УДК 57.087 **DOI:** 10.15622/sp.47.12

Л.В. Уткин, Ю.А. Жук

ПОЛНОГЕНОМНЫЙ ПОИСК АССОЦИАЦИЙ С ИСПОЛЬЗОВАНИЕМ МАТРИЦ ПАРНЫХ СРАВНЕНИЙ

Уткин Л.В., Жук Ю.А. **Полногеномный поиск ассоциаций с использованием матриц парных сравнений.**

Аннотация. Предлагается простой метод определения значимости объектов популяции при установлении ассоциации между однонуклеотидными полиморфизмами и количественными признаками в полногеномном поиске ассоциаций. На первом этапе сравниваются пары объектов популяции с точки зрения расстояния между ними по фенотипу и генотипу. На втором этапе строятся матрицы парных сравнений объектов и вычисляются веса объектов в соответствии с аддитивной и мультипликативной шкалами. Показывается, как можно модифицировать метод Лассо с использованием весов. Числовые эксперименты с реальными данными иллюстрируют предлагаемый метод.

Ключевые слова: биоинформатика, регрессия, полногеномный поиск ассоциаций, парные сравнения, Лассо, аддитивные и мультипликативные шкалы.

Utkin L.V., Zhuk Y.A. **A Genome-Wide Association Study using Pairwise Comparison Matrices.**

Abstract. In this paper, we propose a simple method for assigning importance weights to individuals of a population to determine the association between single nucleotide polymorphisms and quantitative traits in the genome-wide association study. At the first step, pairs of individuals in population are compared in terms of distances between phenotypes and genotypes. At the second step, pairwise comparison matrices of individuals are constructed, and the weights of individuals are computed with respect to additive and multiplicative scales. It is shown how to modify the Lasso method using the weights. Numerical experiments with real data illustrate the proposed method.

Keywords: bioinformatics, regression, genome-wide association study, pairwise comparisons, Lasso, additive and multiplicative scales.

1. Введение. Одним из элементов биоинформатики как совокупности методов и подходов, включающих в себя математические методы компьютерного анализа, алгоритмов и программ в геномике, является полногеномный поиск ассоциаций (a genome-wide association study — GWAS). С точки зрения биологии основная цель GWAS заключается в поиске зависимости между значениями фенотипа объектов некоторой популяции и их генотипом. Другими словами, GWAS определяет, какие элементы ДНК или молекулярные маркеры, называемые однонуклеотидными полиморфизмами (single nucleotide polymorphisms — SNP), оказывают наибольшее влияние на значения фенотипа. С математической точки зрения цель GWAS — это построение зависимости между SNP, которые рассматриваются как независимые переменные, и значениями фенотипа как зависимой переменной. С точки зрения теории машинного обучения, GWAS — это задача отбора признаков, в которой признаками являются SNP, а обучающая

выборка — объекты анализируемой популяции, каждый из которых характеризуется определенным значением фенотипа и множеством значений генотипа (значениями зависимых переменных).

Вследствие важности решения задачи определения значимых маркеров, к настоящему времени разработано огромное количество статистических моделей и алгоритмов. Детальные обзоры существующих подходов можно найти в работах [6, 7, 11, 13, 15], согласно которым большинство моделей и алгоритмов GWAS используют:

1. методы фильтрации [3, 15], которые основаны на статистических свойствах SNP-маркеров для отбрасывания наименее информативных маркеров (критерий Фишера, дисперсионный анализ и т.д.);
2. методы упаковки [9], например, известный метод рекурсивного удаления признаков [10];
3. вложенные методы [14, 16], которые основаны на построении таких регрессионных моделей, как метод Лассо, гребневая регрессия, эластичная сеть, устанавливающих зависимость между SNP (независимые переменные) и количественным признаком или фенотипом (зависимая переменная).

Многими авторами отмечается, что методы, основанные на построении регрессионных моделей, являются наиболее перспективными и интересными, так как построение этих моделей охватывает одновременно все SNP, устанавливая в результате решения определенные веса SNP, которые и являются показателями их значимости в ассоциации.

Однако, несмотря на большое количество моделей и алгоритмов GWAS, в частности регрессионных моделей, большинство из них сталкиваются с рядом сложностей, главной из которых является то, что число SNP-маркеров обычно в десятки или сотни раз превышает число объектов в популяции или в обучающей выборке. Другими словами, размерность задачи значительно превышает количество точек в пространстве признаков. Это существенно снижает доверие к полученным результатам, особенно наиболее интересным регрессионным моделям. Следует также отметить, что значения фенотипа являются случайными величинами, которые определяются не только генотипом, но и условиями внешней среды, например, количеством осадков, количеством солнечных дней, средней температурой, влияние которых в явном виде обычно не известно. Кроме того, во многих случаях исходные данные (матрицы генотипов) могут содержать пропущенные значения, что также снижает качество моделей.

Одним из подходов повышения качества моделей является предварительная обработка имеющихся данных с целью выявления каких-либо закономерностей в данных или их свойств, которые могли

бы использоваться в качестве дополнительной информации при построении стандартных моделей или алгоритмов. В работе [1] предложена модификация метода Лассо, используемого для полногеномного поиска ассоциаций, на примере анализа удвоенных гаплоидных линий ячменя для учета дополнительной информации о целевых значениях фенотипа, определяемого некоторым свойством растений. В рамках модификации формализуется дополнительная информация о свойствах растений в виде пересечения двух множеств весов, приписываемых элементам обучающей выборки. Первое множество образовано при помощи интервальной модели засорения. Второе множество весов образуется последовательностью парных сравнений значений фенотипа. Это позволяет повысить качество модели. Однако данная модификация не учитывает внутренней структуры информации, что делает ее применение не столь эффективным.

Поэтому в представленной работе предлагается простой метод определения значимости объектов популяции или обучающей выборки с точки зрения представительности того или иного объекта в установлении ассоциации между SNP и количественными признаками. Значимость представляется в виде весов объектов и может в дальнейшем использоваться при «взвешивании» элементов обучающей выборки в любых моделях GWAS. Метод основан на двух идеях. Первая была предложена в другом методе GWAS [2], где сравниваются пары объектов популяции с точки зрения расстояния между ними по фенотипу и генотипу. Вторая идея заключается в построении матрицы парных сравнений объектов специальным образом и вычислении весов объектов, например, по аналогии с известным методом анализа иерархий [12]. Предлагаемый метод может применяться не только к объектам биоинформатики, но также использоваться и в других областях, где решается задача поиска значимых элементов, например, в задачах надежности больших систем, где матрица состояний элементов являются аналогом матрицы генотипов, а состояния системы — значения фенотипа.

2. Анализ пар объектов популяции и вычисление их весов. Рассмотрим общее формальное определение задачи поиска ассоциаций с математической точки зрения. Пусть $\mathbf{X} = [X_1, ..., X_p]$ — матрица генотипов для n объектов и p SNP. Со статистической точки зрения SNP можно рассматривать как независимые переменные, т.е. $X_j = (x_{1j}, ..., x_{nj})^{\mathsf{T}}$ — переменная, представляющая j-ый SNP, $j = 1, ..., p$. Каждая переменная x_{ij} — аллель i-го объекта на локусе j-го SNP. Переменная может принимать значения $\{0,1\}$, где 0 и 1 определяются исходя из частоты аллелей. Генотип может быть представлен также

одним из чисел {0, 1, 2}. В частности, для представления гомозиготных аллелей используются обозначения ($AA = 0$) и ($aa = 2$), для гетерозиготных — ($Aa / aA = 1$). Необходимо отметить, что для значений аллелей x_{ij} не установлен порядок, и соответствующие признаки являются номинальными. Поэтому, с одной стороны, вместо чисел 0, 1 или 2 могут использоваться произвольные обозначения. С другой стороны, для построения регрессионной модели необходимо использовать числа 0, 1 и 2. Вектор аллелей, соответствующих *i*-му объекту, является столбцом матрицы генотипов $\mathbf{X}$ и будет обозначаться $\mathbf{x}_i^{\mathsf{T}} = (x_{i1},...,x_{ip})$, $i = 1,...,n.$ Также обозначим вектор значений фенотипа всех объектов $Y = (y_1,...,y_n)$, где $y_i \in \mathbf{R}$.

Таким образом, цель GWAS заключается в поиске значимых SNP в $\mathbf{X}$, от которых сильно зависит Y. Цель данной работы — назначить веса $W = (w_1,...,w_n)$ объектам популяции в соответствии с из значимостью в определении ассоциации между SNP и Y.

Первая основная идея, лежащая в основе определения весов предлагаемого метода, основана на сравнении генотипов пар объектов и сравнение соответствующих значений фенотипов этих объектов в паре. При этом мы используем следующее интуитивное предположение. Если генотипы двух объектов близки друг к другу, и соответствующие значения фенотипов этих двух объектов различны, то SNP-маркеры, которые соответствуют различным элементам рассматриваемых двух генотипов, могут быть значимыми. Действительно, если два объекта различаются малым числом элементов генотипов, то естественно ожидать, что их фенотипы близки или равны. Однако, если соответствующие фенотипы различны, то можно предположить, что это малое число различных элементов генотипов и определяет различие фенотипов. Так как значение фенотипа определяется не только генотипом, но и влиянием внешней среды, то нельзя делать такое заключение по одной паре, а необходимо анализировать все пары объектов. Таким образом, с учетом расстояний между генотипами и фенотипами для каждой пары объектов популяции можно определить некоторый совместный показатель, характеризующий влияние данной пары на определение ассоциации между SNP и количественным признаком.

Вторая идея, положенная в основу предлагаемого метода, заключается в рассмотрении совместного показателя влияния пары как степени предпочтения одного объекта по сравнению с другим с точки зрения увеличения значения фенотипа. Такое представление позволяет построить матрицу попарных сравнений объектов и далее определить веса объектов.

Для формализации приведенных выше рассуждений необходимо определить понятия близости фенотипов и генотипов для каждой пары объектов, характеризующихся векторами аллелей $\mathbf{x}_i$, $\mathbf{x}_j$ и расстояниями y_i, y_j. Обозначим расстояние между двумя векторами $\mathbf{x}_i$ и $\mathbf{x}_j$ в соответствии с некоторой метрикой как $\rho_{ij} = \rho(\mathbf{x}_i, \mathbf{x}_j)$. Учитывая тот факт, что элементы векторов аллелей являются номинальными, наиболее простой метрикой с вычислительной точки зрения является расстояние Хэмминга. Расстояние Хэмминга определяется как число позиций, в которых соответствующие символы двух векторов одинаковой длины различны. Расстояние между значениями фенотипа объектов обозначим d_{ij}, и оно определяется как $d_{ij} = y_i - y_j$.

Одним из вариантов совместного показателя, характеризующего пары объектов, является отношение d_{ij} к ρ_{ij}, обозначаемое r_{ij}, т.е. $r_{ij} = \left(d_{ij}\right)^q / \rho_{ij}$. Здесь разность d_{ij} для общности берется в степени q. Чем больше d_{ij} и меньше ρ_{ij}, тем больше показатель r_{ij}, что соответствует большей степени предпочтения *i*-го объекта перед *j*-ым объектом.

Если теперь построить матрицу $R = \{r_{ij}\}_{n \times n}$ размерности $n \times n$, то полученная матрица имеет свойства, позволяющие говорить о ее согласованности как матрицы парных сравнений. Заметим, что каждый элемент матрицы определяется через разность $y_i - y_j$, поэтому выполняется соотношение $r_{ij} = -r_{ji}$, которое определяет аддитивную шкалу матрицы парных сравнений. Будем говорить, что *i*-ый объект предпочтительнее *j*-го объекта, если $r_{ij} > 0$.

Матрица R является слабо-транзитивной, если из условий $r_{ij} > 0$ и $r_{jk} > 0$ следует $r_{ik} > 0$. Докажем, что матрица R, построенная при помощи предлагаемого алгоритма, является слабо-транзитивной. Без потери общности будем полагать, что $q = 1$. Если объекты упорядочены в порядке убывания, т.е. $y_1 \geq ... \geq y_n$, то можно записать следующее условие для разности фенотипов: $d_{ik} = d_{ij} + d_{jk}$, которое следует из $(y_i - y_j) + (y_j - y_k) = y_i - y_k$. Так как расстояние ρ_{ij} удовлетворяет неравенству треугольника $\rho_{ik} \leq \rho_{ij} + \rho_{jk}$, то:

$$r_{ik} = \frac{d_{ik}}{\rho_{ik}} \geq \frac{d_{ij} + d_{jk}}{\rho_{ij} + \rho_{jk}}.$$

Из условий $r_{ij} > 0$ и $r_{jk} > 0$ следует, что $d_{ij} > 0$ и $d_{jk} > 0$. Отсюда получаем:

$$r_{ik} \geq \frac{d_{ij} + d_{jk}}{\rho_{ij} + \rho_{jk}} > 0,$$

что и следовало доказать.

Условие слабой транзитивности является минимальным требованием согласованности парных сравнений. Оно позволяет нам говорить о вычислении весов объектов w_i, $i = 1,...,n$, которые можно рассматривать, как меры вклада каждого вектора $\mathbf{x}_i$ в ассоциацию между соответствующими SNP и значениями фенотипа. Согласно работе [4], веса объектов при условии аддитивной шкалы определяются как:

$$w_i = \frac{1}{n}\sum_{j=1}^{n} r_{ij}.$$

Больший вес *i*-го объекта означает, что большинство пар объектов с индексами (i, j) при фиксированном *i* и при $j = 1,...,n$ являются определяющими с точки зрения изменений фенотипа.

Другой вариант совместного показателя, характеризующего пары объектов, является:

$$r_{ij} = \exp\left(\left(d_{ij}\right)^q / \rho_{ij}\right).$$

Если $d_{ij} < 0$, то $r_{ij} \in (0,1]$, иначе $r_{ij} > 1$. Кроме того, выполняется соотношение $r_{ij} = 1/r_{ji}$. Также заметим, что $r_{ii} = 1$. Тогда можно говорить о мультипликативной шкале предпочтений. Докажем, что условие слабой транзитивности выполняется, т.е. из условий $r_{ij} > 1$ и $r_{jk} > 1$ следует $r_{ik} > 1$. Без потери общности полагаем, что $q = 1$. Используя снова неравенство треугольника для расстояний, запишем:

$$r_{ik} = \exp\left(\frac{d_{ik}}{\rho_{ik}}\right) \geq \exp\left(\frac{d_{ij} + d_{jk}}{\rho_{ij} + \rho_{jk}}\right).$$

Из условий $r_{ij} > 1$ и $r_{jk} > 1$ следует, что $d_{ij} > 0$ и $d_{jk} > 0$. Отсюда получаем:

$$r_{ik} \geq \exp\left(\frac{d_{ij} + d_{jk}}{\rho_{ij} + \rho_{jk}}\right) > \exp(0) > 1,$$

что и следовало доказать.

Для вычисления весов объектов w_i, $i = 1,...,n$ используем результаты работы [4], где веса объектов при условии мультипликативной шкалы определяются как:

$$w_i = \left(\prod_{j=1}^{n} r_{ij}\right)^{1/n}.$$

После нормировки веса объектов могут использоваться для решения задачи GWAS при применении большинства известных методов, рассмотренных кратко во вводной части статьи. В частности, одним из наиболее успешных является метод Лассо. Поэтому рассмотрим, как изменится этот метод с учетом полученных весов.

3. Метод Лассо с весами. В соответствии с методом Лассо решается задача построения линейной регрессионной модели:

$$y = X\mathbf{b}^{\mathrm{T}} + b_0 + \varepsilon.$$

Здесь ε — ошибка, распределение которой подчиняется нормальному закону с нулевым математическим ожиданием; X — вектор значений SNP-маркеров, для которого необходимо определить значение фенотипа y; $\mathbf{b}$ — вектор неизвестных параметров, каждый элемент которого определяет, как соответствующий SNP влияет на значения фенотипа. Определение вектора $\mathbf{b}$ является основной задачей GWAS. Параметр b_0 — свободный член. Если все переменные центрированы, то можно исключить свободный член из рассмотрения.

Следует отметить, что в общем случае задачу построения регрессионной и классификационной модели можно записать в виде следующей задачи оптимизации:

$$\mathbf{b}^0 = \arg\min_{\mathbf{b}} \left\{ l(Y, \mathbf{X}, \mathbf{b}) + \lambda \cdot Q(\mathbf{b}) \right\},$$

где $l(Y, \mathbf{X}, \mathbf{b})$ — функция потерь, $Q_{\boldsymbol{\lambda}}(\mathbf{b})$ — штрафное слагаемое,

ограничивающее множество возможных решений, $\mathbf{b}^0$ — оптимальные значения параметров $\mathbf{b}$, λ — неотрицательный параметр регуляризации или сглаживания, характеризующий степень влияния второго слагаемого в регрессионной модели.

В методе Лассо используется стандартная функция потерь $\|Y - \mathbf{X}\mathbf{b}\|^2$ и штрафное слагаемое вида:

$$Q(\mathbf{b}) = \|\mathbf{b}\|^1 = \sum_{j=1}^{m} |b_j| .$$

Если обозначить вектор весов объектов как $\mathbf{w} = (w_1, w_2, ..., w_n)$, то метод Лассо с учетом вектора весов можно записать в виде следующей задачи оптимизации:

$$\mathbf{b}^0 = \arg\min_{\mathbf{b}} \left\{ \|Y - \mathbf{X}\mathbf{b}\|^2 \cdot \mathbf{w}^{\mathrm{T}} + \lambda \cdot \|\mathbf{b}\|^1 \right\} .$$

4. Числовые эксперименты. Числовые эксперименты осуществлялись для двух популяций удвоенных гаплоидных (УГ) линий ячменя. Первое множество данных состоит из 93 УГ линий ячменя, описанных в работе [5]. Данные по фенотипам и генотипам можно найти на сайте (http://wheat.pw.usda.gov/ggpages/maps/OWB/). В экспериментах анализировались линии в соответствии с семью количественными признаками: длина колоса (SL) в см.; количество зерен (GN); количество цветков (FN); вес зерен в 100-граммовых порциях (HGW); высота растений (PH) в см.; число колосков (SN); время колошения (HD) в днях. Карта сцепления состоит из 1328 SNP-маркеров.

Второе множество данных состоит из 92 УГ линий ячменя, полученных от скрещивания Dicktoo x Morex и описанных в работе [8]. Данные по фенотипам и генотипам можно найти на сайте http://wheat.pw.usda.gov/ggpages/DxM/. В экспериментах анализировались линии в соответствии с двумя количественными признаками: время колошения с яровизацией и без яровизации с режимом светового периода 8-ч света/16-ч темноты. Карта сцепления состоит из 117 SNP-маркеров.

Пропущенные данные оценивались посредством эвристической процедуры, предложенной в работе [1].

Из каждого множества данных случайным образом выбираются два подмножества: множество обучающих n примеров и множество тестирующих данных в количестве n_{test} примеров для оценки качества

алгоритмов, которое оценивается при помощи средней квадратической ошибки регрессии (СКОР), которая определяется как:

$$E = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{f}(\mathbf{x}_i))^2 ,$$

где $\hat{f}$ — функция, оцененная предложенным методом, $\hat{f}(\mathbf{x}_i)$ — прогнозируемое значение фенотипа y_i для каждого $i \in \{1, \ldots, n_{test}\}$. Показатель ошибки СКОР вычисляется при помощи усреднения результатов многократно повторяющегося случайного выбора тестирующих данных. Чем меньше значение показателя, тем лучше соответствующий метод. Мы используем метод кросс-валидации с одним тестирующим элементом, т.е. $n_{test} = 1$. Это связано с тем, что число линий не велико по сравнению с числом SNP и каждое наблюдение важно.

Значения СКОР для первого множества данных при условии $q = 1$ и выборе аддитивной шкалы построения матрицы парных сравнений приведены в таблице 1. Первый столбец таблицы соответствует семи анализируемым признакам. Столбцы 2 и 3 содержат значения СКОР, полученные с использованием 20 наиболее значимых SNP. Столбцы 4 и 5 содержат значения СКОР, полученные с использованием всех SNP. Сокращения СМ и ПМ означают стандартный метод и предлагаемый метод с весами объектов соответственно. Можно заметить из таблицы 1, что предлагаемый метод обеспечивает лучшие результаты почти для всех количественных признаков. Кроме того, наибольший эффект достигается, когда используются только значимые SNP.

Таблица 1. СКОР при выборе аддитивной шкалы и $q = 1$

	Значимые SNP		Все SNP	
Признаки	СМ	ПМ	СМ	ПМ
SL	0.994	0.942	2.675	2.603
GN	85.914	85.634	140.237	140.242
FN	44.645	36.553	83.954	84.026
HGW	0.124	0.118	0.201	0.191
PH	82.422	72.054	159.735	155.271
SN	12.815	11.728	16.828	16.243
HD	55.537	48.419	90.236	91.116

Для исследования влияния значения параметра q на точность моделирования рассмотрен также случай $q = 2$, для которого результаты анализа приведены в таблице 2. Из таблицы видно, что точность

определения значимых SNP, в этом случаем повышается. Отсюда можно сделать вывод, что «правильный» выбор параметра q может привести к существенно лучшим результатам.

Таблица 2. СКОР при выборе аддитивной шкалы и $q = 2$

	Значимые SNP		Все SNP	
Признаки	СМ	ПМ	СМ	ПМ
SL	0.994	0.933	2.675	2.586
GN	85.914	85.001	140.237	138.954
FN	44.645	34.221	83.954	85.930
HGW	0.124	0.116	0.201	0.189
PH	82.422	70.936	159.735	152.579
SN	12.815	12.249	16.828	16.802
HD	55.537	47.632	90.236	91.130

Для исследования того, как влияет шкала построения матрицы парных сравнений, были проведены эксперименты с использованием мультипликативной шкалы, результаты которых представлены в таблице 3. Из сравнения всех приведенных таблиц нельзя однозначно сделать вывод о преимуществе той или иной шкалы. Для одних количественных признаков, например, для HD, выбор мультипликативной шкалы дает существенное снижение СКОР по сравнению с аддитивной шкалой. Однако для других признаков такого существенного улучшения не наблюдается. Выбор шкалы может рассматриваться в качестве еще одного «параметра» настройки модели на конкретные данные.

Таблица 3. СКОР при выборе мультипликативной шкалы и $q = 1$

	Значимые SNP		Все SNP	
Признаки	СМ	ПМ	СМ	ПМ
SL	0.994	1.008	2.675	2.661
GN	85.914	84.001	140.237	139.971
FN	44.645	35.696	83.954	84.220
HGW	0.112	0.107	0.201	0.201
PH	82.422	78.350	159.735	157.549
SN	12.815	12.023	16.828	16.793
HD	55.537	43.442	90.236	90.008

Рассмотрим второе множество данных. Таблица 4 содержит показатели ошибки для множества данных Dicktoo x Morex при анализе двух количественных признаков: время колошения с яровизацией и без яровизации. Из таблицы 4 следует, что предлагаемый метод показывает значительно лучшие результаты при выборе аддитивной шкалы и для первого признака. Однако это улучшение не столь очевидно для второго признака. В то же время выбор мультипликативной шкалы дает лучшие резуль-

таты, которые представлены в таблице 5, где для первого и второго признаков использовались параметры $q = 1$ и $q = 1/2$ соответственно.

Таблица 4. СКОР при выборе аддитивной шкалы и $q = 2$

	Значимые SNP		Все SNP	
Признаки	СМ	ПМ	СМ	ПМ
без яровизации	29.849	25.261	46.169	43.753
с яровизацией	25.996	23.136	38.241	39.480

Таблица 5. СКОР при выборе мультипликативной шкалы и $q = 1$ и $q = 1/2$

	Значимые SNP		Все SNP	
Признаки	СМ	ПМ	СМ	ПМ
без яровизации	29.849	22.974	46.169	40.331
с яровизацией	25.996	25.295	38.241	37.580

5. Заключение. В работе был предложен простой с вычислительной точки зрения метод назначения весов объектам обучающей выборки, который позволяет учесть целый ряд факторов. Казалось бы, почему не приписать веса объектам в соответствии со значениями фенотипов или в соответствии с некоторой областью вокруг значений фенотипов, как это было сделано в работе [1]. Действительно, такой подход может быть использован. Однако он не учитывает два важных фактора. Во-первых, значения фенотипа являются случайными благодаря внешним условиям. Во-вторых, каждый объект характеризуется генотипом, который может определять изменения фенотипа, а может и не определять. Следует отметить, что второй фактор является намного более важным. Для учета одновременно обоих факторов было предложено сравнивать пары объектов с учетом расстояний между фенотипами и генотипами. Оказывается, что при определенном сравнении пары образуют матрицы парных сравнений, свойства согласованности которых позволяют определить веса объектов. Так как вес получается в результате усреднения по подмножеству пар, то случайные факторы «сглаживаются», что позволяет учесть первый фактор.

Метод был применен к известному алгоритму Лассо, который повсеместно используется в GWAS. Однако это не значит, что предлагаемый метод не может работать с другими инструментами GWAS. Известные методы фильтрации, основанные на проверке статистических гипотез, методы упаковки, например, рекурсивное удаление признаков, с таким же успехом могут быть модифицированы с учетом полученных весов объектов. Эффективность применения метода в сочетании с другими подходами в GWAS является одним из направлений дальнейших исследований.

Заметим, что при получении весов объектов не было использовано никакой дополнительной «внешней» информации, т.е. веса полностью определялись с использованием только матрицы генотипов и вектора фенотипов. Возникает еще одни интересный вопрос, почему, например, метод Лассо не учитывает особенностей предлагаемого метода. Дело в том, что метод Лассо, а также многие другие методы предполагают независимость SNP, т.е. независимость строк матрицы генотипов, что в большинстве случаев является слишком оптимистичным предположением. Определение весов некоторым образом корректирует это предположение, так как в их вычислении используются столбцы матрицы генотипов.

Результаты числовых экспериментов с реальными данными продемонстрировали, что метод позволяет получить более точные характеристики по сравнению со стандартным методом Лассо. Однако эти результаты также показали, что требуется определенная настройка параметров для получения наибольшего эффекта.

Литература

1. *Уткин Л.В., Жук Ю.А., Коолен Ф.* Робастная модификация метода Лассо для полногеномного поиска ассоциаций с учетом целевых значений фенотипа // Научно-технический вестник информационных технологий, механики и оптики. 2016. Вып. 16. № 1. С. 150–160.
2. *Уткин Л.В., Уткина И.Л.* Быстрый алгоритм полногеномного поиска ассоциаций по схеме случай–контроль // Известия Санкт-Петербургской лесотехнической академии. 2015. Вып. 213. С. 263–273.
3. *Altidor W., Khoshgoftaar T.M., Van Hulse J., Napolitano A.* Ensemble feature ranking methods for data intensive computing applications // Handbook of Data Intensive Computing. Springer. New York. 2011. pp. 349–376.
4. *Barzilai J., Golany B.* Deriving weights from pairwise comparison matrices: The additive case // Operations Research Letters. 1990. vol. 9. pp. 407–410.
5. *Cistue L. et. al.* Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes // Theoretical and applied genetics. 2011. vol. 122(7). pp. 1399–1410.
6. *Goddard M.E., Wray N.R., Verbyla K., Visscher P.M.* Estimating effects and making predictions from genome-wide marker data // Statistical Science. 2009. vol. 24(4). pp. 517–529.
7. *Hayes B.* Overview of statistical methods for genome-wide association studies (GWAS) // Methods in Molecular Biology. 2013. vol. 1019. pp. 149–169.
8. *Hayes P. et. al.* The Dicktoo x Morex population // Plant Cold Hardiness. Springer US. 1997. pp. 77–87.
9. *Kohavi R., John G.H.* Wrappers for feature subset selection // Artificial Intelligence. 1997. vol. 97(1–2). pp. 273–324.
10. *Guyon I., Weston J., Barnhill S., Vapnik V.* Gene selection for cancer classification using support vector machines // Machine Learning. 2002. vol. 46. pp. 389–422.
11. *Moore J.H., Asselbergs F.W., Williams S.M.* Bioinformatics challenges for genome-wide association studies // Bioinformatics. 2010. vol. 26(4). pp. 445–455.

12. *Saaty T.L.* Multicriteria Decision Making: The Analytic Hierarchy Process. New York: McGraw Hill. 1980.
13. *Szymczak S. et al.* Machine learning in genome-wide association studies // Genetic Epidemiology. 2009. vol. 33. pp. 51–57.
14. *Tibshirani R.* Regression shrinkage and selection via the Lasso // Journal of the Royal Statistical Society. Series B (Methodological). 1996. vol. 58(1). pp. 267–288.
15. *Zhang X., Huang S., Zhang Z., Wang W.* Chapter 10: Mining Genome-Wide Genetic Markers // PLoS Computational Biology. 2012. vol. 8(12). pp. e1002828.
16. *Zou H., Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. vol. 67(2). pp. 301–320.

References

1. Utkin L.V., Zhuk Y.A., Coolen F. [Robust modification of the Lasso method for genome-wide association study in view of target phenotype values]. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2016. vol. 16. no. 1. pp. 150–160. (In Russ.).
2. Utkin L.V., Utkina I.L. [A fast algorithm of a case-control genome-wide association study]. *Izvestija Sankt-Peterburgskoj lesotehnicheskoj akademii – Izvestia Sankt-Peterburgskoj Lesotehniceskoj Akademii.* 2015. vol. 213. pp. 263–273. (In Russ.).
3. Altidor W., Khoshgoftaar T.M., Van Hulse J., Napolitano A. Ensemble feature ranking methods for data intensive computing applications. Handbook of Data Intensive Computing, Springer. New York, 2011. pp. 349–376.
4. Barzilai J., Golany B. Deriving weights from pairwise comparison matrices: The additive case. *Operations Research Letters*. 1990. vol. 9. pp. 407–410.
5. Cistue L. et. al. Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics*. 2011. vol. 122(7). pp. 1399–1410.
6. Goddard M.E., Wray N.R., Verbyla K., Visscher P.M. Estimating effects and making predictions from genome-wide marker data. *Statistical Science.* 2009. vol. 24(4), pp. 517–529.
7. Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*. 2013. vol. 1019. pp. 149–169.
8. Hayes P. et. al. The Dicktoo x Morex population. Plant Cold Hardiness. Springer US. 1997. pp. 77–87.
9. Kohavi R., John G.H. Wrappers for feature subset selection. *Artificial Intelligence*. 1997. vol. 97(1–2). pp. 273–324.
10. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning.* 2002. vol. 46. pp. 389–422.
11. Moore J.H., Asselbergs F.W., Williams S.M. Bioinformatics chal-lenges for genome-wide association studies. *Bioinformatics.* 2010. vol. 26(4). pp. 445–455.
12. Saaty T.L. Multicriteria Decision Making: The Analytic Hierarchy Process. New York: McGraw Hill. 1980.
13. Szymczak S. et. al. Machine learning in genome-wide association studies. *Genetic Epidemiology.* 2009. vol. 33. pp. 51–57.
14. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996. vol. 58(1). pp. 267–288.
15. Zhang X., Huang S., Zhang Z., Wang W. Chapter 10: Mining Genome-Wide Genetic Markers. *PLoS Computational Biology.* 2012. vol. 8(12). pp. e1002828.
16. Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005. vol. 67(2). pp. 301–320.

Уткин Лев Владимирович — д-р техн. наук, профессор, профессор кафедры телематики (при ЦНИИ РТК), Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: интеллектуальный анализ данных, представление неопределенности, принятие решений при неполной информации, теория надежности, биоинформатика. Число научных публикаций — 328. lev.utkin@gmail.com, http://levvu.narod.ru; Политехническая, 29, Санкт-Петербург, 195251; р.т.: +7(812)5526521.

Utkin Lev Vladimirovich — Ph.D., Dr. Sci., professor, professor of telematics (under Russian state scientific center for robotics and technical cybernetics (RTC)) of Institute of applied mathematics and mechanics, Peter the Great Saint-Petersburg Polytechnic University. Research interests: machine learning, uncertainty representation, decision making under incomplete information, reliability theory, bioinformatics. The number of publications — 328. lev.utkin@gmail.com, http://levvu.narod.ru; 29, Polytechnicheskaya, St. Petersburg, 195251; office phone: +7(812)552-6521.

Жук Юлия Александровна — к-т пед. наук, доцент кафедры компьютерных образовательных технологий, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО). Область научных интересов: когнитивные образовательные технологии, интеллектуальный анализ данных, мультимедийные технологии. Число научных публикаций — 73. zhuk_yua@mail.ru, http://dce.ifmo.ru; Кронверкский пр., 49, Санкт-Петербург, 197101; р.т.: +7(812)233-1961.

Zhuk Yulia Alexandrovna — Ph.D., associate professor of computer educational technologies department, ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics). Research interests: cognitive educational technologies, data mining, multimedia technologies. The number of publications — 73. zhuk_yua@mail.ru, http://dce.ifmo.ru; 49, Kronverksky pr., St. Petersburg, 197101; office phone: +7(812)233-1961.

РЕФЕРАТ

Уткин Л.В., Жук Ю.А. **Полногеномный поиск ассоциаций с использованием матрицы парных сравнений.**

Основная цель полногеномного поиска ассоциаций заключается в определении, какие ДНК-маркеры, называемые однонуклеотидными полиморфизмами, оказывают наибольшее влияние на значения фенотипа. С точки зрения теории машинного обучения имеем задачу отбора признаков.

Одним из подходов повышения качества моделей является предварительная обработка имеющихся данных с целью выявления каких-либо закономерностей в данных, которые могли бы использоваться в качестве дополнительной информации при построении стандартных моделей или алгоритмов. Предлагается простой метод определения значимости объектов популяции или обучающей выборки с точки зрения представительности того или иного объекта в установлении ассоциации между ДНК-маркерами и количественными признаками. Значимость представляется в виде весов объектов и может в дальнейшем использоваться при «взвешивании» элементов обучающей выборки в любых моделях полногеномного поиска ассоциаций. Метод основан на двух идеях. Первая идея — сравнение пар объектов популяции с точки зрения расстояния между ними по фенотипу и генотипу. Вторая идея заключается в построении матрицы парных сравнений объектов специальным образом и вычислении весов объектов, например, по аналогии с известным методом анализа иерархий.

Показано, как найденные веса могут быть использованы в методе Лассо. Результаты числовых экспериментов с реальными данными (удвоенные гаплоидные линии ячменя) продемонстрировали, что метод позволяет получить более точные характеристики по сравнению со стандартным методом Лассо.

SUMMARY

Utkin L.V., Zhuk Y.A. **A Genome-Wide Association Study using Pairwise Comparison Matrices.**

The main aim of a genome-wide association study is to determine which DNA-markers, called the single nucleotide polymorphisms, have the greatest influence on the phenotype values. This is a feature selection problem in the machine learning framework.

One of the approaches to enhance the model quality is preprocessing of available data in order to reveal some dependencies between data, which could be used as additional information for constructing the standard models and algorithms. We propose a simple method for determining the importance of individuals of a population or the training set elements according to the individual representativeness in looking for the association between DNA-markers and quantitative traits. The importance is represented in the form of individual weights and can be used for weighing elements of the training set in arbitrary genome-wide association study models. The method is based on two ideas. The first one is to compare pairs of the individuals of population in terms of distances between phenotypes and between genotypes. The second idea is to construct a pairwise comparison matrix for individuals in a special way and to compute the weights of individuals, for example, similarly to the well-known analytic hierarchy process.

It is shown how the computed weights can be applied to the Lasso method. The results of numerical experiments with real datasets (double haploid lines of barley) illustrated that the proposed method allows us to obtain more accurate measurements in comparison with the standard Lasso method.