УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2016 Том 16 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2016 Vol. 16 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 57.087

РОБАСТНАЯ МОДИФИКАЦИЯ МЕТОДА ЛАССО ДЛЯ ПОЛНОГЕНОМНОГО ПОИСКА АССОЦИАЦИЙ С УЧЕТОМ ЦЕЛЕВЫХ ЗНАЧЕНИЙ ФЕНОТИПА

Л.В. Уткин[a], Ю.А. Жук[b, c], Ф. Коолен[d]

[a] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, 195251, Российская Федерация
[b] Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург, 194021, Российская Федерация
[c] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[d] Университет Дарем, Дарем, DH1 3LE, Великобритания
Адрес для переписки: zhuk_yua@mail.ru

Информация о статье
Поступила в редакцию 13.10.15, принята к печати 02.12.15
doi:10.17586/2226-1494-2016-16-1-150-160
Язык статьи – русский

Ссылка для цитирования: 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. Т. 16. № 1. С. 150–160.

Аннотация

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. Со статистической точки зрения рассматривается модификация задачи линейной регрессии. Предложено формализовать дополнительную информацию 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. Полученное пересечение является выпуклым и полностью определяется его крайними точками, что позволяет свести модифицированный, с точки зрения использования множеств весов наблюдений, метод Лассо к конечному множеству стандартных реализаций Лассо. Результаты числовых экспериментов показали, что модификация позволяет получить более точные характеристики по сравнению со стандартным методом Лассо при малом объеме обучающей выборки.

Ключевые слова

полногеномный поиск ассоциаций, фенотип, регрессия, Лассо, модель засорения, парные сравнения, выпуклое множество

Благодарности

Работа выполнена при поддержке РФФИ, проект № 15-01-01414 и Минобрнауки РФ, проект № 2014/181-2220.

ROBUST MODIFICATION OF THE LASSO METHOD FOR GENOME-WIDE ASSOCIATION STUDY IN VIEW OF TARGET PHENOTYPE VALUES

L.V. Utkin[a], Yu. A. Zhuk[b, c], F. Coolen[d]

[a] Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, 195251, Russian Federation
[b] Saint Petersburg State Forest Technical University, Saint Petersburg, 194021, Russian Federation
[c] ITMO University, Saint Petersburg, 197101, Russian Federation
[d] Durham University, Durham, DH1 3LE, UK
Corresponding author: zhuk_yua@mail.ru

Article info
Received 13.10.15, accepted 02.12.15
doi:10.17586/2226-1494-2016-16-1-150-160
Article in Russian

For citation: Utkin L.V., Zhuk Yu.A., Coolen F. Robust modification of the Lasso method for genome-wide association study in view of target phenotype values. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 1, pp. 150–160.

Abstract

A modification of the Lasso method used for genome-wide association study by examples of double haploid lines of barley is proposed for taking into account the additional information about target values of the phenotype which is defined by some feature of plants. From a statistical point of view, a linear regression problem is studied. It is proposed to formalize the additional information about features of plants as intersection of two sets of weights assigned to the training set elements. The

first set of weights is produced by means of the interval contamination model. The second set is formed by the pair-wise comparisons of phenotype values. The obtained intersection is convex and is totally defined by its extreme points. This feature allows reducing the Lasso method with sets of weights to a finite set of standard Lasso methods. Results of numerical experiments have showed that the modification provides the better accuracy measures in comparison with the standard Lasso when the training set is small.

Keywords
genome-wide association study, phenotype, regression, Lasso, contamination model, pair-wise comparisons, convex set

Acknowledgements
The study was partially supported by RFBR, research project No. 15-01-01414-a and the Ministry of Education and Science of the Russian Federation, project No. 2014/181-2220.

Введение

Цель полногеномного поиска ассоциаций (a genome-wide association study – GWAS) заключается в поиске связи или ассоциации между значениями фенотипа и генотипом. Полногеномный поиск ассоциаций можно рассматривать как один из методов решения известной задачи отбора признаков, где признаками являются однонуклеотидные полиморфизмы (singular nucleotide polymorphisms – SNP) или молекулярные маркеры. Понимание того, как генетические вариации влияют на изменения фенотипа, имеет огромное значение для прогнозирования полезных свойств сельскохозяйственных культур. Как показано в работе [1], имеет место целый ряд сложностей при решении задачи отбора SNP-маркеров. Во-первых, число SNP-маркеров p обычно в десятки или сотни раз превышает число элементов n в обучающей выборке или число анализируемых потомков. Другими словами, размерность задачи значительно превышает количество точек в пространстве признаков. Во-вторых, 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 неизвестно.

Большое число статистических моделей и методов для решения задачи отбора SNP-маркеров было предложено в последнее время. Часть методов можно отнести к методам фильтрации [2, 3], которые используют статистические свойства SNP-маркеров для отбрасывания наименее информативных маркеров. К ним относятся известные методы, основанные на использовании t-статистики, критерия Фишера (F-статистики), χ^2-статистики и многие другие. Фактически методы фильтрации исследуют, насколько значения генотипа каждого SNP-маркера и соответствующие им значения фенотипа образуют различные распределения вероятностей, т.е. насколько различные значения генотипа разделяют значения фенотипа.

Другая часть методов называется методами упаковки [4]. Считается, что они обеспечивают более точное решение по сравнению с методами фильтрации, но одновременно являются более затратными с вычислительной точки зрения [4]. Одним из наиболее известных таких методов является метод, предложенный в работе [5] и называемый рекурсивным удалением признаков. Этот метод был успешно применен для решения задачи отбора генов при классификации онкологических заболеваний.

Методы фильтрации и их модификации, упаковочные методы могут быть достаточно эффективным инструментом для решения задачи отбора SNP-маркеров. В то же время при решении задачи отбора SNP-маркеров, особенно при наличии изменяющихся значений фенотипа, используют регрессионные алгоритмы. Одной из первых и наиболее известных работ, посвященных применению регрессионных моделей для решения задачи отбора SNP-маркеров, является работа [6]. Методы для построения соответствующих регрессионных моделей могут быть отнесены к вложенным методам [7]. Они реализуют отбор SNP-маркеров в процессе обучения моделей и охватывают большое число известных подходов, включая метод Лассо (Least Absolute Shrinkage и Selection Operator – LASSO) [8, 9], который, очевидно, является наиболее популярным и эффективным методом в задачах отбора SNP-маркеров. Основное преимущество использования Лассо заключается в том, что он осуществляет отбор переменных и классификацию (или построение регрессии) одновременно.

Множество подходов, использующих метод Лассо и его модификации, было разработано для решения задачи отбора SNP-маркеров в рамках полногеномного поиска ассоциаций (см., например, работы [10]). В работе [11] приведен исчерпывающий обзор статистических методов решения задач полногеномного поиска ассоциаций.

В настоящей работе мы модифицируем GWAS и исследуем предлагаемую модификацию на примере анализа удвоенных гаплоидных (УГ) линий ячменя. В соответствии с УГ методом, только два типа генотипов имеют место для каждой пары аллелей. Со статистической точки зрения мы решаем задачу линейной регрессии. Решение основано на использовании метода Лассо. Мы предлагаем формализовать дополнительную информацию о свойствах культур в виде оригинального пересечения двух множеств весов элементов обучающей выборки. Первое множество образовано при помощи интервальной модели ε-засорения [12]. Второе множество весов образуется множеством парных сравнений значений фенотипа или элементов обучающей выборки. В результате этого множества компенсируют друг друга. Мы

показываем, что полученное пересечение является выпуклым и полностью определяется его крайними точками, что позволяет свести модифицированный, с точки зрения использования множеств весов наблюдений, метод Лассо к конечному множеству стандартных реализаций Лассо. Крайние точки пересечения получены в работе в явном виде, что делает реализацию метода достаточно простой задачей.

Метод Лассо

Мы анализируем n УГ линий ячменя или популяцию n потомков. С точки зрения регрессионных моделей SNP-маркеры могут рассматриваться как независимые входные переменные, т.е. $\mathbf{X}_j = (x_{1j},...,x_{nj})^{\mathrm{T}}$ – вектор значений j-го SNP-маркера, $j = 1,...,p$. Здесь x_{ij} – двоичная переменная, т.е. $x_{ij} \in \{0,1\}$. Числовой признак или множество значений фенотипа $y_i \in \mathbf{R}$, $i = 1,...,n$, можно рассматривать как выходной вектор $\mathbf{Y} = (y_1,...,y_n)^{\mathrm{T}}$. Мы также обозначим матрицу генотипов для n линий $\mathbf{X} = [\mathbf{X}_1,...,\mathbf{X}_p]$, вектор аллелей, соответствующих i-ой линии $\mathbf{x}_i^{\mathrm{T}} = (x_{i1},...,x_{ip})$, $i = 1,...,n$.

Наша цель – построить стандартную линейную регрессионную модель

$$y = \sum_{i=1}^{p} \beta_i \mathbf{X} + \beta_0 + \varepsilon = \mathbf{X}\boldsymbol{\beta}^{\mathrm{T}} + \beta_0 + \varepsilon.$$

Здесь ε – центрированный шум, т.е. случайная величина, имеющая нормальное распределение с математическим ожиданием, равным 0; $\mathbf{X}_0$ – вектор значений SNP-маркеров нового потомка, для которого необходимо определить значение фенотипа y; $\boldsymbol{\beta} = (\beta_1,...,\beta_p)$ – вектор неизвестных параметров, каждый элемент которого β_i определяет, как SNP-маркер с номером *i* влияет на значения вектора фенотипов $\mathbf{Y}$. Чем больше значение β_i, тем более значимым является *i*-й SNP-маркер. Таким образом, определение вектора $\boldsymbol{\beta}$ является основной задачей GWAS. Параметр β_0 – свободный член. Если все переменные центрированы, то можно исключить свободный член β_0 из рассмотрения [13].

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

$$\boldsymbol{\beta}^0 = \arg\min_{\beta \in \mathbf{R}^p} \left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}) + Q_\lambda(\boldsymbol{\beta}) \right\}, \quad (1)$$

где $l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta})$ – функция потерь; $Q_\lambda(\boldsymbol{\beta})$ – штрафное слагаемое; $\boldsymbol{\beta}^0$ – оптимальные значения параметров $\boldsymbol{\beta}$.

Наиболее популярной и эффективной регрессионной моделью, осуществляющей обработку данных при условии $p > n$, является метод Лассо [8]. Модель оценивает параметры $\boldsymbol{\beta}$ на основе минимизации следующей функции потерь $l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta})$:

$$\boldsymbol{\beta}^0 = \arg\min_{\beta \in \mathbf{R}^p} l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}) = \arg\min_{\beta \in \mathbf{R}^p} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|^2.$$

В методе Лассо ограничения на значения вектора $\boldsymbol{\beta}$ имеют вид

$$\left\| \boldsymbol{\beta} \right\|^1 = \sum_{j=1}^{p} \left| \beta_j \right| \le s, \; s \ge 0.$$

Здесь s – величина ограничения, выбираемая в зависимости от конкретной задачи или настраиваемая в результате экспериментов. Двойственная задача записывается как

$$\boldsymbol{\beta}^0 = \arg\min_{\beta \in \mathbf{R}^p} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|^1, \quad (2)$$

т.е. $Q_\lambda(\boldsymbol{\beta}) = \lambda \left\| \boldsymbol{\beta} \right\|^1$; λ – неотрицательный параметр регуляризации или сглаживания, характеризующий степень влияния второго слагаемого в регрессионной модели.

Мотивация для разработки нового метода

Существующие методы полногеномного поиска ассоциации не используют в полной мере дополнительную информацию, которая обычно имеет место. Для многих приложений цель полногеномного поиска ассоциации заключается не в построении наилучшей в некотором смысле регрессионной модели по имеющимся данным, а в отборе SNP-маркеров, которые влияют на наименьшие (наибольшие) значения фенотипа, исходя из практических соображений. Например, при анализе влияния SNP-маркеров на время колошения ячменя нас прежде всего интересуют минимальные значения данного признака, т.е. наша цель заключается в определении SNP-маркеров, влияющих на уменьшение времени колошения. Аналогичные зависимости можно определить для большинства признаков, определяющих полезные

свойства растений. Ниже для определенности будем рассматривать признаки, требующие минимизации их значений.

На первый взгляд, использование представленной дополнительной информации является тривиальной задачей и может быть осуществлено различными способами. Например, первый наиболее простой способ – оставить для построения регрессионной модели только линии с минимальными значениями фенотипа. Однако количество линий обычно настолько мало по сравнению с количеством SNP-маркеров, что этот подход приведет к совершенно неудовлетворительным результатам. Более того, оставшуюся выборку необходимо разделять на две части: обучающую и контрольную. Другой очевидный способ – приписать каждой линии некоторый вес, например, используя правило Фишберна, в соответствии с упорядоченными значениями фенотипа. Однако необходимо отметить, что значения фенотипа являются случайными в результате воздействия внешних факторов. В связи с этим, если одно значение фенотипа меньше другого, то это не означает, что данное соотношение имеет место благодаря соответствующим значениям генотипа. Исходя из этого, использование такого подхода также может привести к некорректным результатам. На самом деле первый способ есть частный случай второго способа, когда веса равны либо 0 (для отбрасываемых наблюдений) или 1 для наблюдений с минимальными значениями фенотипа.

Метод Лассо и множества весов наблюдений

Вернемся к рассмотрению задачи оптимизации (1) для регрессионной модели. Данная задача была записана при условии, что все элементы обучающей выборки равнозначны и имеют одинаковый вес. Для обобщения модели можно использовать веса $\mathbf{w} = (w_1, ..., w_n)$ наблюдений, на которые умножается функция потерь. Однако точные веса обычно неизвестны, или их назначение является бесполезным. В то же время вместо точных весов можно определить некоторое выпуклое множество весов $\mathbf{W}$, которое ослабляет жесткие условия, связанные с назначением точных весов.

Запишем задачу оптимизации при условии наличия множества весов. При этом необходимо, прежде всего, определить стратегию принятия решений, т.е. необходимо определить правило оптимального выбора весов из множества $\mathbf{W}$. Наиболее распространенной стратегией является минимаксная или пессимистическая стратегия, в соответствии с которой выбирается наихудший в некотором смысле вектор весов. В соответствии с этой стратегией выбирается такой вектор весов из множества $\mathbf{W}$, который обеспечивает максимум функции потерь для каждого фиксированного вектора параметров $\boldsymbol{\beta}$. Минимаксная стратегия интерпретируется как страховка против наихудшего случая. Тогда задача оптимизации может быть переписана следующим образом:

$$\boldsymbol{\beta}^0 = \arg\min_{\beta \in \mathbf{R}^p} \max_{\mathbf{w} \in \mathbf{W}} \left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}) \cdot \mathbf{w}^{\mathrm{T}} + Q_\lambda(\boldsymbol{\beta}) \right\}. \quad (3)$$

Так как оптимальный вектор весов зависит от параметров $\boldsymbol{\beta}$, то задачу нельзя решить непосредственно. Однако ее можно свести к конечному числу простых задач оптимизации, используя свойства множества $\mathbf{W}$ и задач линейного программирования.

Зафиксируем вектор параметров $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Тогда мы получаем задачу линейного программирования с переменными $\mathbf{w}$:

$$\max_{\mathbf{w} \in \mathbf{W}} \left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \cdot \mathbf{w}^{\mathrm{T}} + Q_\lambda(\boldsymbol{\beta}^*) \right\}.$$

Решение задачи оптимизации можно искать среди крайних точек множества возможных решений. В данном случае таким множеством является $\mathbf{W}$. Обозначим его крайние точки $\mathbf{E}(\mathbf{W})$. Тогда можно записать задачу оптимизации как

$$\max_{\mathbf{w}^{(k)} \in \mathbf{E}(\mathbf{W})} \left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \cdot \mathbf{w}^{(k)\mathrm{T}} + Q_\lambda(\boldsymbol{\beta}^*) \right\},$$

где $\mathbf{w}^{(k)}$ – k-ая крайняя точка множества $\mathbf{W}$.

Если число крайних точек равно r, то получаем r целевых функций или задач безусловной оптимизации. Таким образом, задача (3) может быть переписана в виде r стандартных задач оптимизации

$$\boldsymbol{\beta}^0 = \arg\max_{k=1,...,r} \min_{\beta \in \mathbf{R}^p} \left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}) \cdot \mathbf{w}^{(k)\mathrm{T}} + Q_\lambda(\boldsymbol{\beta}) \right\}. \quad (4)$$

Подставляя в (4) крайние точки $\mathbf{w}^{(k)}$ множества $\mathbf{W}$, получаем r задач (1). Оптимальное решение достигается при наибольшем значении целевой функции $\left\{ l(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}) \cdot \mathbf{w}^{(k)\mathrm{T}} + Q_\lambda(\boldsymbol{\beta}) \right\}$. Таким образом, зная крайние точки множества $\mathbf{W}$ и возвращаясь к методу Лассо (см. задачу (2)), задачу оптимизации (3) можно свести к r стандартным задачам Лассо.

Множества весов наблюдений и их крайние точки

Идея назначения весов наблюдениям может быть полезна и позволит повысить точность отбора SNP-маркеров, если ослабить некоторым образом их жесткое назначение. Это можно сделать, рассмотрев два множества весов.

Первое множество образовано при помощи интервальной модели ε-засорения [12]. Использование этой модели ослабляет условие назначения одинаковых весов, которое неявно имеет место в традиционном использовании метода Лассо, когда используются все наблюдения и они равнозначны.

Второе множество весов образуется множеством парных сравнений. Это множество позволяет ослабить жесткое упорядочивание и назначение точных весов, например, в соответствии с правилом Фишберна. Мы не назначаем точные веса для ранжирования элементов обучающей выборки. В то время как использование первого множества является достаточно общим методом для создания робастных статистических моделей, устойчивых к шумам, имеющим место особенно при ограниченности размера обучающей выборки, применение второго множества позволяет моделировать информацию о различной значимости элементов обучающей выборки. Здесь важно отметить, что мы не назначаем жестко веса, сравнивая значения фенотипа. Мы предлагаем использовать множество весов, согласованных с имеющейся информацией о цели полногеномного поиска ассоциации. Это существенное отличие от известных подходов.

Каждое множество весов в отдельности имеет свои положительные стороны. Однако эти множества, особенно второе, могут быть слишком большими, что либо делает их использование бесполезным, либо может привести к слишком пессимистическим решениям. В связи с этим мы используем оба множества, а точнее, их пересечение. Оба множества компенсируют друг друга, и их пересечение существенно сокращает их размер. Более того, оба множества являются выпуклыми, а следовательно, и их пересечение является выпуклым множеством, которое полностью определяется своими крайними точками. Следовательно, можно использовать эти крайние точки для решения задач оптимизации (4).

Рассмотрим более детально два множества весов.

Интервальная модель засорения. Первое множество весов образуется при помощи интервальной модели ε-засорения, которая была предложена Уолли [12] в качестве обобщения известной робастной модели ε-засорения [14]. Пусть $\mathbf{q} = (q_1, ..., q_n)$ – оценка некоторого распределения весов. Интервальная модель ε-засорения образует вокруг q такое множество весов $\mathbf{P}(\varepsilon, \mathbf{q})$, что $w_i = (1-\varepsilon)q_i + \varepsilon h_i$ для каждого фиксированного $\varepsilon \in (0,1)$ и q_i, где h_i – произвольные веса, на которые накладывается единственное ограничение $h_1 + ... + h_n = 1$. Другими словами, $\mathbf{h} = (h_1, ..., h_n)$ является произвольным распределением весов из единичного симплекса, обозначенного $S(1,n)$. Параметр засорения ε отражает меру неопределенности в оценке q. Таким образом, множество $\mathbf{P}(\varepsilon, \mathbf{q})$ является подмножеством единичного симплекса $S(1,n)$ и представляет собой симплекс меньшего размера. Только при $\varepsilon = 1$ множество $\mathbf{P}(\varepsilon, \mathbf{q})$ совпадает с единичным симплексом. Обозначим множество $\mathbf{P}(\varepsilon, \mathbf{q})$ при $\mathbf{q} = (1/n, ..., 1/n)$ как $\mathbf{P}(\varepsilon)$. В этом случае множество $\mathbf{P}(\varepsilon)$ образовано $n+1$ плоскостями или ограничениями

$$w_i \geq (1-\varepsilon)n^{-1},\ i = 1, ..., n,\ w_1 + ... + w_n = 1. \quad (5)$$

Множество, образованное сравнительной информацией. Рассмотрим следующую информацию в виде парных сравнений элементов обучающей выборки:

$$w_1 \leq w_2 \leq ... \leq w_n \quad (6)$$

при условии $w_1 + ... + w_n = 1$.

Множество $\mathbf{M}$ весов $\mathbf{w} = (w_1, ..., w_n)$, которые образованы $n-1$ неравенствами вида $w_i - w_{i-1} \geq 0$, а также n неравенствами $w_i \geq 0$ и одним равенством $w_1 + ... + w_n = 1$, является выпуклым, так как все неравенства и равенство являются линейными. Следовательно, множество $\mathbf{M}$ полностью определяется крайними точками следующего вида:

w_1	w_2	...	w_3	w_4	w_n
0	0	...	0	0	1
0	0	...	0	1/2	1/2
0	0	...	1/3	1/3	1/3
...	...	...	...	...	...
1/n	1/n	...	1/n	1/n	1/n

.

Интересно отметить, что крайние точки соответствуют ситуациям, когда только часть элементов обучающей выборки используется для построения регрессионной модели. Более того, два крайних случая имеют место. Первый случай заключается в том, что распределение весов имеет только один ненулевой элемент. Это означает, что модель строится на основе одного элемента с минимальным значением

фенотипа. Второй случай совпадает с равномерным распределением, которое неявно присутствует в стандартном подходе без учета дополнительной информации о свойствах растений.

Пересечение двух множеств. Перед тем как рассмотреть общий случай пересечения множеств $\mathbf{P}(\varepsilon)$ и $\mathbf{M}$, проанализируем случай $n=3$, используя стандартное представление весов при помощи единичного симплекса. Рис. 1 иллюстрирует симплекс, каждая точка которого – это вектор весов (w_1, w_2, w_3). Множество $\mathbf{M}$ соответствует области, ограниченной треугольником ABC. Множество $\mathbf{P}(\varepsilon)$ соответствует области, ограниченной малым симплексом. Их пересечение ограничено треугольником BED (заштрихованная область). Можно заметить, что полученное множество весов небольшое. Кроме того, веса смещены в направлении последней (третьей) вершины симплекса. В то же время малый симплекс, соответствующий $\mathbf{P}(\varepsilon)$, не позволяет назначать слишком большой вес как третьей вершине, так и другим вершинам. Другими словами, $\mathbf{P}(\varepsilon)$ является ограничителем весов, полученных парными сравнениями.

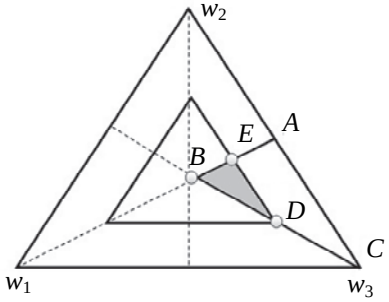

Рис. 1. Симплекс

Утверждение 1. Множество крайних точек пересечения $\mathbf{P}(\varepsilon)\cap\mathbf{M}$ состоит из n элементов вида

$$w_1 = w_2 = ... = w_{i-1} = \frac{1}{n} - \frac{\varepsilon}{n},\ w_i = w_{i+1} = ... = w_n = \frac{1}{n} + \frac{\varepsilon}{n}\cdot\frac{i-1}{n-i+1},\ i = 1,...,n.$$

Доказательство. Обозначим P – неравенства (5) и M – неравенства (6). Неравенства $w_i \geq 0$ не входят в P, так как все точки с $w_i = 0$, по крайней мере, для одного i не принадлежат пересечению $\mathbf{P}(\varepsilon)\cap\mathbf{M}$. Рассмотрим систему $2n-1$ неравенств из P и M. Известно, что каждая крайняя точка определяется $n-1$ равенствами из $P\cap M$. Рассмотрим следующие случаи.

Случай 1. Все $n-1$ неравенств принадлежат P. Очевидно, что $w_1 = ... = w_n = 1/n$.

Случай 2. $n-2$ неравенств принадлежат P, и 1 ограничение принадлежит M. Это значит, что существуют одно строгое неравенство $w_{i-1} < w_i$ из P и одно равенство $w_k = n^{-1} - \varepsilon n^{-1}$ из M. Рассмотрим два частных случая. Первый случай – $k \geq i$. Тогда

$$w_{i-1} < w_i = w_k = n^{-1} - \varepsilon n^{-1}.$$

Однако $w_i = w_{i+1} = ... = w_n$. Отсюда следует, что $w_1 + ... + w_n = 1 - \varepsilon < 1$. Получили противоречие. Поэтому этот случай не дает крайних точек. Второй случай – $k < i$. Тогда

$$w_1 = w_2 = ... = w_{i-1} = n^{-1} - \varepsilon n^{-1}.$$

Отсюда

$$w_i + ... + w_n = 1 - \left(n^{-1} - \varepsilon n^{-1}\right) i.$$

В итоге получаем

$$w_i = \frac{1 - \left(n^{-1} - \varepsilon n^{-1}\right) i}{n-i+1} = \frac{1}{n} + \frac{\varepsilon}{n}\cdot\frac{i-1}{n-i+1}.$$

Получена крайняя точка для фиксированного i.

Случай 3. $n-3$ неравенств принадлежат P, и 2 ограничения принадлежат M. Это означает, что существуют два неравенства $w_{i-1} < w_i$ и $w_{j-1} < w_j$, $i < j$, из P и два равенства $w_k = n^{-1} - \varepsilon n^{-1}$ и $w_l = n^{-1} - \varepsilon n^{-1}$, $k < l$. Случай $k \geq i$ и $l \geq i$ не рассматривается здесь (см. аналогичную ситуацию выше). Предположим, что $k < i$ и $l < i$. Тогда $w_1 = w_2 = ... = w_{i-1} = n^{-1} - \varepsilon n^{-1}$.

Предположим, что $w_s = a$, $s = i,..., j-1$, и $w_s = b$, $s = j,...,n$, т.е. $w_i = ... = w_{j-1} = a$, $w_j = ... = w_n = b$. Здесь благодаря неравенствам $w_{i-1} < w_i$ и $w_{j-1} < w_j$ запишем $n^{-1} - \varepsilon n^{-1} < a < b$.

Числа a и b удовлетворяют следующим очевидным условиям:

$$\left(n^{-1}-\varepsilon n^{-1}\right)(i-1)+a(j-i)+b(n-j+1)=1.$$

Существует бесконечное множество значений a и b, удовлетворяющих условиям выше. Это означает, что мы получили ребро соответствующего многогранника. То же самое получается для других случаев, когда мы берем $n-r$ равенств из P, и $r-1$ ограничений из M. Следовательно, случай 2 полностью определяет все крайние точки, что требовалось доказать. ■

Рассмотрим частные случаи множества $\mathbf{P}(\varepsilon)\cap\mathbf{M}$. Если $\varepsilon=0$, то множество $\mathbf{P}(\varepsilon)\cap\mathbf{M}$ сужается до точки $(1/n,...,1/n)$. Этот случай соответствует стандартному методу Лассо. Если $\varepsilon=1$, то $\mathbf{P}(1)\cap\mathbf{M}=\mathbf{M}$, т.е. построение регрессии основано только на использовании парных сравнений.

Утверждение 1 позволяет в явном виде получить крайние точки множества $\mathbf{P}(\varepsilon)\cap\mathbf{M}$, которые необходимы для решения $s=n$ задач оптимизации в рамках метода Лассо. Каждая задача оптимизации определяется одной из крайних точек $\mathbf{w}^{(k)}$, $k=1,...,n$. В результате получаем n стандартных задач Лассо, решение которых не представляет трудностей.

Числовые эксперименты

Множества данных. Числовые эксперименты осуществлялись для двух популяций удвоенных гаплоидных (УГ) линий ячменя.

1. Первое множество данных состоит из 93 УГ линий ячменя, описанных в работах [15, 16]. Данные по фенотипам и генотипам можно найти на сайте (http://wheat.pw.usda.gov/ggpages/maps/OWB/). Мы анализировали линии в соответствии с тремя фенотипическими свойствами: длина колоса (SL) в см; количество зерен (GN); время колошения (HD) в днях. Карта сцепления состоит из 1328 SNP-маркеров.
2. Второе множество данных состоит из 92 УГ линий ячменя, полученных от скрещивания Dicktoo×Morex и описанных в работе [17]. Данные по фенотипам и генотипам можно найти на сайте http://wheat.pw.usda.gov/ggpages/DxM/. Мы анализировали линии в соответствии с двумя фенотипическими свойствами: 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. Пропущенные данные во всех множествах данных оцениваются посредством следующей эвристической процедуры, которая может рассматриваться как некоторая модификация известного метода K ближайших соседей. Предположим, что вектор $\mathbf{X}_i$, соответствующий i-му SNP, имеет пропущенное данное на k-ой позиции, т.е., x_{ik} пропущено. Используя удельное расстояние Хэмминга между вектором $\mathbf{X}_i$ и всеми векторами $\mathbf{X}_j$, $j=1,...p$, $j\neq i$, мы отбираем K ближайших соседей $\mathbf{X}_{i_1},...,\mathbf{X}_{i_K}$ или K ближайших векторов. Для учета пропущенных значений они исключаются из вычисления расстояния Хэмминга. Вот почему используется удельное расстояние Хэмминга для сравнения векторов с различным числом пропущенных значений, т.е. мы вычисляем расстояние, приведенное к одному элементу $\mathbf{X}_i$. Доопределяемое значение принимается таким, которое соответствует наибольшему числу K значений на k-ой позиции всех ближайших соседей $X_{i_1k},...,X_{i_Kk}$. Значение K было выбрано 7.

Показатель ошибки. Из каждого множества данных случайным образом выбираются два подмножества: множество обучающих n примеров и множество тестирующих данных в количестве n_{test} примеров. Качество алгоритмов оценивается при помощи средней квадратической ошибки регрессии (СКОР), которая определяется как

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

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 невелико по сравнению с числом SNP и каждое наблюдение важно.

Первое множество данных. Сначала рассмотрим УГ линии ячменя из первой группы. Значения СКОР для первого множества данных приведены в табл. 1, где первый столбец соответствует трем анализируемым признакам, столбцы 2–5 содержат значения СКОР, используя 40 наиболее значимых SNP. При этом изучались случаи, когда точность определяется для всех линий (все линии), и для первых 10 линий с наименьшими значениями фенотипа (первые 10 линий). Сокращения СМ и ПМ соответствуют стандартному методу и предлагаемому методу с весами наблюдений соответственно. В качестве зна-

чений СКОР для предлагаемого метода выбраны наименьшие значения, полученные при различных ε в пределах от 0 до $0,5$. Можно заметить, что предлагаемый метод обеспечивает лучшие результаты в основном для первых 10 линий. При этом наибольший эффект достигается для времени колошения. В то же время первые два признака, длина колоса и количество зерен, не зависят от назначения весов и не демонстрируют какого-либо существенного уменьшения ошибки регрессии. Данный вывод имеет место также при использовании всех SNP-маркеров.

Наиболее значимые SNP					Все SNP			
	все линии		первые 10 линий		все линии		первые 10 линий	
Свойства	СМ	ПМ	СМ	ПМ	СМ	ПМ	СМ	ПМ
SL	1,44	1,36	1,552	1,392	2,684	2,684	2,45	2,45
GN	85,9	82,94	58,53	58,18	140,4	139,6	104,3	104,3
HD	47,72	33,01	60,72	16,866	98,58	94,92	92,38	86,82

Таблица 1. СКОР для стандартного и предлагаемого методов

В качестве иллюстрации зависимости СКОР от параметра засорения ε, определяющего множество $\mathbf{P}(\varepsilon)$, на рис. 2 показаны графики, полученные для признака SL. График с треугольными маркерами показывает зависимость при использовании всех линий, а график с прямоугольными маркерами иллюстрирует, как меняется ошибка регрессии при рассмотрении только первых 10 линий. Из рис. 2 можно увидеть, что существует некоторое оптимальное значение ε, при котором СКОР E достигает своего минимума. В рассматриваемом случае оптимальное значение ε для обоих графиков совпадает. В то же время рис. 2 показывает, что использование множества весов может даже снизить точность модели, когда используются все SNP-маркеры для моделирования. Важно отметить, что случай $\varepsilon = 0$ соответствует стандартному методу Лассо. Исходя из этого, ниже мы будем сравнивать полученные значения СКОР при различных $\varepsilon > 0$ со значением при $\varepsilon = 0$, т.е. со стандартным Лассо. Аналогичные графики для времени колошения представлены на рис. 3. Из рисунка видно существенное снижение ошибки регрессии. При использовании всех SNP-маркеров такое снижение практически отсутствует и имеет место только для $\varepsilon = 0,05$.

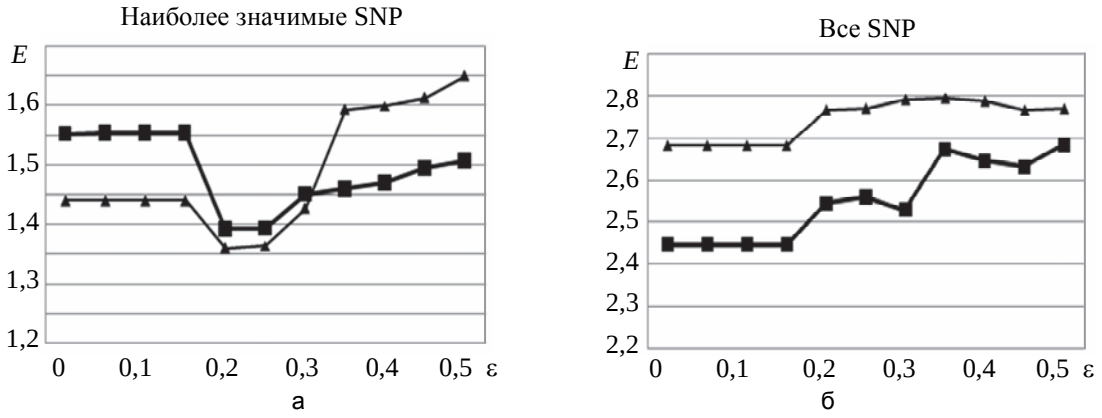

Рис. 2. Зависимости СКОР ($см^2$) от ε (отн.ед.) для наиболее значимых SNP(а) и для всех SNP(б), полученные для признака SL; треугольные маркеры – при использовании всех линий; прямоугольные маркеры – только первых 10 линий

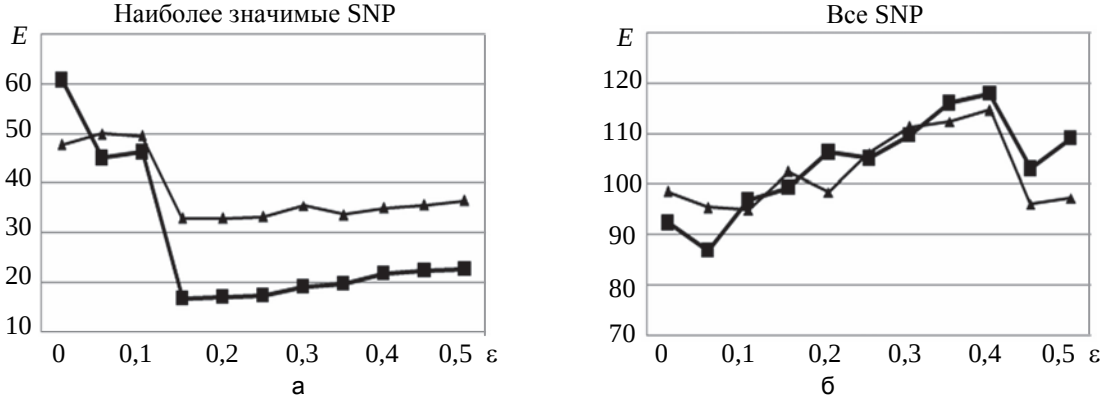

Рис. 3. Зависимости СКОР (количество дней^2) от ε (отн.ед.) для наиболее значимых SNP(а) и для всех SNP(б), полученные для времени колошения; треугольные маркеры – при использовании всех линий; прямоугольные маркеры – только первых 10 линий

Второе множество данных. Рассмотрим множество данных, состоящее из 92 УГ линий ячменя, полученное из скрещивания Dicktoo×Morex. Табл. 2 содержит показатели ошибки для множества данных Dicktoo×Morex при анализе двух признаков: время колошения с яровизацией и без яровизации. Из табл. 2 следует, что предлагаемый метод практически не влияет на точность модели, т.е. стандартный метод Лассо дает такие же результаты. Это также видно из рис. 4, где только для наиболее значимых SNP и при рассмотрении первых 10 линий наблюдается незначительное уменьшение СКОР.

	Наиболее значимые SNP				Все SNP			
	все линии		первые 10 линий		все линии		первые 10 линий	
Свойства	СМ	ПМ	СМ	ПМ	СМ	ПМ	СМ	ПМ
без яровизации	32,23	32,23	19,6	19,31	46,41	46,25	23,74	23,36
с яровизацией	29,05	28,3	24,6	23,4	38,52	38,55	27,4	27,02

Таблица 2. СКОР для Dicktoo×Morex

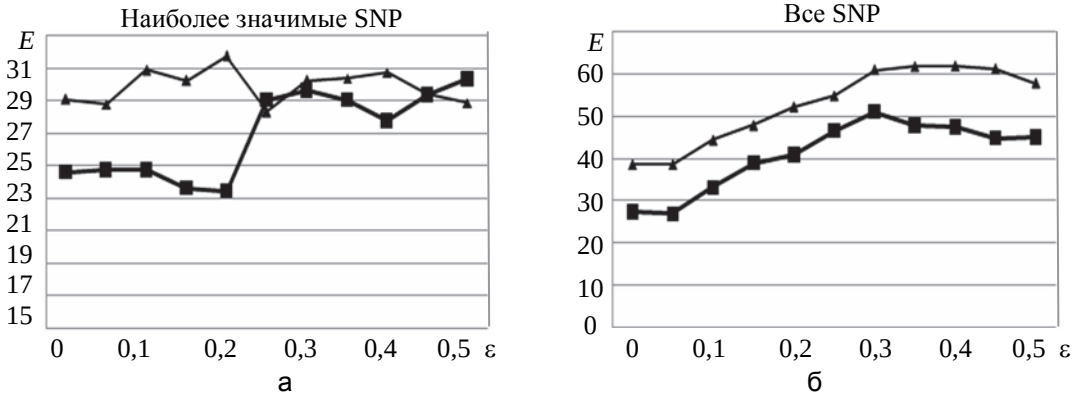

Рис. 4. Зависимости СКОР (количество дней^2) от ε (отн.ед.) для наиболее значимых SNP(а) и для всех SNP(б), полученные для Dicktoo × Morex; треугольные маркеры – при использовании всех линий; прямоугольные маркеры – только первых 10 линий

Сократим размер обучающей выборки для данных Dicktoo×Morex до $n = 30$, удалив из обучающего множества 62 элемента. Результаты анализа при различных условиях и $n = 30$ приведены в табл. 3. Мы можем наблюдать совершенно другую картину. Практически для всех условий и для двух признаков предлагаемый метод значительно превосходит стандартный. Это также можно видеть из рис. 5.

	Наиболее значимые SNP				Все SNP			
	все линии		первые 10 линий		все линии		первые 10 линий	
Свойства	СМ	ПМ	СМ	ПМ	СМ	ПМ	СМ	ПМ
без яровизации	18,88	18,88	19,30	18,62	45,37	44,55	51,23	46,52
с яровизацией	6,76	5,88	8,74	6,83	39,66	33,07	64,47	54,39

Таблица 3. СКОР для Dicktoo×Morex при малой выборке

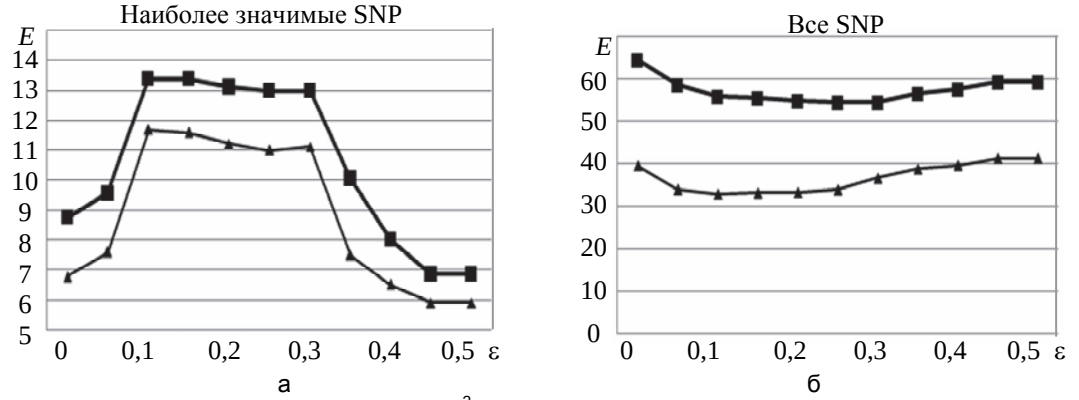

Рис. 5. Зависимости СКОР (количество дней^2) от ε (отн.ед.) для наиболее значимых SNP(а) и для всех SNP(б), полученные для Dicktoo×Morex при малой выборке; треугольные маркеры – при использовании всех линий; прямоугольные маркеры – только первых 10 линий

Заключение

Результаты числовых экспериментов и логика, лежащая в основе предлагаемого метода, продемонстрировали, что метод позволяет получить более точные характеристики по сравнению со стандартным методом Лассо, особенно, когда объем обучающей выборки мал.

Необходимо отметить, что полногеномный поиск ассоциаций был выполнен в качестве приложения для удвоенных гаплоидных линий ячменя. Однако метод может применяться без каких-либо изменений и для других объектов, например, для полногеномного поиска ассоциаций других организмов, для анализа экспрессии генов в геномах живых организмов на основе данных экспрессионных микрочипов.

В работе был предложен один из способов учета дополнительной информации в виде пересечения двух множеств весов наблюдений, образованных интервальной моделью засорения и парными сравнениями. При этом ключевым элементом, позволяющим использовать множества в задачах оптимизации, была возможность получения конечного набора крайних точек пересечения. Аналогичным образом можно разработать другие, возможно, более эффективные подходы для учета дополнительной информации, образующие множества весов. Это является одним из направлений дальнейших исследований.

References

1. 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. doi: 10.1214/09-STS306
2. Altidor W., Khoshgoftaar T.M., Van Hulse J., Napolitano A. Ensemble feature ranking methods for data intensive computing applications. In *Handbook of Data Intensive Computing*. NY, Springer, 2011, pp. 349–376. doi: 10.1007/978-1-4614-1415-5_13
3. Lee I.-H., Lushington G.H., Visvanathan M. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics*, 2011, vol. 1, no. 11 , art. 11. doi: 10.1186/2043-9113-1-11
4. Kohavi R., John G.H. Wrappers for feature subset selection. *Artificial Intelligence*, 1997, vol. 97, no. 1–2, pp. 273–324.
5. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002, vol. 46, no. 1–3, pp. 389–422. doi: 10.1023/A:1012487302797
6. Lander E.S., Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 1989, vol. 121, no. 1, pp. 185–199.
7. Lal T.N., Chapelle O., Weston J., Elisseeff A. Embedded methods. In *Feature Extraction*. Springer, 2006. V. 207. P. 137–165. doi: 10.1007/978-3-540-35488-8_6
8. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, vol. 58, no. 1, pp. 267–288.
9. 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, no. 2, pp. 301–320. doi: 10.1111/j.1467-9868.2005.00503.x
10. Gu X., Yin G., Lee J.J. Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary Clinical Trials*, 2013, vol. 36, no. 2, pp. 642–650. doi: 10.1016/j.cct.2013.09.009
11. Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*, 2013, vol. 1019, pp. 149–169. doi: 10.1007/978-1-62703-447-0-6
12. Walley P. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall, 1991, 362 p.
13. Draper N., Smith H. Applied Regression Analysis. 2nd ed. NY, John Wiley and Sons, 1981, 709 p.
14. Huber P.J. *Robust Statistics*. NY, Wiley, 1981, 320 p.
15. Chutimanitsakun Y., Nipper R.W., Cuesta-Marcos A., Cistue L., Corey A., Filichkina T., Johnson E.A., Hayes P.M. Construction and application for qtl analysis of a restriction site associated DNA (rad) linkage map in barley. *BMC Genomics*, 2011, vol. 12, art. 4. doi: 10.1186/1471-2164-12-4
16. Cistue L., Cuesta-Marcos A., Chao S., Echavarri B., Chutimanitsakun Y., Corey A., Filichkina T., Garcia-Marino N., Romagosa I., Hayes P.M. Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and Applied Genetics*, 2011, vol. 122, no. 7, pp. 1399–1410. doi: 10.1007/s00122-011-1540-9
17. Hayes P., Chen F.Q., Corey A., Pan A., Chen T.H.H., Baird E., Powell W., Thomas W., Waugh R., Bedo Z., Karsai I., Blake T., Oberthur L. The Dicktoo x Morex population. *Plant Cold Hardiness*, 1997, pp. 77–87. doi: 10.1007/978-1-4899-0277-1_8

Уткин Лев Владимирович – доктор технических наук, профессор, профессор, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, 195251, Российская Федерация, lev.utkin@gmail.com

Жук Юлия Александровна – кандидат педагогических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; доцент, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, Санкт-Петербург, 194021, Российская Федерация, zhuk_yua@mail.ru

Коолен Франк – PhD, профессор, профессор, Университет Дарем, Дарем, DH1 3LE, Великобритания, frank.coolen@durham.ac.uk

Lev V. Utkin – D.Sc., Professor, Professor, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, 195251, Russian Federation, lev.utkin@gmail.com

Yulia A. Zhuk – PhD, Associate professor, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation; Associate professor, Saint Petersburg State Forest Technical University, Saint Petersburg, 194021, Russian Federation, zhuk_yua@mail.ru

Frank Coolen – PhD, Professor, Professor, Durham University, Durham, DH1 3LE, UK, frank.coolen@durham.ac.uk