu ij - \frac{\nu i. \nu .j}{n}}{\frac{\nu i. \nu .j}{n}}$ при справедливости H_0 имеет распределение $\chi^2_{(d1-1)(d2-1)}$.

Статистический критерий $\varphi(\overrightarrow{x})$, постороенный на χ^2, принимает H_0, если $\chi^2 \leq x_\alpha$ и отвергает ее при $\chi^2 > x_\alpha$. Значение x_α выбирается исходя из распределения $\chi^2_{(d1-1)(d2-1)}$ таким образом, что $P_{H_0}(\chi^2 \leq x_\alpha)=1-\alpha$. pV значение – преобразование статистики критерия исходя из ее распределения при справедливости H_0:

$pV=1 - K_{(d1-1)(d2-1)}(\chi^2)$, где $K_{(d1-1)(d2-1)}$ – функция распределения $\chi^2_{(d1-1)(d2-1)}$. Согласно преобразованию Смирнова, распределение pV – равномерное U(0,1). Критерий может быть переписан в виде: $\varphi(\overrightarrow{x})$ принимает H_0, если $pV \geq \alpha$ и отвергает ее иначе. Значение pV часто используют в качестве показателя статистической значимости отклонения от H_0.

При проведении множества статистических тестов установлено, что стандартный уровень, ограничивающий ошибку I рода, применим для каждого теста в отдельности, но неприменим для всей совокупности тестов. Для выявления статистической значимости отклонения от нулевых гипотез в задаче

интерпретации результатов многочисленных статистических тестов принята поправка Бонферрони. В основе лежит неравенство Буля: $P(\bigcup_i Ai) \leq \sum_i P(Ai)$.

Выбор уровня значимости $\alpha_c = \alpha/m$, m – число статистических тестов, гарантирует, что вероятность ошибочного выявления хотя бы одного нарушения основной гипотезы не превосходит α, т.к.:

Пусть A_l – отвержение H_{ol}, l=1,…,m, $P_{Ho}(A_l)$ – вероятность ошибочного отвержения основной гипотезы в l-м тесте. Если $P(A_l) \leq \alpha/m$, то

$$P(\bigcup_{l=1}^{m} Al) \leq \sum_{l=1}^{m} P(Al) \leq m\frac{\alpha}{m} = \alpha .$$

Таким образом, важно иметь формулу, дающую высокую точность вычислений вероятностей $P(A_l)$ при H_o. В рассматриваемом случае это формула для $P(A_l)=P(\chi^2 \geq \chi\alpha/m)$ при больших $\chi\alpha/m$.

Чтение и запись GDS-файлов

GDS – специальный формат, используемый только в R-Bioconductor. В языке C нет стандартных функций для чтения и записи файлов этого формата, в то время как в R работа с GDS-файлами реализована оптимально и быстро. Поэтому решено читать и записывать GDS-файлы средствами R и передавать полученные данные модулю C, используя .C - стандартный интерфейс вызова модуля из R.

Алгоритм вычисления

1) Вектор phenotype длины n и таблицу genotype размером n строк *c столбцов модуль получает на входе из R. Элементы столбца – числа 1:k (для разных входных файлов значение k разное, в большинстве случаев k=2), элементы таблицы – числа 0:3. n~1000, c~1000000. Дальнейшие вычисления выполняются для ***каждого столбца*** таблицы ***отдельно***.
2)

	phen	gen
0	1	2
1	1	1
2	2	0
...		
j	u	v
n-1		

Составляется небольшая таблица contTable[1:k , 0:2], заполняемая по следующему принципу: в ячейке contTable[u,v] стоит количество таких j, для которых phen[j]=u и gen[j]=v. Значения, где gen[j]=3, просто отсеиваются, так как 3 означает, что генетические данные не определены и анализировать нечего. Эта таблица называется таблицей сопряженности. Дальнейшие операции выполняются именно с ней.

3) Далее вызывается функция, принимающей на вход таблицу contTable, ее размеры row и col, и не до конца заполненный вектор чисел степеней свободы.
4) Для contTable считаются суммы в строках sumrow и суммы в столбцах sumcol, сумма всех элементов total.
5) Число степеней свободы deg_of_freedom – важное значение, без которого не обойтись. Для таблицы размера r*c deg_of_freedom=(r-1)*(c-1). На этом шаге deg_of_freedom:=(row-1)*(col-1).
6) Проверка для очень часто встречающегося случая, когда contTable имеет размер 2*3.
   Если одна из строк contTable полностью заполнена нулями (один из двух

элементов sumrow нулевой), то статистические тесты неприменимы, возвращается значение NaN и счет для данного столбца gen закончен.
Если обе строки ненулевые, но два из трех столбцов contTable нулевые (два из трех элементов sumcol нулевые), то статистические тесты также неприменимы, возвращается значение NaN и счет для столбца gen заканчивается.
Если обе строки ненулевые и только один столбец contTable нулевой, то статистические тесты применяются к таблице contTable, из которой вычеркнут нулевой столбец. При этом число степеней свободы deg_of_freedom:=(row-1)*(col-2).

7) Далее вычисляется chisq как сумма $\frac{(contTable[i,j]-E[i,j])^2}{E[i,j]}$ по всем невычеркнутым клеткам contTable[i,j], где E[i,j]= $\frac{sumrow[i]*sumcol[j]}{total}$.
8) Работа функции закончена, управление передается вызвавшей ее программе. Программа получает chisq и deg_of_freedom.

Модуль возвращает вектор chisq, содержащий значения хи-квадрат статистики для каждого столбца таблицы genotype и вектор значений deg_of_freedom. Далее встроенной в R функцией pchisq вычисляется вектор значений pValue, а затем этот вектор упаковывается в новый gds-файл.

Инструменты для вычисления специальных статистических функций

В ходе алгоритма необходимо с высокой точностью вычислять значение «правых хвостов» функции распределения хи-квадрат. Реализация этого вычисления на C – отдельная сложная задача. В R это вычисление реализовано с требуемой точностью и оптимизировано по времени выполнения (функция pchisq). Отсюда возникает идея не писать эту функцию заново на языке C, что не даст существенного выигрыша во времени, а использовать функцию языка R.

Тестирование результата

В таблице приведено время выполнения вычислений на специальных серверах, предназначенных для биостатистических вычислений, для таблицы genotype размера 750*1829336. Время посчитано без учета чтения/записи gds-файлов, так как работа с ними выполняется средствами R и занимает в обоих случаях одинаковое время.

Средство	Время выполнения, с
Стандартные функции R	1944.494
Моя реализация	6.277

Заключение

Полученная в результате работы реализация существенно выигрывает по времени выполнения у реализации на языке R, что и было основным требованием, выдвинутым в постановке задачи.

Список литературы

1. Боровков А.А. Математическая статистика. СПб.: Лань, 2010. — 704 с.

2. Классические методы статистики: критерий хи-квадрат. –
URL: http://r-analytics.blogspot.ru/2012/08/blog-post.html#.VocMWOQ2xH4