Теоретические основы информатики

Theoretical foundations of computer science

УДК 57.087.1

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ЭКЗОНОВ ГЕНОВ ЧЕЛОВЕКА

А. В. ВОЛКОВ[1)]***, Н. Н. ЯЦКОВ***[1)]***, В. В. ГРИНЕВ***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассмотрена задача сокращения размерности пространства признаков экзонов человека с целью определить их генную принадлежность. Для оценки эффективности алгоритмов отбора признаков проведены вычислительные эксперименты на примерах экзонов 14 известных генов человека. Установлено, что экзоны четко разделимы относительно генной принадлежности. Алгоритмы автоматического отбора чувствительны к шумовым признакам и позволяют оценить количество таких признаков. Сокращение числа последних улучшает производительность вычислений и потребление памяти, а также позволяет получать значительно более простые прогностические модели и повышает их интерпретируемость. Показано, что тренировка алгоритмов индуктивного обучения на признаках фланкирующих интронов обеспечивает более высокую предсказательную способность в сравнении с обучением алгоритмов на признаках экзонов. Результаты представленной работы открывают новые возможности для изучения организации генов человека с помощью алгоритмов машинного обучения.

Ключевые слова: экзон; интрон; биоинформатика; отбор признаков; имитационное моделирование; алгоритм классификации.

Образец цитирования:
Волков АВ, Яцков НН, Гринев ВВ. Отбор информативных признаков экзонов генов человека. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;1:77–89.
https://doi.org/10.33581/2520-6508-2019-1-77-89

For citation:
Volkau AU, Yatskou MM, Grinev VV. Selecting informative features of human gene exons. *Journal of the Belarusian State University. Mathematics and Informatics*. 2019;1:77–89. Russian.
https://doi.org/10.33581/2520-6508-2019-1-77-89

Авторы:
Андрей Владимирович Волков – аспирант кафедры системного анализа и компьютерного моделирования факультета радиофизики и компьютерных технологий. Научный руководитель – Н. Н. Яцков.
Николай Николаевич Яцков – кандидат физико-математических наук, доцент; доцент кафедры системного анализа и компьютерного моделирования факультета радиофизики и компьютерных технологий.
Василий Викторович Гринев – кандидат биологических наук, доцент; доцент кафедры генетики биологического факультета.

Authors:
Andrei U. Volkau, postgraduate student at the department of system analysis and computer simulation, faculty of radiophysics and computer technologies.
andrei@cybergizer.com
Mikalai M. Yatskou, PhD (physics and mathematics), docent; associate professor at the department of system analysis and computer simulation, faculty of radiophysics and computer technologies.
yatskou@bsu.by
Vasily V. Grinev, PhD (biology), docent; associate professor at the department of genetics, faculty of biology.
grinev_vv@bsu.by

Журнал Белорусского государственного университета. Математика. Информатика. 2019;1:77–89
Journal of the Belarusian State University. Mathematics and Informatics. 2019;1:77–89

SELECTING INFORMATIVE FEATURES OF HUMAN GENE EXONS

A. U. VOLKAU [a], ***M. M. YATSKOU*** [a], ***V. V. GRINEV*** [a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: A. U. Volkau (andrei@cybergizer.com)

Dimensionality reduction of the human gene exon feature space is considered with the aim of gene identification. To evaluate the performance of various feature selection algorithms, computational experiments were carried out using the examples of exons of 14 known human genes. It is proven that exons are clearly separable regarding gene affiliation. Feature selection algorithms are sensitive to noise features and allow to estimate their number. Reducing the number of features improves CPU-time, memory usage as well as reduces the complexity of a model and makes it easier to interpret. Our findings indicate that utilizing of features of flanking intronic sequences leads to better prediction models in comparison with utilizing of exon features. The results of the research provide new opportunities for study of human gene data using machine learning algorithms.

Key words: exon; intron; bioinformatics; feature selection; simulation modeling; classification algorithm.

Введение

Исследование организации и функционирования генов (в том числе и онкогенов) человека является важной задачей биоинформатики [1]. Гены состоят из экзонов и интронов. Особый интерес представляют экзоны, поскольку из них формируются зрелые молекулы РНК, а на основе последних происходит синтез белков в клетке.

Экзон может быть описан с помощью набора признаков. Признаками экзонов являются длины нуклеотидных последовательностей, биофизические свойства экзонов, измеренные экспериментальным путем, признаки фланкирующих нуклеотидных участков [2]. Экзон характеризуется большим количеством признаков (более 1000), в то же время число экзонов, принадлежащих гену, невелико (как правило, менее 200).

Проблема большого числа признаков и относительно малого числа объектов наблюдений характерна для всей области биоинформатики в целом. К примеру, для предсказания альтернативных транскриптов генов человека необходимо точно классифицировать экзоны согласно генной принадлежности [1], однако обучение алгоритмов классификации затрудняется «проклятием размерности». Возможным решением данной проблемы является использование алгоритмов снижения размерности пространства признаков [3].

Алгоритмы снижения размерности данных делятся на две группы: алгоритмы преобразования признаков (англ. *feature extraction*) и алгоритмы автоматического отбора признаков (англ. *feature selection*). Алгоритмы преобразования признаков проецируют исходное пространство признаков в новое пространство низкой размерности, которое часто является линейным или нелинейным преобразованием исходного пространства. Среди алгоритмов преобразования признаков следует выделить методы главных компонент, факторного анализа, многомерного шкалирования, линейного дискриминантного анализа, канонического корреляционного анализа и сингулярного разложения [4; 5]. Алгоритмы отбора признаков осуществляют непосредственный выбор наиболее релевантных признаков из исходного множества. Отсутствие каких-либо преобразований над исходными признаками позволяет сохранить их физический смысл. Данное свойство является особенно значимым в биоинформатических приложениях, поскольку каждый из признаков имеет уникальный биологический смысл, важный для эксперта в этой области. Применение алгоритмов отбора для выделения наиболее информативных признаков экзонов генов человека может существенно улучшить эффективность анализа геномных данных, а именно: увеличить производительность вычислений, повысить эффективность определения генной принадлежности экзонов, улучшить точность прогностических моделей [3].

Цель данного исследования – выяснение принципиальной возможности предсказания генной принадлежности экзонов по их признакам, выделение наиболее информативных признаков экзонов, определение эффективных алгоритмов отбора признаков индуктивного обучения в контексте решаемой задачи. В работе представлен краткий обзор существующих алгоритмов автоматического отбора признаков объектов, выполнен сравнительный анализ наиболее эффективных алгоритмов на примере экзонов 14 генов человека, выделены наилучшие по информативности группы признаков экзонов.

Алгоритмы автоматического отбора признаков объектов

Большинство алгоритмов отбора признаков основаны на принципе выбора более информативных (релевантных) признаков и удалении всех остальных. Интуитивно признак может считаться релевантным, если содержит некоторую информацию о метке класса характеризуемого объекта. Формальное определение предложено в работе [6] и может быть сформулировано следующим образом. Признак X является *строго релевантным*, если удаление признака X из обучающей выборки приводит к ухудшению предсказательной способности оптимального байесовского классификатора. Признак X имеет *слабую релевантность*, если он не является строго релевантным и при этом существует подмножество признаков S таких, что предсказательная способность оптимального байесовского классификатора, обученного на наборе признаков S, является менее точной по сравнению с предсказательной способностью при обучении на $S \cup X$. Признак следует считать *нерелевантным*, если он не является строго или слабо релевантным.

Помимо типов признаков на основе релевантности, используется определение избыточности. Ее строгое определение представлено в [7]: признак следует считать *избыточным*, если он имеет слабую релевантность и образует покрытие Маркова с подмножеством остальных признаков. Оптимальное подмножество признаков должно содержать все строго релевантные признаки и слабо релевантные признаки, которые не являются избыточными.

Алгоритмы автоматического отбора признаков можно разделить на семейства контролируемого и неконтролируемого отбора. При контролируемом отборе выбирается подмножество признаков с учетом метки классов объектов данных, в то время как при неконтролируемом отборе такая метка не учитывается [8]. В настоящей работе рассмотрены алгоритмы контролируемого отбора признаков.

Алгоритмы автоматического отбора признаков, в соответствии с используемой стратегией поиска релевантных наборов признаков, формируют четыре большие группы:

- фильтрующие (от англ. *filter*);
- оберточные (от англ. *wrapper*);
- встроенные (от англ. *embedded*);
- гибридные (от англ. *hybrid*) [9].

Фильтрующие алгоритмы выбирают подмножество признаков без применения процедуры индуктивного обучения, что позволяет эффективно использовать их в задачах анализа большого количества признаков (более 100). Однако вследствие данного подхода результирующие наборы признаков не являются в равной степени оптимальными для различных алгоритмов индуктивного обучения.

Оберточные алгоритмы итеративно используют процедуру индуктивного обучения для оценки точности предсказания на подмножествах признаков. Алгоритмы характеризуются высокой вычислительной сложностью и склонны к переобучению при анализе выборок небольших размеров.

Встроенные алгоритмы осуществляют отбор признаков непосредственно в ходе процесса индуктивного обучения, что способствует определению оптимального набора признаков для используемого индуктивного метода. В сравнении с оберточными алгоритмами встроенные алгоритмы имеют значительно бóльшую вычислительную эффективность, поскольку не нуждаются в итеративной оценке наборов признаков.

Гибридные алгоритмы создаются на основе комбинации фильтрующих и оберточных алгоритмов. Автоматический отбор признаков выполняется в два шага. На первом шаге выбирается подмножество признаков с использованием процедуры фильтрующего алгоритма. На втором шаге осуществляется отбор признаков с помощью оберточного алгоритма. В гибридных алгоритмах сведены к минимуму недостатки фильтрующих и оберточных алгоритмов.

Алгоритмы отбора признаков подразделяются на универиативные и мультивариативные. Мультивариативный подход учитывает отношения множества признаков значительно эффективнее универиативного и позволяет устранять избыточность признаков [10].

По результатам проведенного литературного обзора в настоящей работе было принято решение об использовании фильтрующих алгоритмов, так как они:

- обладают высоким быстродействием, что обеспечивает проведение вычислительного эксперимента при малых вычислительных затратах;
- не зависят от типа алгоритма индуктивного обучения;
- относительно просты для программной реализации.

Алгоритмы Relief/ReliefF (на основе критерия Фишера и индекса Джини) были выбраны как репрезентативные представители группы фильтрующих алгоритмов.

Алгоритм на основе критерия Фишера относится к классу универиативных алгоритмов. Наиболее информативными являются признаки, имеющие близкие значения для экзонов, принадлежащих

к одному гену, и отличные значения для экзонов, принадлежащих разным генам. Алгоритм позволяет ранжировать признаки по информативности для классификации. По причине того что алгоритм на основе критерия Фишера оценивает каждый признак независимо от остальных, он не может отфильтровывать избыточные признаки.

Положим число признаков равным M. Оценка Фишера для признака f_i ($i = 1, 2, \ldots, M$):

$$F(f_i) = \frac{\sum_{j=1}^{c} n_j \left(\mu_{i,j} - \mu_i\right)^2}{\sum_{j=1}^{c} n_j \sigma(i,j)^2},$$

где c – число классов; n_j – число наблюдений в классе j; $\mu_{i,j}$ – среднее значение признака f_i для объектов наблюдения, соответствующих классу j; μ_i – среднее значение признака f_i; $\sigma(i,j)^2$ – значение дисперсии признака f_i для объектов наблюдения, соответствующих классу j. В качестве наилучших выбираются k признаков с наибольшими оценками Фишера.

Алгоритм Relief впервые предложен в работе [11] для анализа двухклассовых наборов данных и относится к классу мультивариативных. Положим, что l экзонов случайным образом выбраны среди n экзонов, тогда критерий *Relief* для признака f_i рассчитывается следующим образом:

$$Relief(f_i) = \frac{1}{2}\sum_{j=1}^{l} d\left(X(j,i) - X\left(NM(j), i\right)\right) - d\left(X(j,i) - X\left(NH(j), i\right)\right),$$

где $d(\cdot)$ – метрика расстояния (наиболее часто используется расстояние Евклида); $X \in \mathbb{R}^{n \times M}$ – матрица данных с n экзонами и M признаками; $NM(j)$ – ближайшие соседние экзоны к экзону x_j для случая одного класса; $NH(j)$ – ближайшие соседние экзоны к экзону x_j для случая разных классов.

Алгоритм ReliefF представляет собой адаптацию алгоритма Relief к анализу многоклассовых данных с высоким уровнем шума [12]. Значение критерия *ReliefF* вычисляется как

$$ReliefF(f_i) = \frac{1}{c}\sum_{j=1}^{l}\left(-\frac{1}{m_j}\sum_{x_r \in NH(j)} d\left(X(j,i) - X(r,i)\right) + \sum_{y \neq y_j}\frac{1}{h_{jy}}\frac{p(y)}{1-p(y)}\sum_{x_r \in NM(j,y)} d\left(X(j,i) - X(r,i)\right)\right),$$

где c – число классов; $NH(j)$ – ближайшие экзоны к экзону x_j в рамках одного класса; m_j – размерность $NH(j)$; $d(\cdot)$ – метрика расстояния; $X \in \mathbb{R}^{n \times M}$ – матрица данных с n экзонами и M признаками; $NM(j,y)$ – ближайшие экзоны к экзону x_j в рамках разных классов; h_{jy} – размерность $NM(j,y)$; $p(y)$ – относительное количество объектов, имеющих метку класса y.

Алгоритм отбора признаков на основе индекса Джини является унивариативным [13]. Он определяет степень дискриминирующей способности признака по отношению к метке класса (генной принадлежности) с помощью критерия-индекса Джини. Выражение для индекса Джини:

$$G(r) = 1 - \sum_{j=1}^{c}\left[p(j|r)\right]^2,$$

где c – число классов; $p(j|r)$ – относительная частота j-го класса в узле дерева r (или по выделенному признаку).

Алгоритм оценивает каждый из признаков независимо, поэтому не может выделить избыточные признаки.

Программные реализации алгоритмов автоматического отбора признаков

Многие популярные алгоритмы автоматического отбора признаков входят в состав программного пакета для интеллектуального анализа данных *scikit-feature*. Среди них следует выделить алгоритмы на основе критериев Фишера, индекса Джини и прироста информации, ReliefF, минимальной избыточности – максимальной релевантности. Пакет разработан на языке программирования *Python* и тесно интегрирован с пакетами *scikit-learn*, *Numpy* и *Scipy*.

Сравнительный анализ алгоритмов отбора признаков экзонов генов является сложной задачей в связи с отсутствием априорной информации об оптимальном наборе релевантных признаков, наличием

избыточных, нерелевантных признаков, большим числом признаков, поэтому всестороннее исследование алгоритмов часто проводят на искусственно сгенерированных данных, для которых установлены наборы релевантных признаков. В работе [14] представлены результаты сравнительного анализа 7 фильтрующих, 2 оберточных и 2 встроенных алгоритмов на примерах 11 имитационно смоделированных наборов данных с нерелевантными (шумовыми) и избыточными признаками или с числом признаков, намного превышающим количество объектов наблюдения. Установлено, что алгоритм Relief/ReliefF в среднем дает наилучшие результаты для наборов данных с различными свойствами.

В работе [15] показано, что характеристики исследуемых наборов данных (такие как число признаков, число наблюдений и т. п.) сильно влияют на стабильность алгоритма отбора признаков. Для оценки последней выбран алгоритм расчета индекса стабильности [16].

Алгоритмы индуктивного обучения

Эффективность алгоритмов контролируемого отбора признаков определяется с помощью измерения информативности отбираемых наборов признаков, для оценки которой используется предсказательная способность алгоритма индуктивного обучения, обученного на отобранных признаках.

В работе рассмотрены три алгоритма индуктивного обучения: наивный байесовский классификатор, метод *k* ближайших соседей и машина опорных векторов с линейным ядром [17]. Выбор алгоритмов обусловлен их популярностью, разнородностью подходов к индуктивному обучению, отсутствием встроенных механизмов отбора признаков.

Предсказание вероятности принадлежности экзона к гену алгоритмом классификации позволяет использовать так называемые скоринговые правила (или правила счета) для оценки качества вероятностного прогноза. Строго корректные скоринговые правила, включая скоринговое правило Брайера и логарифмическое скоринговое правило, являются стандартными метриками вероятностных прогнозов [18]. В настоящей работе применяется скоринговое правило Брайера, которое для двухклассовых наборов данных имеет вид

$$Оценка\_Брайера = \frac{1}{N}\sum_{t=1}^{N}\left(f_t - o_t\right)^2,$$

где N – число экзонов в тестовой выборке; f_t – вероятность принадлежности экзона t к заданному гену; o_t – метка принадлежности экзона t к заданному гену (единица означает принадлежность, нуль – отсутствие).

Скоринговое правило Брайера многоклассовых наборов данных (R классов) определяется следующим образом:

$$Оценка\_Брайера = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}\left(f_{ti} - o_{ti}\right)^2,$$

где N – число экзонов в тестовой выборке; R – число генов; f_{ti} – вероятность принадлежности экзона t к гену i; o_{ti} – метка принадлежности экзона t к гену i.

Минимальное значение скорингового правила Брайера соответствует наилучшему или более точному вероятностному прогнозу.

Экспериментальные данные

Экспериментальные данные получены из базы Ensembl [19] и содержат 1762 уникальных экзона, принадлежащих к 14 произвольно отобранным генам (названия генов обозначены в номенклатуре Ensembl): ENSG00000239665, ENSG00000166444, ENSG00000165795, ENSG00000205336, ENSG00000196628, ENSG00000226674, ENSG00000231898, ENSG00000237298, ENSG00000236172, ENSG00000228486, ENSG00000242808, ENSG00000228956, ENSG00000242086 и ENSG00000154556.

Каждый экзон был охарактеризован с помощью 1198 численных признаков: 429 признаков непосредственно самих экзонов и 769 признаков фланкирующих участков нуклеотидных последовательностей (длина цепи составляет 100 нуклеотидов). Различия между участками гена, соответствующими признакам экзонных нуклеотидных последовательностей, и участками гена, соответствующими признакам фланкирующих экзонов последовательностей интронов, показаны на рис. 1. Признаки фланкирующих экзонов нуклеотидных участков представляют особый интерес в ходе тестирования реализованных алгоритмов, так как интроны не содержат структурной информации о белке, но играют важную роль во время сплайсинга (соединения) экзонов, что, предположительно, должно позволить дополнительно проверить точность прогнозирования разработанных алгоритмов. Пример набора данных для генов ENSG00000239665 и ENSG00000237298 демонстрируется в табл. 1.

Число экзонов для каждого из изученных генов приведено в табл. 2. Индекс сбалансированности классов IR представляет собой отношение числа объектов наблюдения (в данном случае – экзонов) в доминирующем классе к числу объектов наблюдения в минорном классе [20].

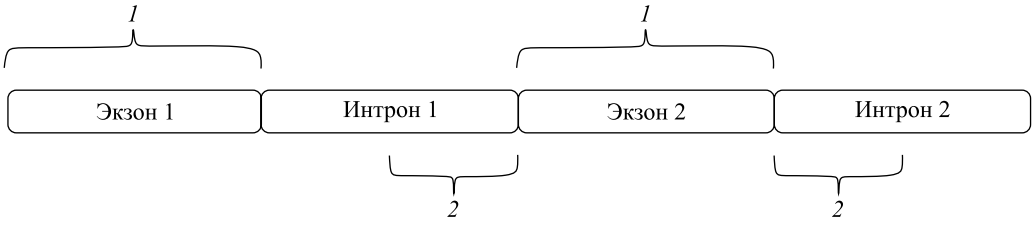

Рис. 1. Области соответствия признакам экзонных нуклеотидных последовательностей (*1*) и фланкирующих интронов (*2*)

Fig. 1. Regions of compliance with the features of exon nucleotide sequence (*1*) and of flanking introns (*2*)

Таблица 1

Примеры признаков экзонов генов человека

Table 1

Examples of exon features of human genes

Экзон	Признак			Ген
	phastCons_TE	phyloP_TE	Exon_length	
chr10:13631143-13631363	0,004 061 71	−0,394 81	221	ENSG00000239665
chr10:13631288-13631363	0,000 103 605	−0,532 445	76	ENSG00000239665
chr10:13631296-13631363	0,000 115 794	−0,623 316	68	ENSG00000239665
chr10:13631333-13631363	0	−0,654 904	31	ENSG00000239665
chr2:178764186-178764368	0,620 799	2,441 66	183	ENSG00000237298
chr2:178764186-178764612	0,464 401	1,987 85	427	ENSG00000237298
chr2:178773275-178773335	0,935 975	4,421 6	61	ENSG00000237298
chr2:178774227-178774319	0,916 519	4,305 95	93	ENSG00000237298
chr2:178774227-178774658	0,544 692	2,450 13	432	ENSG00000237298

Примечание. Признаки phastCons_TE и phyloP_TE показывают эволюционный консерватизм (неизменность в ходе эволюции) экзонов, рассчитаны с помощью алгоритмов phastCons и phyloP соответственно. Признак Exon_length является мерой длины экзона, выраженной в количестве нуклеотидов.

Таблица 2

Характеристики наборов данных, включающих по 2 гена, для случая бинарной классификации экзонов

Table 2

Characteristics of data sets comprising 2 genes each in case of binary exon classification

Номер пары генов	Составные гены	Число экзонов	IR	Совокупное число экзонов
1	ENSG00000239665	93	1,37	220
	ENSG00000237298	127		
2	ENSG00000166444	121	1,04	247
	ENSG00000226674	126		
3	ENSG00000165795	106	1,01	213
	ENSG00000236172	107		

Окончание табл. 2
Ending table 2

Номер пары генов	Составные гены	Число экзонов	IR	Совокупное число экзонов
4	ENSG00000205336	156	1,23	283
	ENSG00000231898	127		
5	ENSG00000228486	93	1,35	219
	ENSG00000226674	126		
6	ENSG00000242808	97	1,34	227
	ENSG00000154556	130		
7	ENSG00000228956	118	1,10	225
	ENSG00000236172	107		
8	ENSG00000239665	93	2,38	314
	ENSG00000242086	221		
9	ENSG00000196628	140	1,44	237
	ENSG00000242808	97		
10	ENSG00000166444	121	1,08	251
	ENSG00000154556	130		

Методология исследования

Блок-схема организации вычислительного эксперимента для исследования алгоритмов автоматического отбора признаков экзонов показана на рис. 2. Представленный подход позволяет исследовать зависимость оценки предсказательной способности алгоритмов индуктивного обучения по скоринговому правилу Брайера для ранжированного ряда признаков с помощью алгоритмов автоматического отбора.

В блоке 1 формируется набор данных для анализа. В ходе работы рассмотрены наборы данных с числом генов $M = 2, 3, \ldots, 14$. Для каждого из экзонов указана принадлежность к модельному гену человека. В блоке 2 производится формирование выборки набора данных для M различных генов. Количество разных выборок установлено равным 10.

В блоке 3 осуществляется разбиение данных на подмножества тренировочной и тестируемой выборок, использующиеся далее для перекрестной проверки. В блоке 4 выполняется стандартизация тренировочного и тестового множеств.

В блоке 5 признаки ранжируются по значимости, рассчитанной на основе применения заданного алгоритма их автоматического отбора. В блоке 6 значение индекса N (количество наилучших признаков) полагается равным единице.

В блоке 7 осуществляется обучение алгоритма индуктивного обучения на N признаках и оценка релевантности выбранного набора признаков с помощью подсчета оценки Брайера. В блоке 8 проверяется условие исследования всех признаков. Если не все признаки исследованы, происходит переход к блоку 9, в котором значение счетчика индекса N увеличивается на единицу. В случае если все признаки исследованы, проверяется условие исследования всех множеств, сгенерированных для 10-кратной перекрестной проверки (блок 10). Если не все множества исследованы, осуществляется переход к блоку 3, если все множества исследованы – к блоку 11.

В блоке 11 производится проверка условия исследования всех выборок для M различных генов. Переход к блоку 2 для формирования новой выборки из M различных генов происходит, если не все выборки для M различных генов исследованы. Иначе осуществляется переход к блоку 12, в котором выполняется усреднение значений точности классификации, полученных для всех выборок из M генов.

В табл. 2 представлены характеристики десяти исследованных пар генов для случая бинарной классификации экзонов. Особенностью сформированных наборов является доминирование числа признаков над числом объектов данных (в среднем число признаков в 5 раз превосходит число объектов).

Сравнительный анализ алгоритмов отбора признаков с целью оценить стабильность такого отбора проведен с помощью алгоритма 10-кратной перекрестной проверки. Отбираемые на тренировочных множествах ранжированные ряды из 25 наилучших признаков поступали на вход алгоритма оценки стабильности отбора признаков. Для получения достоверных результатов эксперимент повторялся 10 раз для различных комбинаций генов, после чего данные усреднялись.

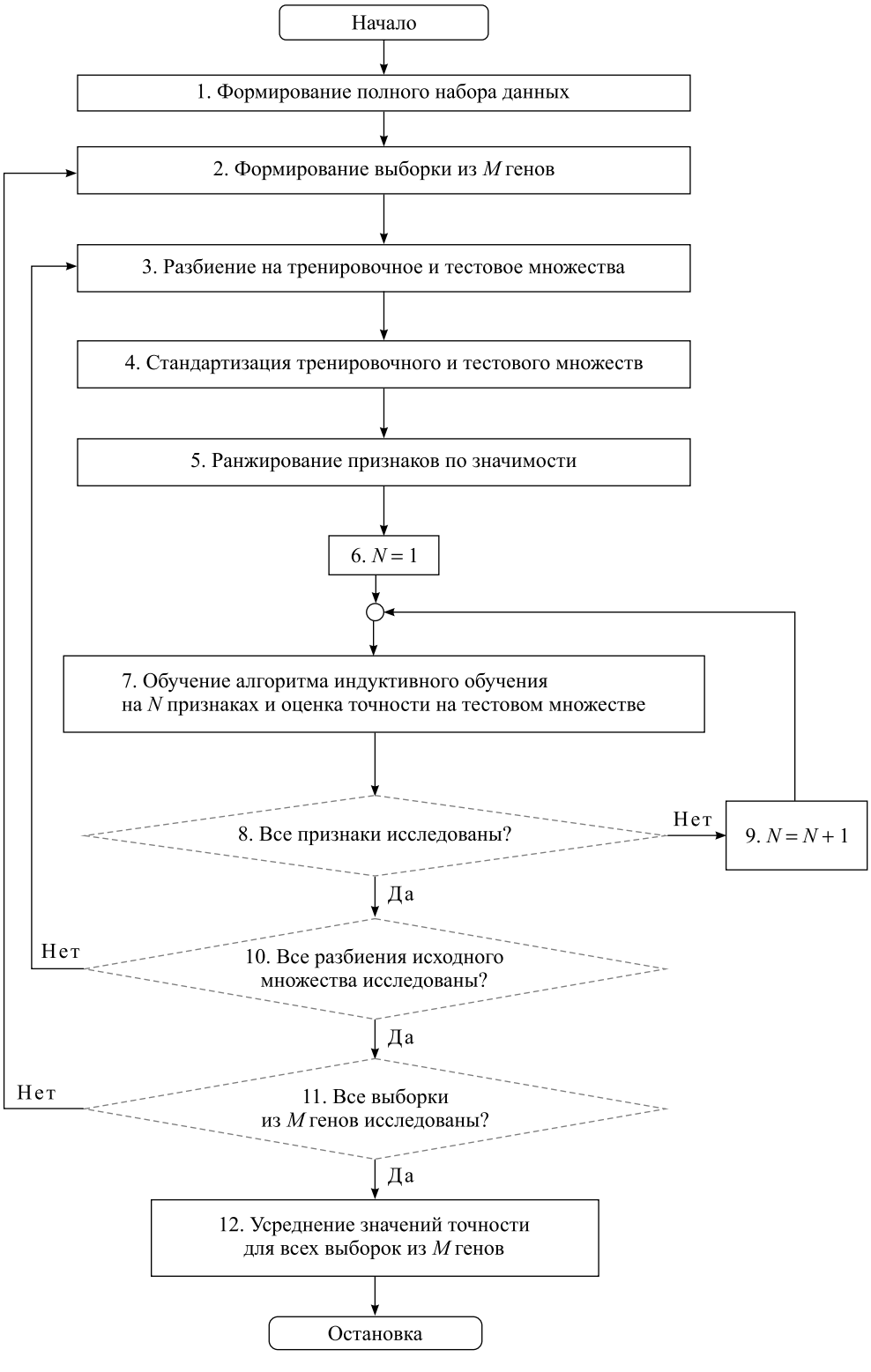

Рис. 2. Блок-схема организации вычислительного эксперимента
Fig. 2. Flow chart of the computing experiment

Компьютерная генерация дополнительных признаков экзонов

Для проверки релевантности признаков экзонов и чувствительности алгоритмов автоматического отбора нерелевантных признаков реализована имитационная модель [21]. В ней производится генерация дополнительных нерелевантных признаков, представляющих собой шумовые (здесь и далее используется прямой перевод термина с англ. *noisy features*) или неинформативные некоррелированные признаки.

Рассмотрим задачу генерации набора данных дополнительных признаков x_{ij}, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, L$, представляющих единый кластер с центром μ и дисперсией D (разброс объектов) в пространстве признаков. Реализации дополнительных координат экзонов x_{ij} задаются выражением

$$x_{ij} = \mu + z\sigma,$$

где z – реализация нормальной стандартизированной случайной величины $N(0, 1)$.

Выбор нормального распределения для генерации шумовых признаков обусловлен тем, что большинство признаков экзонов имеют указанное распределение.

Результаты и их обсуждение

Выполнен сравнительный анализ алгоритмов случайного отбора признаков (алгоритмов на основе критерия Фишера, индекса Джини, Relief/ReliefF) при использовании алгоритмов индуктивного обучения (метод k ближайших соседей, алгоритм машины опорных векторов с линейным ядром и алгоритм наивной байесовской классификации) на примерах сформированных наборов данных (см. табл. 2). Под случайным отбором понимается выборка признаков, полученная случайным образом. Зависимость оценки по скоринговому правилу Брайера для ранжированных алгоритмами рядов признаков при варьировании алгоритмов индуктивного обучения представлена на рис. 3.

Для рассмотренных алгоритмов классификации характерно резкое убывание оценки ошибки до 50–100 признаков, затем слабое убывание (метод опорных векторов) или возрастание (k ближайших соседей и наивный байесовский классификатор) от 100 до 1200 признаков. Полученные значения оценок ошибок по скоринговому правилу Брайера позволяют сделать важный вывод о разделимости экзонов, принадлежащих различным генам. Достаточно первых 100 ранжированных признаков для точного предсказания генной принадлежности (оценка Брайера 0,02–0,04).

Наилучшую точность среди исследованных алгоритмов индуктивного обучения демонстрируют метод k ближайших соседей (1 ближайший сосед) и машина опорных векторов с линейным ядром. Минимальное достигаемое значение оценки по скоринговому правилу Брайера для обоих случаев составляет 0,02. Приведенные зависимости свидетельствуют о том, что автоматический отбор позволяет выбрать относительно небольшое число наиболее информативных признаков (до 50 признаков), приводящее к резкому уменьшению оценки Брайера и ее существенному отличию от случайного отбора выборок из такого же количества признаков. При увеличении числа признаков наблюдается лишь незначительное преимущество автоматического отбора признаков над случайным. Это свидетельствует о том, что абсолютное большинство признаков могут быть отнесены к классу избыточных. Следует отметить, что увеличение числа признаков приводит к небольшому ухудшению точности классификации алгоритмов k ближайших соседей и наивного байесовского классификатора. Среди исследованных алгоритмов классификации наихудшие результаты демонстрирует наивный байесовский классификатор, что обусловлено отсутствием учета корреляционных зависимостей между признаками в алгоритме классификатора.

На рис. 4 представлена зависимость оценок по скоринговому правилу Брайера для 100 наиболее информативных признаков при варьировании числа классов (генов) и алгоритма отбора признаков. При этом использованы алгоритмы индуктивного обучения, показывающие наилучшие результаты: машина опорных векторов с линейным ядром и метод одного ближайшего соседа. При увеличении числа классов характер зависимостей счетов Брайера сохраняется, однако значения изменяются от 0,1–0,14 для трех генов до 0,17–0,22 для шести генов, что обусловлено увеличением вероятности принятия классификатором неверного решения. Алгоритмы автоматического отбора признаков позволяют в среднем на 2,5 % точнее предсказывать генную принадлежность экзонов в сравнении со случайным отбором признаков.

В результате анализа фланкирующих участков нуклеотидных последовательностей обнаружено (рис. 5), что более низкие значения оценок Брайера, равные 0,02–0,03, соответствуют обучению алгоритмов на признаках фланкирующих интронов в сравнении с величиной 0,07–0,08 для признаков экзонных нуклеотидных последовательностей.

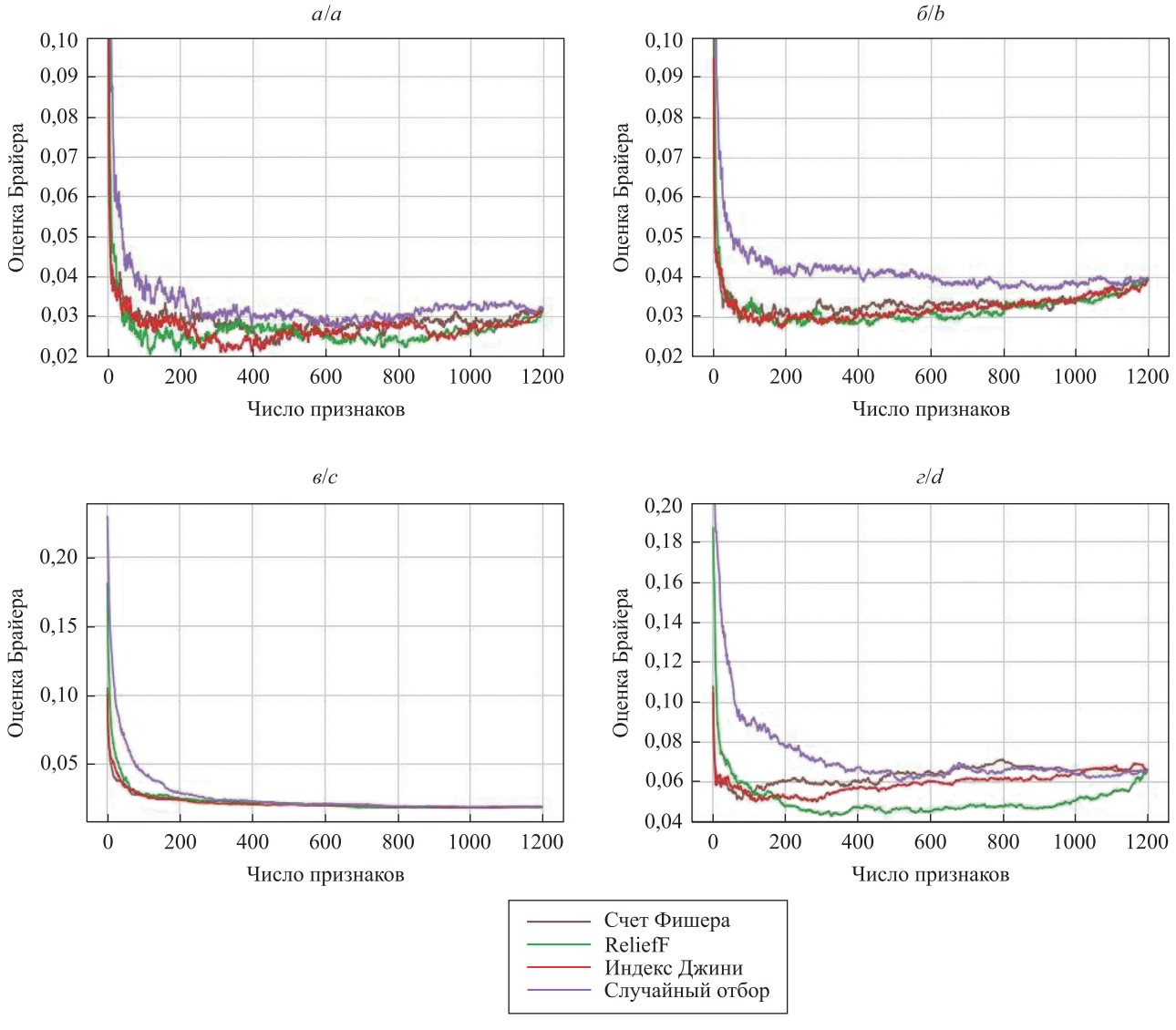

Рис. 3. Зависимость оценки по скоринговому правилу Брайера
для ранжированного ряда признаков (2 гена):
а – метод одного ближайшего соседа; *б* – метод трех ближайших соседей;
в – машина опорных векторов с линейным ядром; *г* – наивный байесовский классификатор

Fig. 3. The Brier score for a ranked list of features (case of 2 genes):
a – 1-nearest neighbor algorithm; *b* – 3-nearest neighbors algorithm;
c – support vector machine with linear kernel; *d* – naive Bayes classifier

Рассмотрим работоспособность алгоритмов отбора признаков на примере добавления к экспериментальным данным дополнительных смоделированных признаков для 2 генов. В ходе анализа исследован набор 178 лучших признаков, отобранных экспертным путем, а также 1020 смоделированных шумовых признаков, сгенерированных с использованием нормального распределения $N(0, 1)$. Для рассмотренных закономерностей характерны три фазы на графике оценки ошибки (рис. 6):

1) резкое убывание ошибки до величины 0,03–0,04 для первых 20 признаков;

2) фаза участка «плато» для признаков 20–180, на котором значение ошибки остается практически неизменным: 0,03–0,05;

3) линейное увеличение ошибки 0,05–0,1 для признаков от 180 до 1198.

Фазы 1) и 2) хорошо согласуются с результатами анализа полного экспериментального набора признаков экзонов без учета смоделированных шумовых признаков. Влиянием последних обусловлена фаза 3). Это позволяет предположить, что добавление шумовых, нерелевантных, некоррелированных признаков должно приводить к линейному увеличению ошибки, рассчитанной согласно скоринговому правилу Брайера.

Зависимость оценки ошибки для случайного отбора признаков имеет монотонно убывающий характер и значительно превышает среднюю величину оценки ошибки для алгоритмов автоматического отбора признаков. Значения ошибок для всех рассмотренных алгоритмов совпадают для итогового набора 1999 экспериментальных и смоделированных признаков. Различия ошибок для алгоритмов случайного и неслучайного отбора признаков позволяют сделать следующие выводы:

- реализованные алгоритмы отбора признаков крайне чувствительны к наличию шумовых признаков;
- признаки экзонов не являются шумовыми;
- наличие шумовых признаков приводит к линейному увеличению ошибки правила Брайера;
- количество шумовых признаков может быть оценено по графику ошибки Брайера и соответствует числу признаков, для которых наблюдается увеличение ошибки.

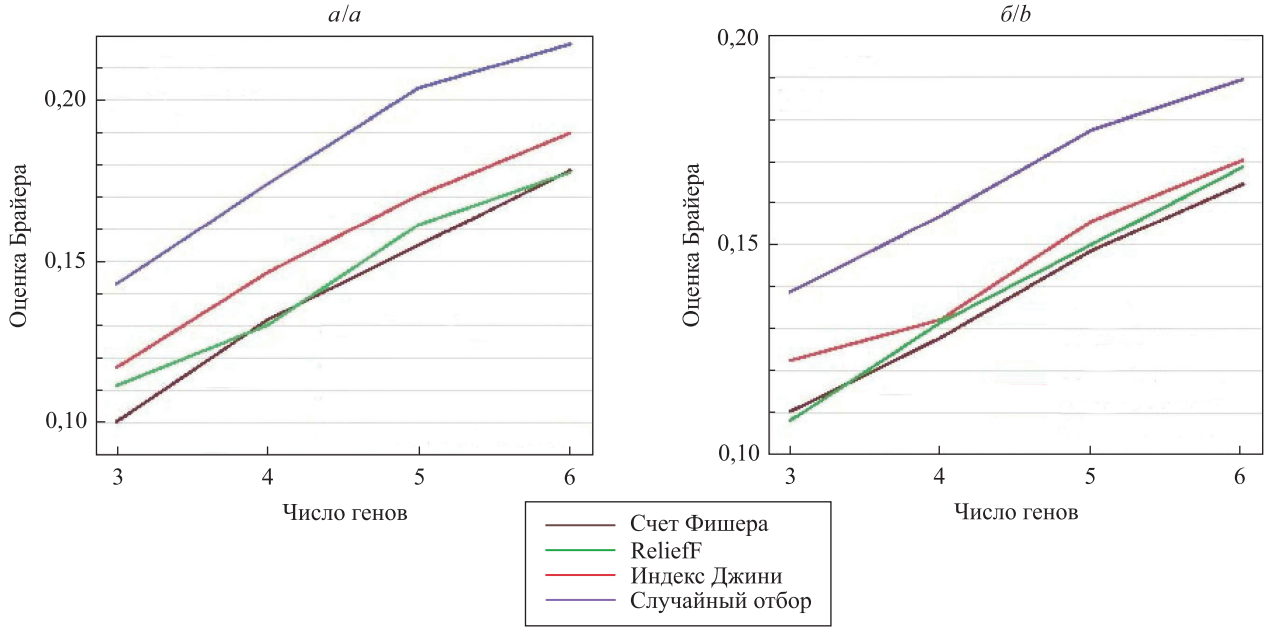

Рис. 4. Зависимость оценки по скоринговому правилу Брайера
для 100 признаков при варьировании числа классов:
а – машина опорных векторов с линейным ядром; *б* – метод одного ближайшего соседа

Fig. 4. The Brier score for 100 features versus the number of genes:
a – support vector machine with linear kernel; *b* – 1-nearest neighbor algorithm

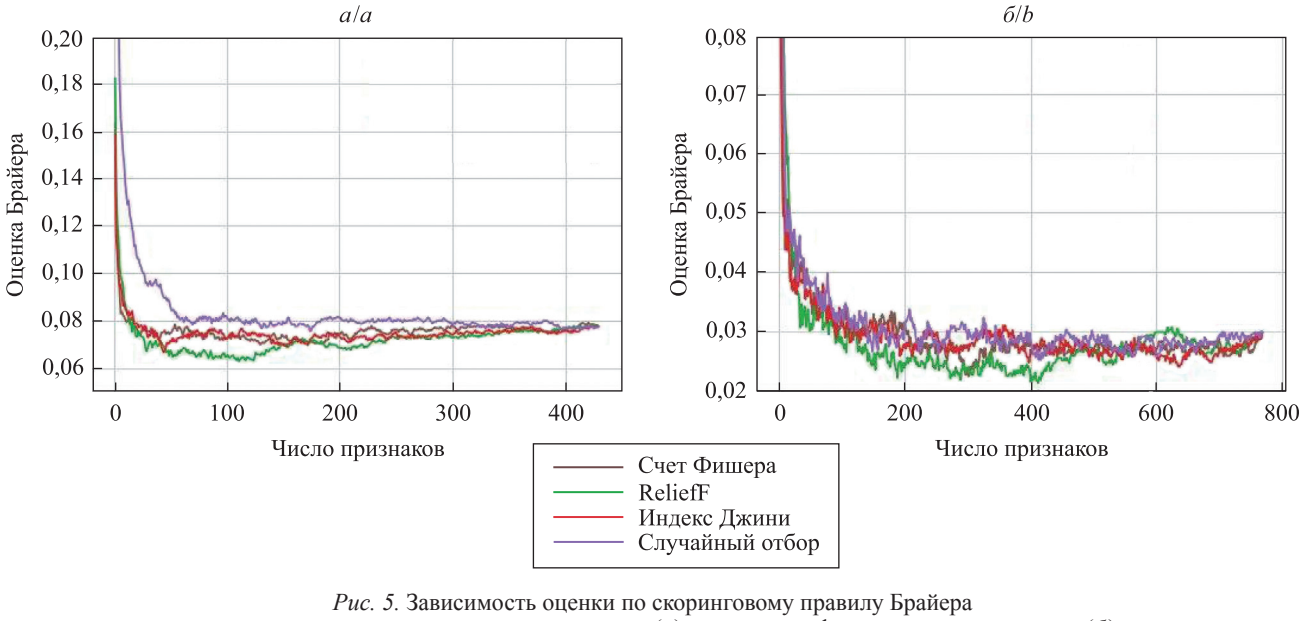

Рис. 5. Зависимость оценки по скоринговому правилу Брайера
для ранжированного ряда признаков экзонов (*а*) и признаков фланкирующих интронов (*б*)

Fig. 5. The Brier score for a ranked list of exon features (*a*)
and a ranked list of features of flanking nucleotide sequences (*b*)

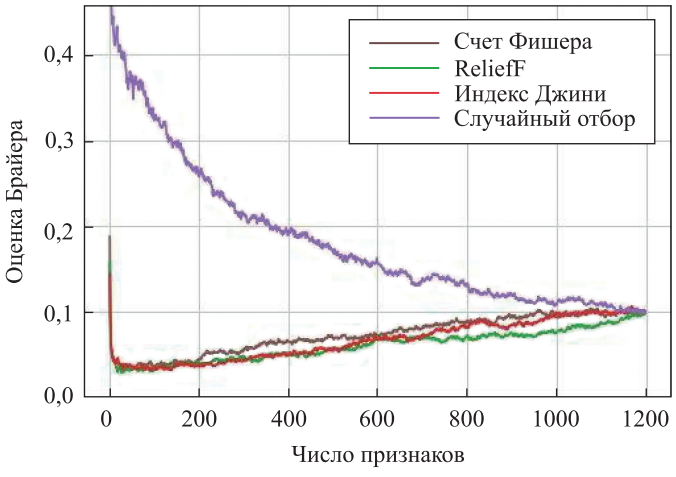

Рис. 6. Зависимость оценки по скоринговому правилу Брайера (метод одного ближайшего соседа) для ранжированного ряда признаков в случае добавления 1020 шумовых признаков

Fig. 6. The Brier score for a ranked list of features, when 1020 artificial noisy features are added

В завершение рассмотрим поведение индекса стабильности, рассчитанного в ходе анализа (табл. 3). Реализованные алгоритмы демонстрируют высокие значения (0,75–1) индекса стабильности для алгоритмов автоматического отбора признаков. Стабильность случайного отбора признаков равна нулю. Алгоритм на основе критерия Фишера и алгоритм отбора признаков на основе индекса Джини обладают наивысшими значениями индекса (0,85–1).

Таблица 3

Значения индексов стабильности

Table 3

Stability index values

Количество генов	Алгоритм на основе критерия Фишера	Relief/ReliefF	Индекс Джини
2	0,863	0,794	0,850
3	0,864	0,759	0,853
5	0,881	0,759	1
14	0,926	0,835	1

Заключение

Исследована эффективность алгоритмов отбора признаков (алгоритмы на основе критериев Фишера, Relief/ReliefF, индекса Джини) и алгоритмов индуктивного обучения (наивный байесовский классификатор, метод k ближайших соседей с одним и тремя ближайшими соседями, машина опорных векторов с линейным ядром) на примерах классификации экзонов генов человека. Установлен факт существенной разделимости между экзонами, принадлежащими разным генам. Наилучшая точность классификации достигается для наборов, состоящих из экзонов 2 генов, когда значение оценки по скоринговому правилу Брайера достигает 0,02.

Автоматический отбор признаков позволяет выбрать относительно небольшое число (до 100) наиболее информативных признаков. Среди исследованных алгоритмов отбора признаков алгоритм на основе критерия Фишера демонстрирует наивысшую вычислительную эффективность при практически идентичных показателях точности и, как следствие, является наилучшим для отбора признаков экзонов. В контексте стабильности отбора признаков экзонов наилучшим является алгоритм, основанный на вычислении индекса Джини.

В ходе работы проведено исследование наборов данных с добавлением смоделированных шумовых признаков. Установлено, что алгоритмы автоматического отбора признаков крайне чувствительны к наличию шумовых признаков, которое приводит к линейному увеличению ошибки по скоринговому

правилу Брайера. Количество шумовых признаков может быть определено по графику оценки Брайера и соответствует числу признаков, для которых наблюдается увеличение ошибки. При этом признаки экзонов в большинстве своем не являются шумовыми. Тренировка алгоритмов индуктивного обучения на признаках фланкирующих интронов обеспечивает более высокую предсказательную способность классификаторов по сравнению с обучением на признаках экзонов. Это наблюдение представляет большой интерес и требует дальнейшего детального изучения с помощью методов биоинформатики, а также экспериментальных методов молекулярной биологии.

Библиографические ссылки / References

1. Grinev VV, Migas AA, Kirsanava AD, Mishkova OA, Siomava N, Ramanouskaya TV, et al. Decoding of exon splicing patterns in the human RUNX1–RUNX1T1 fusion gene. *International Journal of Biochemistry & Cell Biology.* 2015;68:48–58. DOI: 10.1016/j.biocel.2015.08.017.

2. Zhang M. Statistical features of human exons and their flanking regions. *Human Molecular Genetics.* 1998;7(5):919–932. DOI: 10.1093/hmg/7.5.919.

3. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007;23(19):2507–2517. DOI: 10.1093/bioinformatics/btm344.

4. Cox TF. Ch. 16. Multidimensional scaling in process control. *Handbook of Statistics.* 2003;22:609–623. DOI: 10.1016/s0169-7161(03)22018-6.

5. Martinez AM, Kak AC. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2001;23(2):228–233. DOI: 10.1109/34.908974.

6. John GH, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. In: Cohen WW, Hirsh H, editors. *Machine Learning Proceedings. Proceedings of the Eleventh International Conference; 1994 July 10–13; New Brunswick, Canada*. New Brunswick: Rutgers University; 1994. p. 121–129. DOI: 10.1016/b978-1-55860-335-6.50023-4.

7. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research.* 2004;5:1205–1224.

8. Ang JC, Mirzal A, Haron H, Hamed HNA. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology nad Bioinformatics.* 2016;13(5):971–989. DOI: 10.1109/tcbb.2015.2478454.

9. Belanche LA, Gonzalez FF. Review and evaluation of feature selection algorithms in synthetic problems [Internet]. [Cited 2018 September 12]. Available from: http://arxiv.org/abs/1101.2320.

10. Wang L, Lei Y, Zeng Y, Tong L, Yan B. Principal feature analysis: a multivariate feature selection method for fMRI data. *Computational and Mathematical Methods in Medicine.* 2013;2013:1–7. DOI: 10.1155/2013/645921.

11. Kira K, Rendell LA. A practical approach to feature selection. In: *Machine Learning Proceedings. Proceedings of the Ninth International Workshop on Machine Learning; 1992 July 1–3; Aberdeen, Scotland.* Aberdeen: ML; 1992. p. 249–256. DOI: 10.1016/b978-1-55860-247-2.50037-1.

12. Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: *Machine Learning: ECML-94. European Conference; 1994 April 6–8; Catania, Italy.* Berlin: Springer; 1994. p. 171–182. DOI: 10.1007/3-540-57868-4_57.

13. Singh SR, Murthy HA, Gonsalves TA. Feature selection for text classification based on Gini coefficient of inequality. *The Fourth Workshop on Feature Selection in Data Mining*. 2010;10:76–85.

14. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data. *Knowledge and Information Systems.* 2012;34(3):483–519. DOI: 10.1007/s10115-012-0487-8.

15. Kalousis A, Prados J, Hilario M. Stability of Feature Selection Algorithms: a study on high dimensional spaces. *Knowledge and information System*. 2007;12(1):95–116.

16. Nogueira S, Sechidis K, Brown G. On the stability of feature selection. *Journal of Machine Learning Research.* 2018;18(174):1–54.

17. Nilsson NJ. Artificial intelligence: A modern approach. *Artificial Intelligence. Elsevier BV.* 1996;82(1–2):369–380. DOI: 10.1016/0004-3702(96)00007-0.

18. Merkle EC, Steyvers M. Choosing a Strictly Proper Scoring Rule. *Decision Analysis.* 2013;10(4):292–304. DOI: 10.1287/deca.2013.0280.

19. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. *Database.* 2016;2016:baw093. DOI: 10.1093/database/baw093.

20. Orriols-Puig A, Bernadó-Mansilla E. Evolutionary rule-based systems for imbalanced data sets. *Soft Computing.* 2008;13(3):213–225. DOI: 10.1007/s00500-008-0319-7.

21. Qiu W, Joe H. Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification.* 2006;23(2):315–334. DOI: 10.1007/s00357-006-0018-y.

Статья поступила в редколлегию 18.09.2018.
Received by editorial board 18.09.2018.