*Математическая биология и биоинформатика*
*2020. Т. 15. № 1. С. 4–19. doi: 10.17537/2020.15.4*

========== **БИОИНФОРМАТИКА** ==========

# Подход к отбору значимых признаков при решении биомедицинских задач бинарной классификации данных с микрочипов

**Бойко И.Ю.*** **Анисимов Д.С., Смолякова Л.Л., Рязанов М.А.**

*Алтайский государственный университет, Барнаул, Россия*

***Аннотация.*** В ряде современных биомедицинских исследований, направленных на поиск методов ранней диагностики онкологических заболеваний, используются микрочипы, содержащие определенную биологическую информацию о пациентах. На основе этих данных происходит отнесение пациентов к одному из двух классов, соответствующих наличию и отсутствию у пациента некоторого диагноза. При решении данной задачи отбор значимых признаков является одним из этапов, оказывающих решающее влияние на качество классификации. В данной работе предлагается критерий отбора значимых признаков, основанный на использовании ledge-коэффициента корреляции, введенного ранее для оценки степени взаимосвязи числового и бинарного признаков. Для двух наборов данных с микрочипов приведены сравнительные примеры их бинарной классификации с использованием трех алгоритмов отбора признаков, трех методов уменьшения размерности, шести моделей классификации. Использование ledge-критерия отбора признаков позволило получить качество классификации, сравнимое с результатами использования распространенных методов отбора признаков, таких как: *t*-критерий, *U*-критерий. Для рассмотренного в работе набора данных с пептидных микрочипов ранее была выявлена эффективность применения метода проекции на латентные структуры. Использование этого метода в сочетании с отбором значимых признаков ledge-критерием позволило получить более высокий показатель качества классификации.

***Ключевые слова:*** *отбор признаков, ledge-коэффициент, бинарная классификация, микрочипы, ROC-кривая, метод проекции на латентные структуры*.

## ВВЕДЕНИЕ

Развитие технологий сбора и хранения информации, используемых в современных биомедицинских исследованиях, приводит к увеличению объема данных, подлежащих анализу, что вызывает потребность в развитии методов их обработки. Одним из актуальных примеров является класс задач обработки многомерных данных, полученных с пептидных, либо ДНК-микрочипов [1, 2]. Среди задач данного типа большое распространение имеют проблемы бинарной классификации, которые исследуются, например, с целью поиска методов ранней диагностики онкологических заболеваний [3, 4]. Задачи бинарной классификации состоят в отнесении каждого пациента некоторого данного множества к одному из двух классов, исходя из информации, отражающей результаты определенной диагностики состояния его организма. В таких случаях, как правило, данные представлены числовыми значениями очень большого набора биомедицинских признаков. Их количество может доходить до сотен тысяч, что многократно увеличивает время анализа данных. В связи с этим существует следующая актуальная проблема.

*jehll@ya.ru

Для качественного решения задач бинарной классификации в рассматриваемой предметной области важно определить, какой набор числовых признаков из имеющихся в наличии может наилучшим образом разделять классы, то есть выступать в качестве «индикатора» принадлежности рассматриваемого объекта к одному из классов [5]. Решение рассматриваемой проблемы позволит существенно снизить объем вычислений при выполнении классификации. Кроме того, исключение признаков, шумовая составляющая которых значительно преобладает над информационной, часто содержит в себе потенциал для того, чтобы получить более высокое качество классификации.

Для решения задачи отбора признаков существует множество алгоритмов. Эти методы делят на три группы: алгоритмы фильтрации, алгоритмы обертки, встроенные алгоритмы [6, 7].

Концепция фильтров состоит в том, что задается некоторая мера, согласно которой происходит оценка и упорядочивание признаков. Затем с использованием некоторого отсекающего правила выбирается искомое подмножество. В отличие от алгоритмов остальных групп, в фильтрах не используются методы машинного обучения, поэтому алгоритмы фильтрации наиболее вычислительно эффективны [6]. Благодаря данному преимуществу фильтры широко применяются к наборам данных с большим количеством признаков. Однако, алгоритмы фильтрации зачастую не предназначены для выявления взаимосвязей между признаками и совместного влияния нескольких признаков на целевой, поэтому не всегда в полном объеме отбирают потенциально полезную информацию [7, 8].

Для выбора значимых признаков из данных с микрочипов наиболее часто используются алгоритмы фильтрации в силу высокой скорости их работы. В настоящее время широко распространено применение фильтров, основанных на использовании *t*-критерия Стьюдента [9–15] и *U*-критерия Манна – Уитни (Wilcoxon rank-sum test) [10–13]. Часто применяют фильтры, основанные на корреляции [9], например, используют коэффициенты Пирсона или Спирмена [16]. Также распространено применение метода Холла (Correlation-based Feature Selection) и Fast Correlation-Based Filter [17].

В алгоритмах обертки (wrappers) происходит построение определенной модели машинного обучения на различных подмножествах признаков, получаемых, например, путем использования генетического алгоритма или последовательного добавления признаков [6, 18]. Искомым подмножеством является такое из рассмотренных, на котором модель достигает максимального качества. Алгоритмы обертки, в отличие от фильтров, нередко позволяют получить более высокое качество классификации, но являются наименее вычислительно эффективными, потому что на каждой итерации алгоритма происходит повторное обучение выбранной модели. Следовательно, такие методы практически неприменимы к данным с микрочипов, состоящим из большого количества признаков.

Во встроенных алгоритмах отбор признаков выполняется в ходе построения модели машинного обучения. Такие методы работают быстрее, чем алгоритмы обертки, но медленнее, чем фильтры. Встроенные алгоритмы для отбора значимых признаков с микрочипов основаны на использовании таких подходов как: метод опорных векторов с рекурсивным исключением признаков (SVM-RFE) [19, 20], случайный лес (Random Forest) [21], Лассо (LASSO) [22].

В настоящее время распространены гибридные методы, которые заключаются в последовательном использовании алгоритмов нескольких групп, что позволяет получить преимущества различных подходов к отбору значимых признаков. Например, из большого количества исходных признаков путем фильтрации выбирается небольшой набор, который затем передается в алгоритм обертки для выявления искомого множества [23–25]. Более того, набирает популярность подход к построению моделей данного типа, который заключается в отборе значимых признаков путем создания ансамбля фильтров

[26]. Таким образом удается снизить нестабильность результатов классификации различных наборов данных, которая может иметь место при использовании лишь одного некоторого выбранного исследователем алгоритма фильтрации [27, 28].

При работе с информацией о биомаркерах выход имеющегося в наличии числового биомедицинского признака за определенные границы часто является доводом в пользу идентификации паталогического состояния организма пациента, что свидетельствует о наличии связи между числовым и бинарным признаками [11, 29]. Иными словами, важно выделить те числовые признаки, которые имеют статистически значимую взаимосвязь с бинарным, характеризующим наличие, либо отсутствие паталогического состояния организма пациента [8, 11]. Вышеизложенные рассуждения приводят к выводу о необходимости исследования взаимосвязи числового и бинарного признаков при решении проблемы отбора значимых признаков из данных с микрочипов. Распространенные на практике в настоящее время методы отбора признаков не вполне сосредоточены на выявлении связи рассматриваемого типа в силу ее существенной нелинейности, поэтому в следующих работах был введен ledge-коэффициент корреляции для оценивания силы такой связи, а также алгоритмы по его вычислению [30, 31]. Далее нами предлагается критерий отбора значимых признаков из данных с микрочипов, который представляет собой алгоритм фильтрации, основанный на использовании ledge-коэффициента корреляции. Затем для двух наборов данных, собранных, соответственно, с пептидных и ДНК-микрочипов выполняется бинарная классификация с использованием как предложенного метода, так и распространенных подходов к отбору значимых признаков.

## МАТЕРИАЛЫ И МЕТОДЫ

### Ledge-коэффициент корреляции

Числовые медицинские признаки часто имеют нормативные интервалы, выход за которые может являться доводом в пользу идентификации паталогического состояния организма пациента. На основе данного положения в работе [30] введена связь типа «ступенька», которую мы для краткости будем называть связью типа ledge (от английского ledge – ступенька). Далее приведем ее описание. Пусть у каждого из $n$ объектов измеряются две характеристики, то есть имеем две связанные выборки $\mathbf{X}$ и $\mathbf{Y}$ объема $n$, где $\mathbf{X}$ – числовой медицинский признак, $\mathbf{Y}$ – бинарный, характеризующий, например, наличие, либо отсутствие некоторого паталогического состояния у данного наблюдаемого. Эти выборки имеют вид

$$\mathbf{X} = (x_1, x_2, ..., x_n);\ \mathbf{Y} = (y_1, y_2, ..., y_n).$$

Измеренные значения пары характеристик $i$-го объекта обозначим $(x_i, y_i)$. При этом $x_i \in \mathbb{R}$, $y_i \in \{0,1\}$, $i = 1,...,n$. Если существует связь типа ledge признаков $\mathbf{X}$ и $\mathbf{Y}$, то при нахождении числовых значений в некоторых границах $a, b \in \mathbf{X}$ бинарный признак в основном должен принимать одно и то же значение, а в случае максимально сильной связи значения выборки $\mathbf{Y}$ имеют вид $y_i = y_{a,b}(x_i)$, $i = 1,...,n$ и выражаются формулой

$$y_{a,b}(x_i) = \begin{cases} 1, & a \le x_i \le b, \\ 0 & \text{иначе.} \end{cases} \quad (1)$$

Также в статье [30] вводится ledge-коэффициент корреляции, который оценивает, насколько точно облако точек $(x_i, y_i)$ при наилучшем выборе границ $a$, $b$ соответствует идеальному случаю рассматриваемой связи, описанному формулой (1).

Отметим, что для исследования связи типа ledge числовые значения выборки **X** не важны. Выполняется сортировка набора пар $(x_i, y_i)$ по возрастанию первой компоненты. Из полученных в результате такого упорядочивания значений второй компоненты составляется бинарная цепочка **Y**, которая и изучается. Числовые значения теперь можно опустить, заменив их порядковыми номерами (рангами). То есть, будем считать, что $x_i = i, i = 1,...,n$. Согласно работе [30] ledge-коэффициент корреляции задается формулой

$$L_E(\mathbf{Y}) = 1 - \frac{S(\mathbf{Y})}{S_{k,m}},$$

где $S(\mathbf{Y}) = \min\limits_{a,b} \sum\limits_{j=1}^{n} (y_j - y_{a,b}(j))^2$ характеризует отклонение наблюдаемой картины от идеальной, выраженной (1), и называется числом ошибок, а $S_{k,m}$ определяется формулой $S_{k,m} = \begin{cases} k-1, & k < m+1 \\ m, & k \geq m+1 \end{cases}$, где $k$ и $m$ – это, соответственно, число единиц и нулей в рассматриваемой бинарной цепочке. Чем больше значение коэффициента $L_E$, тем сильнее связь.

## Проверка гипотезы о наличии связи типа ledge

Рассмотрим вопрос о том, насколько большим должно быть значение ledge-коэффициента для имеющейся бинарной цепочки, чтобы сделать вывод о наличии, либо отсутствии изучаемой связи. Обычно подобные задачи решаются путем сравнения некоторой статистики с ее критическим значением. Опишем алгоритм поиска такого критического значения для ledge-коэффициента.

Пусть бинарная цепочка отсортирована по увеличению значений соответствующего ей числового признака. Требуется установить, насколько сильно связаны рассматриваемые выборки. Пусть связи нет вовсе (назовем это предположение основной гипотезой $H_0$). Тогда можно считать, что бинарная цепочка формируется случайным образом. Разумеется, при этом она может получиться такой, что ошибок в рассматриваемом нами смысле будет немного, что может вынудить нас принять гипотезу о наличии изучаемой связи (это – альтернативная гипотеза $H_1$). Потому нас интересует вопрос: насколько большим должно быть значение ledge-коэффициента для имеющейся у нас (вообще говоря, случайной) бинарной цепочки, чтобы отвергнуть гипотезу относительно ее чистой случайности и прийти к выводу о наличии связи нужного вида на статистически значимом уровне?

Отсюда возникает самая общая задача. Задан размер выборки $n$ и разумно малое положительное $\varepsilon > 0$. Указать такое критическое значение $L_E(\varepsilon, n)$, что при выполнении основной гипотезы $H_0$: $P(L_E > L_E(\varepsilon, n)) \leq \varepsilon$.

Если такое критическое значение $L_E(\varepsilon, n)$ будет найдено, а рассчитанный по нашей бинарной цепочке коэффициент окажется больше него, то следует признать установленным факт связи типа ledge между числовым и бинарным признаками с вероятностью ошибки $\varepsilon$.

Для полностью корректного решения поставленной задачи, таким образом, необходимо знать распределение этого коэффициента как дискретной случайной величины в предположении полной случайности той бинарной цепочки, по которой он рассчитывается.

Рассмотрим упорядоченное по убыванию множество значений коэффициента $L_E$, рассчитанных по достаточно большому количеству $N$ бинарных цепочек. Выделим в нем подмножество $F_n$ из первых $\lfloor \varepsilon N \rfloor$ элементов (где символ $\lfloor \ \rfloor$ означает округление вниз до ближайшего целого). Это означает, что мы выбрали $\varepsilon \cdot 100\,\%$ из имеющихся коэффициентов, которым соответствуют бинарные цепочки с наиболее сильной связью. Значение последнего (т.е. минимального) элемента, попавшего во множество $F_n$ и будет искомым критическим значением ledge-коэффициента $L_E(\varepsilon, n)$ для заданных $n$ и $\varepsilon$.

Идея нахождения $L_E(\varepsilon, n)$ состоит в том, чтобы при заданном $n$ сгенерировать множество из $N$ бинарных цепочек, вычисляя для каждой $L_E$, а затем по изложенному выше правилу отыскать $L_E(\varepsilon, n)$ для выбранного $\varepsilon$. Далее будем генерировать $M$ раз по $N$ $(1 \ll N \ll 2^n)$ случайных бинарных цепочек из $2^n$ возможных, находить $L_E^i(\varepsilon, n),\ i = 1, ..., M$ для каждого из $M$ этих наборов цепочек, а затем в качестве итогового критического значения $L_E(\varepsilon, n)$ выберем медиану получившегося распределения значений $L_E^i(\varepsilon, n),\ i = 1, ..., M$. Необходимость выбора $N$, $M$ здесь связана с трудностью использования метода полного перебора при больших значениях $n$ из-за высокой скорости возрастания функции количества возможных бинарных цепочек $y = 2^n$, а значит значительного увеличения количества операций алгоритма и, следовательно, времени его выполнения, с ростом $n$.

Также на практике вполне возможна ситуация, когда при достаточно большом размере выборок найдутся элементы числовой выборки **X** с одинаковыми значениями. Причем, соответствующие им значения **Y**, вообще говоря, могут быть различны. Поэтому при обработке данных перед заменой значений **X** на их порядковые номера следует каждому набору повторяющихся элементов сопоставить долю единиц от общего числа соответствующих элементов **Y** при данном числовом значении (значения **Y**, равные единице, указывают на отсутствие заболевания). После этого выборка **Y** формально тоже является числовой, поэтому перед вычислением ledge-коэффициента следует выполнить бинаризацию ее значений. Порог бинаризации представляет собой долю единиц (здоровых пациентов) среди значений **Y** с одинаковым соответствующим значением **X**, которую мы принимаем достаточной для того, чтобы считать данное числовое значение находящимся в пределах «нормы». Следовательно, при решении задач анализа биомедицинских данных более естественным является выбор порога бинаризации, близкого к единице.

**Описание наборов данных**

Для исследования были использованы два набора данных. Первый набор – данные пептидных микрочипов Российско-Американского противоракового центра (далее данные 330К) [32]. Данные состоят из двух групп пациентов – с диагнозом рак молочной железы (РМЖ) и контрольных доноров (КД). Группа РМЖ содержит 40 доноров, группа КД – 41 донор. Для каждого донора было проведено два технических повтора. При этом каждый микрочип состоит из 330034 пептидов. В качестве количественной меры нами использовалась медиана светимости пептида на длине волны 532 нм. Итоговый набор данных состоит из 162 объектов, и 330034 признаков с диапазоном значений по каждому признаку от 0 до 65535.

Второй набор данных основан на анализе ДНК-микрочипов. Группа ученых во главе с Э. Гравье в своем исследовании [33] выполнили сбор, предварительную подготовку и анализ данных о 168 пациентах, у которых был диагностирован рак молочной железы (более точная терминология представлена в цитированной выше работе). По результатам

5 лет наблюдений после диагностики объекты данных были размечены следующим образом. Класс A сопоставили 111 пациентам (за время наблюдений не произошло появление метастазов). Класс B был сопоставлен остальным 57 пациентам (произошло появление метастазов). Кроме того, авторы приведенной выше работы разработали классификатор, предсказывающий группу риска появления метастазов для объектов исследования (LR – группа низкого риска, HR – группа высокого риска). Классификатор был обучен на 78 объектах исходных данных (53 объекта класса A, 25 объектов класса B), а затем проверен на остальных 90 объектах (58 объектов класса A, 32 объекта класса B), в результате чего авторы получили следующие результаты классификации: чувствительность – 66 % (21 из 32 пациентов), специфичность – 84 % (49 из 58 пациентов), точность классификации – 78 % (70 из 90 пациентов).

## Описание эксперимента

Перед применением алгоритмов, предложенных в данной работе, выполнялась предобработка данных, разработанная под специфику решаемой задачи. В частности, для данных 330K предварительно применялось логарифмирование по основанию два, и затем медианная нормализация [32]. Также были проведены эксперименты без использования нормализации. Данные Гравье были использованы с авторской методикой предобработки.

Отбор значимых признаков и последующая классификация были выполнены с использованием перекрестной проверки типа «один против всех», т.е. на каждой $i$-й ее итерации тестовая выборка состоит только из данных соответствующих $i$-у пациенту. Помимо вышеизложенного критерия проверки значимости значения ledge-коэффициента для выделения информативных признаков был применен $t$-критерий Стьюдента [34], который является достаточно мощным в случае, когда данные в каждой группе распределены нормально и $U$-критерий Манна – Уитни [35], который используют в случаях, когда распределение данных является произвольным. Признаки отбирались с уровнем значимости 0.05. В таблице 1 представлено, какая в среднем доля от общего числа признаков была отобрана с помощью вышеуказанных методов.

**Таблица 1.** Доля значимых признаков, отобранных с использованием $t$-критерия, $U$-критерия, и ledge-критерия при уровне значимости 0.05

| Отбор признаков | Данные 330K (без нормализации) | Данные 330K (после нормализации) | Данные Гравье |
|---|---|---|---|
| $t$-критерий | 6.11 % | 9.08 % | 22.03 % |
| $U$-критерий | 5.03 % | 8.42 % | 22.69 % |
| ledge-критерий | 11.73 % | 18.62 % | 9.71 % |

Методы выделения информативных признаков $t$-критерий и $U$-критерий подразумевают наличие линейного различия значения признаков между классами. Также надо отметить, что для вышеуказанных методов стоит использовать линейные классификаторы. Исходя из предыдущих работ [32, 36] выявлена эффективность применения метода проекции на латентные структуры (PLS) [37]. Дополнительно используем алгоритмы логистической регрессии (LogisticRegression) [38] и линейного метода опорных векторов (linear-SVM) [39], которые имеют возможность настройки параметров регуляризации, позволяющих модели не адаптироваться к редким представителям выборки (выбросам).

Поскольку связь типа ledge не является линейной, то использование линейных классификаторов может нивелировать эффективность использования метода отбора

информативных признаков, основанного на ledge-коэффициенте корреляции. В дополнение к линейным классификаторам также будем рассматривать и следующие нелинейные: *k*-ближайших соседей (*k*-NN) [40], случайный лес (Random Forest) [41], метод опорных векторов с ядром в виде радиально-базисных функций (rbf-SVM) [42].

## Результаты эксперимента

Для численных расчётов были использованы следующие программные средства: язык программирования Python 3.6, среда разработки Spyder и библиотеки NumPy, SciPy, scikit-learn. Для настройки параметров применялся модуль Hyperopt 0.1.1 [43]. С использованием данной библиотеки осуществляется поиск оптимальных аргументов скалярной функции (функции ошибки на тестовой выборке), при этом выполняется более оптимальный перебор параметров, в отличие от поиска с использованием полного перебора параметров. Hyperopt позволяет более подробно описывать пространство поиска, например, указать распределение параметров (равномерное, нормальное и т.д.).

**Таблица 2.** Качество классификации тестовых образцов данных Гравье в зависимости от методов уменьшения размерности и классификаторов. При отборе признаков использовался фиксированный уровень значимости равный 0.05. Отдельно отмечены наилучшие результаты по каждому методу отбора признаков

| № | Метод | Классификатор | Без отбора | | *t*-критерий | | *U*-критерий | | Ledge-критерий | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* |
| 1 | - | PLS-R | 0.278 | 0.782 | 0.222 | 0.769 | 0.311 | 0.793 | 0.344 | 0.637 |
| 2 | - | linear-SVM | 0.289 | 0.769 | 0.233 | 0.791 | 0.233 | 0.768 | 0.322 | 0.629 |
| 3 | - | rbf-SVM | **0.256** | 0.825 | 0.211 | **0.864** | 0.267 | **0.845** | **0.278** | 0.769 |
| 4 | - | *k*-NN | 0.356 | 0.659 | 0.356 | 0.754 | 0.333 | 0.757 | 0.356 | 0.623 |
| 5 | - | LogisticRegression | 0.322 | 0.725 | 0.267 | 0.778 | 0.267 | 0.781 | 0.344 | 0.673 |
| 6 | - | RandomForest | 0.311 | 0.820 | 0.311 | 0.828 | 0.311 | 0.821 | 0.333 | 0.726 |
| 7 | PCA | PLS-R | 0.267 | 0.790 | 0.333 | 0.767 | 0.311 | 0.793 | 0.322 | 0.675 |
| 8 | PCA | linear-SVM | 0.311 | 0.763 | 0.333 | 0.771 | **0.222** | 0.787 | 0.356 | 0.640 |
| 9 | PCA | rbf-SVM | 0.333 | 0.804 | 0.278 | 0.796 | 0.300 | 0.783 | 0.322 | **0.771** |
| 10 | PCA | *k*-NN | 0.344 | 0.599 | 0.356 | 0.767 | 0.322 | 0.763 | 0.367 | 0.730 |
| 11 | PCA | LogisticRegression | 0.289 | 0.796 | 0.300 | 0.763 | 0.278 | 0.781 | 0.333 | 0.733 |
| 12 | PCA | RandomForest | 0.344 | 0.786 | 0.322 | 0.810 | 0.289 | 0.804 | 0.311 | 0.751 |
| 13 | PLS | PLS-R | 0.278 | 0.782 | 0.300 | 0.823 | 0.278 | 0.820 | 0.344 | 0.637 |
| 14 | PLS | linear-SVM | 0.267 | 0.779 | 0.200 | 0.747 | 0.256 | 0.718 | 0.378 | 0.636 |
| 15 | PLS | rbf-SVM | 0.300 | 0.758 | 0.244 | 0.822 | 0.378 | 0.653 | 0.356 | 0.606 |
| 16 | PLS | *k*-NN | 0.356 | 0.723 | 0.311 | 0.800 | 0.322 | 0.792 | 0.344 | 0.671 |
| 17 | PLS | LogisticRegression | 0.267 | 0.813 | **0.178** | 0.841 | 0.278 | 0.810 | 0.311 | 0.680 |
| 18 | PLS | RandomForest | 0.333 | **0.864** | 0.244 | 0.821 | 0.244 | 0.766 | 0.356 | 0.649 |

В наших расчётах использовалась оптимизируемая функция $F = (1 - AUC)$, где $AUC$ – это площадь под ROC-графиком, вычисленная по результатам пятикратной перекрёстной проверки. После чего итоговая оценка качества вычислялась с использованием найденных оптимальных параметров и перекрёстной проверки «один против всех».

Качество классификации тестовых образцов данных Гравье в зависимости от методов уменьшения размерности и классификаторов показано в таблице 2**Таблица 2**. При этом использовались различные методы отбора признаков при фиксированном уровне значимости равном 0.05.

Из таблицы 2 можем сделать вывод, что в данном эксперименте ledge-критерий сработал хуже (показатель *AUC*) по сравнению с другими методами отбора признаков. Однако качество значительно выше при использовании нелинейных классификаторов (RandomForest и SVM с ядром rbf) по сравнению с линейными. Также качество классификации снижается при использовании метода уменьшения размерности PLS, при котором латентные переменные строятся с учётом линейной разделимости классов (путём максимизации ковариации).

Возможным объяснением полученного результата является то, что при использовании на рассматриваемом наборе данных ledge-критерия с уровнем значимости 0.05 большая доля отобранных признаков была выбрана при учете нелинейной связи с метками классов, и по малой доле признаков с линейной разделимостью линейные классификаторы не способны построить устойчивую дискриминационную модель.

**Таблица 3.** Качество классификации тестовых образцов данных Гравье в зависимости от методов уменьшения размерности и классификаторов. При отборе признаков использовался фиксированный уровень значимости равный 0.1. Отдельно отмечены наилучшие результаты по каждому методу отбора признаков

| № | Метод | Классификатор | Без отбора | | *t*-критерий | | *U*-критерий | | Ledge-критерий | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* |
| 1 | – | PLS-R | 0.278 | 0.782 | 0.244 | 0.737 | 0.278 | 0.752 | 0.322 | 0.737 |
| 2 | – | linear-SVM | 0.344 | 0.769 | 0.267 | 0.796 | **0.244** | 0.781 | 0.333 | 0.727 |
| 3 | – | rbf-SVM | **0.256** | 0.825 | 0.222 | **0.859** | **0.244** | **0.848** | 0.278 | 0.805 |
| 4 | – | k-NN | 0.356 | 0.659 | 0.344 | 0.694 | 0.333 | 0.736 | 0.356 | 0.579 |
| 5 | – | LogisticRegression | 0.322 | 0.725 | 0.278 | 0.774 | 0.300 | 0.769 | 0.300 | 0.755 |
| 6 | – | RandomForest | 0.311 | 0.820 | 0.322 | 0.821 | 0.311 | 0.790 | 0.311 | 0.790 |
| 7 | PCA | PLS-R | 0.278 | 0.800 | 0.278 | 0.833 | 0.256 | 0.796 | 0.256 | 0.756 |
| 8 | PCA | linear-SVM | 0.344 | 0.756 | 0.233 | 0.836 | 0.322 | 0.778 | 0.311 | 0.698 |
| 9 | PCA | rbf-SVM | 0.322 | 0.800 | 0.322 | 0.831 | 0.256 | 0.831 | 0.322 | 0.627 |
| 10 | PCA | k-NN | 0.333 | 0.615 | 0.344 | 0.758 | 0.289 | 0.745 | 0.356 | 0.713 |
| 11 | PCA | LogisticRegression | 0.289 | 0.788 | **0.211** | 0.834 | **0.244** | 0.805 | 0.256 | 0.772 |
| 12 | PCA | RandomForest | 0.344 | 0.779 | 0.267 | 0.807 | 0.278 | 0.822 | 0.322 | 0.725 |
| 13 | PLS | PLS-R | 0.278 | 0.782 | 0.244 | 0.737 | 0.289 | 0.740 | 0.322 | 0.737 |
| 14 | PLS | linear-SVM | 0.267 | 0.779 | 0.278 | 0.765 | 0.311 | 0.729 | 0.367 | 0.704 |
| 15 | PLS | rbf-SVM | 0.300 | 0.756 | 0.278 | 0.781 | 0.289 | 0.761 | 0.322 | **0.810** |
| 16 | PLS | k-NN | 0.356 | 0.723 | 0.311 | 0.825 | 0.333 | 0.805 | 0.333 | 0.743 |
| 17 | PLS | LogisticRegression | 0.267 | 0.814 | 0.233 | 0.841 | **0.244** | 0.825 | **0.233** | 0.753 |
| 18 | PLS | RandomForest | 0.333 | **0.864** | 0.267 | 0.819 | 0.311 | 0.801 | 0.311 | 0.763 |

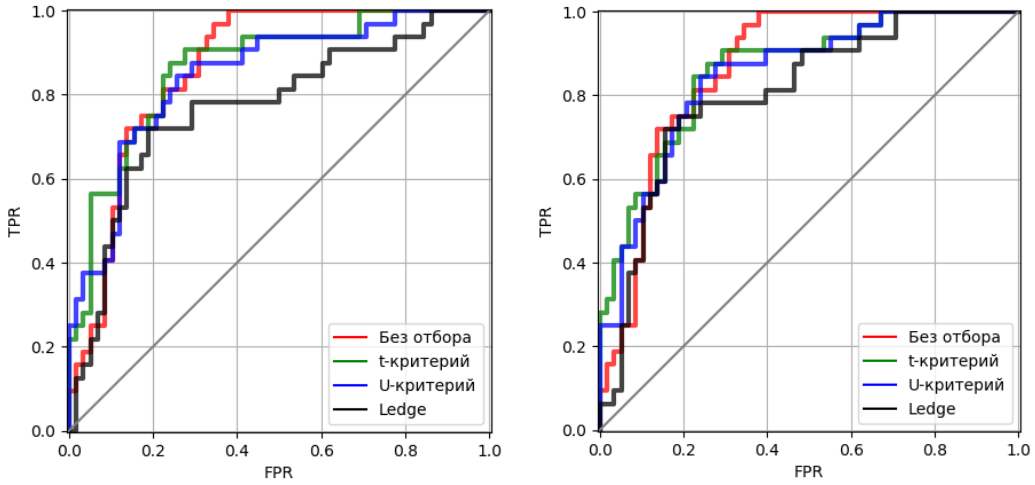
Из таблицы 3 видно, что при повышении уровня значимости до 0.1 показатель качества *AUC,* рассчитанный с использованием метода отбора признаков на основе ledge-коэффициента, повышается с 0.771 до уровня 0.810 при оптимальных настройках алгоритмов классификации. При использовании других методов отбора значимых признаков, показатель *AUC* изменился незначительно: с 0.864 до 0.859 и с 0.845 до 0.848 для *t*-критерия и *U*-критерия соответственно.

При использовании линейных методов отбора признаков (*t*-критерий, *U*-критерий) или без отбора признаков, качество классификации нелинейными методами в среднем выше, чем линейными, независимо от метода уменьшения размерности. Что может свидетельствовать о наличии некоторой «полезной» нелинейной информации в данных.

На рисунке 1 представлены ROC-графики, характеризующие качество классификации тестового множества данных Гравье. Визуализированы только

результаты с наибольшим показателем *AUC* по каждому методу отбора информативных признаков. Графики приведены для отбора признаков с уровнем значимости 0.05 (рис. 1,а) и с уровнем значимости 0.1 (рис. 1,б).

На графиках видно, что при увеличении уровня значимости использование ledge-критерия позволяет получить качество классификации, стремящееся к качеству, полученному при использовании классических методов отбора признаков. По нашему мнению, качество классификации, полученное ledge-критерием, обусловлено тем, что данный метод отбирает меньше информативных признаков, как показано в таблице 1, и можем сделать вывод, что ledge-критерий является менее предназначенным для линейно разделимых данных.




**а)** **б)**

**Рис. 1.** ROC-графики характеризующие классификаторы на данных Гравье: **а)** – отбор признаков с уровнем значимости 0.05, **б)** – отбор признаков с уровнем значимости 0.1. По каждому методу отбора признаку выбран классификатор с максимальным показателем *AUC* из таблицы 2 (**а**) и таблицы 3 (**б**).

В оригинальной работе [33], авторы использовали порог срабатывания классификаторов, полученный в соответствии с критерием Юдена [44], то есть путём максимизации суммы чувствительности и специфичности. При этом авторами были получены следующие качественные показатели: чувствительность – 66 % (21 из 32 пациентов), специфичность – 84 % (49 из 58 пациентов), точность классификации – 78 % (70 из 90 пациентов).

В нашем случае в соответствии с критерием Юдена получены следующие результаты: чувствительность – 91 % (29 из 32 пациентов), специфичность – 72 % (42 из 58 пациентов), точность – 79 % (71 из 90 пациентов). Указанных результатов удалось достичь при отборе значимых признаков с использованием *t*-критерия Стьюдента с уровнем значимости 0.05 и при классификации методом SVM с rbf-ядром. Относительно результатов, полученных в исходной статье, при приблизительно одинаковой точности классификации наблюдается перевес в сторону повышения чувствительности при снижении специфичности.

Рассмотрим так же порог срабатывания классификатора, обеспечивающий тот же уровень ложноположительных решений, что и у Гравье (специфичность 84 %). При этом чувствительность равна 72 % (23 из 32 пациентов), а точность – 80 % (72 из 90 пациентов). Хотя данные результаты не являются оптимальными в смысле критерия Юдена, наблюдается выигрыш в чувствительности (на 6 %) и точности (на 2 %).

Классификатор RandomForest с уменьшением размерности PLS, является оптимальным с точки зрения показателя *AUC*, при этом чувствительность равна 100 %

(32 из 32 пациентов), специфичность – 62 % (36 из 58 пациентов) и точность – 80 % (72 из 90 пациентов).

Применение вышеописанных методов к данным, изложенным в работе Гравье [33] позволило выявить применимость указанных методов к данным биологических микрочипов. Из представленных выше расчётов видно, что для отбора значимых признаков, применение ledge-критерия даёт результаты, сравнимые с классическими критериями, хотя и при малом уровне значимости уступает им в качестве. К тому же применение метода проекции на латентные структуры как алгоритма уменьшения размерности или классификатора позволяет получить качество классификации выше, чем в оригинальной работе, хотя различия в точности классификации не являются значимыми.

Далее рассмотрим применение набора методов к данным 330K изложенным в работе [32]. Данное исследование было проведено с целью уменьшения размерности данных без существенной потери качества классификации. В статье [32] был описан метод классификации на основе алгоритма проекции на латентные структуры, однако, по нашему мнению, существуют методы, которые требуют глубокого анализа и сравнения с методом PLS.

**Таблица 4.** Качество классификации тестовых образцов данных 330K с медианной нормализацией в зависимости от методов уменьшения размерности и классификаторов. При отборе признаков использовался фиксированный уровень значимости равный 0.05. Отдельно отмечены наилучшие результаты по каждому методу отбора признаков

| № | Метод | Классификатор | Без отбора | | *t*-критерий | | *U*-критерий | | Ledge-критерий | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* |
| 1 | – | PLS-R | **0.247** | **0.813** | 0.469 | 0.528 | 0.358 | 0.661 | 0.383 | 0.631 |
| 2 | – | linear-SVM | 0.407 | 0.627 | 0.383 | 0.649 | **0.346** | 0.660 | 0.377 | 0.644 |
| 3 | – | rbf-SVM | 0.432 | 0.543 | 0.401 | 0.640 | 0.407 | 0.636 | 0.451 | 0.600 |
| 4 | – | *k*-NN | 0.463 | 0.553 | 0.463 | 0.552 | 0.420 | 0.574 | 0.543 | 0.454 |
| 5 | – | LogisticRegression | 0.401 | 0.643 | 0.438 | 0.606 | 0.389 | **0.679** | 0.383 | 0.660 |
| 6 | – | RandomForest | 0.469 | 0.516 | 0.475 | 0.559 | 0.488 | 0.531 | 0.469 | 0.519 |
| 7 | PCA | PLS-R | 0.259 | 0.786 | 0.377 | 0.637 | 0.370 | 0.652 | 0.420 | 0.562 |
| 8 | PCA | linear-SVM | 0.488 | 0.508 | 0.395 | 0.666 | 0.420 | 0.635 | 0.432 | 0.591 |
| 9 | PCA | rbf-SVM | 0.531 | 0.319 | 0.586 | 0.384 | 0.525 | 0.501 | 0.432 | 0.555 |
| 10 | PCA | *k*-NN | 0.494 | 0.491 | 0.469 | 0.553 | 0.444 | 0.559 | 0.512 | 0.490 |
| 11 | PCA | LogisticRegression | 0.426 | 0.585 | 0.407 | **0.669** | 0.401 | 0.664 | 0.475 | 0.571 |
| 12 | PCA | RandomForest | 0.512 | 0.456 | 0.401 | 0.639 | 0.463 | 0.561 | 0.457 | 0.515 |
| 13 | PLS | PLS-R | **0.247** | **0.813** | 0.469 | 0.528 | 0.358 | 0.661 | **0.309** | **0.740** |
| 14 | PLS | linear-SVM | 0.290 | 0.753 | 0.500 | 0.540 | 0.352 | 0.671 | 0.315 | 0.717 |
| 15 | PLS | rbf-SVM | 0.494 | 0.516 | 0.469 | 0.491 | 0.506 | 0.528 | 0.383 | 0.600 |
| 16 | PLS | *k*-NN | 0.265 | 0.806 | 0.457 | 0.557 | 0.395 | 0.604 | 0.358 | 0.685 |
| 17 | PLS | LogisticRegression | **0.247** | 0.803 | **0.346** | 0.653 | 0.358 | 0.668 | 0.321 | 0.734 |
| 18 | PLS | RandomForest | 0.278 | 0.770 | 0.500 | 0.510 | 0.432 | 0.569 | 0.389 | 0.688 |

Из таблицы 4 видно, что с использованием медианной нормализации наилучшее качество классификации (*AUC* = 0.813) достигается с использованием метода PLS без отбора значимых признаков. Однако можно заметить, что отбор значимых признаков ledge-критерием хотя и не позволяет достичь качества классификации сравнимого с результатами, полученными без отбора признаков, но даёт лучшие результаты (*AUC* = 0.740) по сравнению с отбором признаков с использованием *t*-критерия (*AUC* = 0.669) и *U*-критерия (*AUC* = 0.679).

Ранее, в работе [32] было показано, что применение классификатора на основе проекции на латентные структуры к данным без использования нормализации позволяет получить более качественные результаты, чем с использованием медианной нормализации. Результаты данного подхода, дополненные методами, описанными выше, представлены в таблице 5.

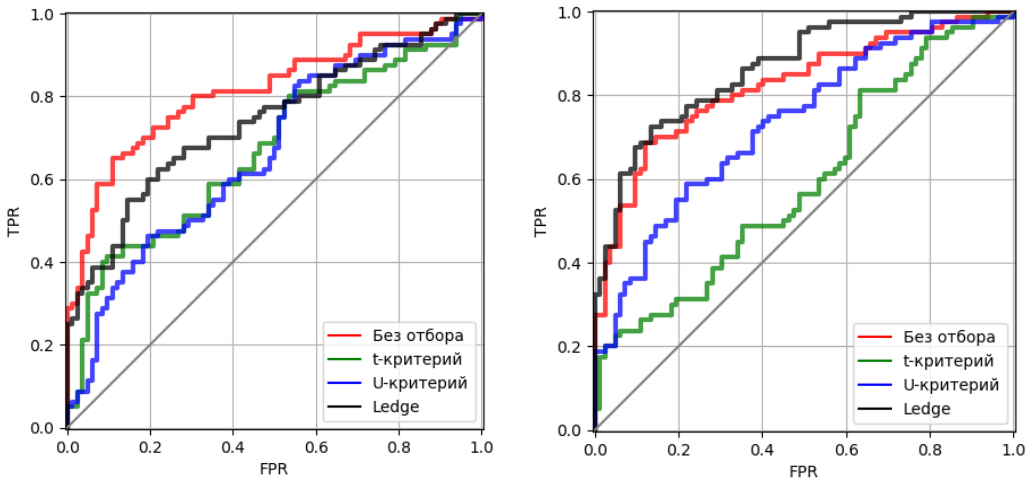
**Таблица 5.** Качество классификации тестовых образцов данных 330K без использования нормализации в зависимости от методов уменьшения размерности и классификаторов. При отборе признаков использовался фиксированный уровень значимости равный 0.05. Отдельно отмечены наилучшие результаты по каждому методу отбора признаков

| № | Метод | Классификатор | Без отбора | | *t*-критерий | | *U*-критерий | | Ledge-критерий | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* | ошибка | *AUC* |
| 1 | – | PLS-R | **0.222** | **0.824** | 0.444 | 0.604 | 0.444 | 0.564 | 0.389 | 0.654 |
| 2 | – | linear-SVM | 0.401 | 0.634 | 0.389 | 0.604 | 0.414 | 0.604 | 0.370 | 0.646 |
| 3 | – | rbf-SVM | 0.469 | 0.441 | 0.475 | 0.517 | 0.467 | 0.506 | 0.500 | 0.465 |
| 4 | – | *k*-NN | 0.519 | 0.339 | 0.457 | 0.482 | 0.519 | 0.415 | 0.685 | 0.265 |
| 5 | – | LogisticRegression | 0.414 | 0.644 | 0.407 | **0.650** | 0.346 | 0.711 | 0.407 | 0.619 |
| 6 | – | RandomForest | 0.605 | 0.378 | 0.549 | 0.478 | 0.506 | 0.483 | 0.593 | 0.393 |
| 7 | PCA | PLS-R | 0.259 | 0.802 | 0.444 | 0.608 | 0.494 | 0.565 | 0.395 | 0.625 |
| 8 | PCA | linear-SVM | 0.426 | 0.568 | 0.389 | 0.643 | 0.407 | 0.614 | 0.383 | 0.630 |
| 9 | PCA | rbf-SVM | 0.556 | 0.271 | 0.430 | 0.584 | 0.414 | 0.568 | 0.420 | 0.555 |
| 10 | PCA | *k*-NN | 0.457 | 0.470 | 0.525 | 0.509 | 0.512 | 0.426 | 0.593 | 0.318 |
| 11 | PCA | LogisticRegression | 0.475 | 0.578 | 0.407 | 0.585 | 0.426 | 0.611 | 0.457 | 0.585 |
| 12 | PCA | RandomForest | 0.463 | 0.542 | 0.463 | 0.525 | 0.426 | 0.562 | 0.463 | 0.515 |
| 13 | PLS | PLS-R | 0.321 | 0.728 | 0.438 | 0.617 | **0.333** | **0.734** | **0.222** | **0.866** |
| 14 | PLS | linear-SVM | 0.241 | 0.810 | 0.444 | 0.611 | 0.463 | 0.558 | 0.432 | 0.611 |
| 15 | PLS | rbf-SVM | 0.469 | 0.526 | 0.401 | 0.577 | 0.444 | 0.586 | 0.426 | 0.621 |
| 16 | PLS | *k*-NN | 0.352 | 0.703 | 0.432 | 0.566 | 0.475 | 0.523 | 0.420 | 0.607 |
| 17 | PLS | LogisticRegression | 0.253 | 0.809 | **0.377** | 0.618 | 0.401 | 0.596 | 0.370 | 0.661 |
| 18 | PLS | RandomForest | 0.284 | 0.800 | 0.414 | 0.595 | 0.432 | 0.579 | 0.457 | 0.606 |

Из таблицы 5 видно, что качество классификации в смысле показателя *AUC* повысилось для всех методов отбора признаков, кроме *t*-критерия. При этом наилучшее качество достигается при отборе признаков ledge-критерием, уменьшении размерности и классификации методом проекции на латентные структуры.

ROC-графики, соответствующие параметрам с максимальным *AUC* для каждого способа отбора признаков, представлены на рисунке 2. На рисунке 2,б видно, что ROC-график, соответствующий классификации с использованием ledge-критерия, находится выше, чем полученные ранее результаты без использования отбора признаков. То есть при любой фиксированной чувствительности алгоритма его специфичность с использованием ledge-критерия не ниже, чем при классификации на полном наборе признаков. При этом качество, полученное в соответствии с критерием Юдена, выражено следующими величинами: чувствительность – 73 % (58 из 80 образцов), специфичность – 87 % (71 из 82 образцов), точность – 80 % (129 из 162 образцов).

Использование *t*-критерия и *U*-критерия с уровнем значимости 0.05 не позволяет добиться результатов близких к тем, что получены с использованием ledge-критерия или без отбора признаков, при заданном выше множестве классификаторов и алгоритмов уменьшения размерности.




**а)** **б)**

**Рис. 2.** ROC-графики характеризующие классификаторы на данных 330K: **а**) отбор признаков с уровнем значимости 0.05, **б**) – отбор признаков с уровнем значимости 0.1. По каждому методу отбора признаку выбран классификатор с максимальным показателем AUC из таблицы 4 (**а**) и таблицы 5 (**б**).

## ЗАКЛЮЧЕНИЕ

Для отбора значимых признаков из данных с микрочипов, обладающих высокой размерностью, наиболее применимы алгоритмы фильтрации в силу высокой скорости работы. Однако, с их помощью не всегда удается получить достаточно высокое качество бинарной классификации. Распространенные в настоящее время методы не вполне сосредоточены на выявлении нелинейной связи между числовым и бинарным признаками, свойственной значимым признакам, разделяющим классы. Поэтому исследование взаимосвязи данного типа содержит в себе потенциал для отбора более полезной информации и, следовательно, улучшения качества бинарной классификации.

Для решения вышеизложенной задачи в данной работе предложен алгоритм отбора значимых признаков из данных с микрочипов, который был назван ledge-критерием. Этот метод представляет собой алгоритм фильтрации, основанный на ранжировании признаков по значению ledge-коэффициента корреляции, который предназначен для оценки силы связи числового и бинарного признаков. Использование данного метода продемонстрировано на двух наборах данных пептидных и ДНК-микрочипов, к которым помимо ledge-критерия применялись два широко распространенных алгоритма фильтрации (*t*-критерий Стьюдента, *U*-критерий Манна – Уитни). Для бинарной классификации использовались шесть известных моделей машинного обучения. В ряде экспериментов применение ledge-критерия отбора значимых признаков позволило получить качество классификации, сравнимое и превосходящее результаты, полученные при использовании традиционных методов.

Перспективным направлением для развития предложенного алгоритма является использование ledge-критерия отбора значимых признаков при создании комплексных методов бинарной классификации данных с микрочипов, применение которых набирает популярность на практике в настоящее время. Речь идет о гибридных алгоритмах, использующих ансамбль фильтров для более эффективного отбора относительно небольшого количества значимых признаков, к которым затем применяется алгоритм обертки. Использование ансамбля фильтров позволяет улучшить эффективность модели и получать более стабильные результаты на различных наборах данных, потому что отбор выполняется на основе нескольких критериев. Таким образом исключается ситуация полного влияния одного алгоритма на выбор значимых признаков. Использование комбинации разнообразных фильтров предоставляет возможность для

того, чтобы получать больше полезной информации на основе всесторонней оценки признаков с помощью различных соответствующих критериев отбора. В настоящее время часто в ансамбли фильтров включают широко распространенные алгоритмы. Добавление к ним ledge-критерия содержит в себе потенциал для улучшения эффективности модели за счет выявления признаков, имеющих статистически значимую нелинейную связь с целевым бинарным признаком, свойственную биомаркерам. Рассмотренный выше подход позволяет использовать преимущества как алгоритмов фильтрации, так и алгоритмов обертки. Таким образом, формирование комплексной гибридной модели вышеизложенным образом потенциально способно улучшить качество бинарной классификации данных с микрочипов.

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 17-04-00321.

## СПИСОК ЛИТЕРАТУРЫ

1. Renard B.Y., Löwer M., Kühne Y., Reimer U., Rothermel A., Türeci O., Castle J.C., Sahin U. Rapmad: Robust analysis of peptide microarray data. *BMC Bioinformatics*. 2011. V. 12. doi: 10.1186/1471-2105-12-324.
2. Önskog J., Freyhult E., Landfors M., Rydén P., Hvidsten T.R. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinformatics*. 2011. V. 12. doi: 10.1186/1471-2105-12-390.
3. Mohammed A., Biegert G., Adamec J., Helikar T. CancerDiscover: An integrative pipeline for cancer biomarker and cancer class prediction from high-throughput sequencing data. *Oncotarget*. 2018. V. 9. № 2. P. 2565–2573. doi: 10.18632/oncotarget.23511.
4. Alanni R., Hou J., Azzawi H., Xiang Y. A novel gene selection algorithm for cancer classification using microarray datasets. *BMC Med Genomics*. 2019. V. 12. doi: 10.1186/s12920-018-0447-6.
5. Xi M., Sun J., Liu L., Fan F., Wu X. Cancer Feature Selection and Classification Using a Binary Quantum-Behaved Particle Swarm Optimization and Support Vector Machine. *Computational and Mathematical Methods in Medicine*. 2016. V. 2016. P. 1–9. doi: 10.1155/2016/3572705.
6. Hira Z., Gillies D. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*. 2015. V. 2015. P. 1-13. doi: 10.1155/2015/198363.
7. Saeys Y., Inza I., Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007. V. 23. № 19. P. 2507-2517. doi: 10.1093/bioinformatics/btm344.
8. Lazar C., Taminau J., Meganck S., Steenhoff D., Coletta A., Molter C., de Schaetzen V., Duque R., Bersini H., Nowe A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2012. V. 9. № 4. P. 1106-1119. doi: 10.1109/TCBB.2012.33.
9. Jafari P., Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*. 2006. V. 6. doi: 10.1186/1472-6947-6-27.
10. Nguyen T., Khosravi A., Creighton D., Nahavandi S. Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification. *PLoS ONE*. 2015. V. 10. № 3. doi: 10.1371/journal.pone.0120364.
11. Shahjaman M., Rahman M., Islam S., Mollah M. A Robust Approach for Identification of Cancer Biomarkers and Candidate Drugs. *Medicina*. 2019. V. 55. № 6. doi: 10.3390/medicina55060269.

12. Maniruzzaman M., Rahman J., Ahammed B., Abedin M., Suri H., Biswas M., El-Baz A., Bangeas P., Tsoulfas G., Suri J. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Computer Methods and Programs in Biomedicine*. 2019. V. 176. P. 173-193. doi: 10.1016/j.cmpb.2019.04.008.
13. Momenzadeh M., Sehhati M., Rabbani H. A novel feature selection method for microarray data classification based on hidden Markov model. *Journal of Biomedical Informatics*. 2019. V. 95. doi: 10.1016/j.jbi.2019.103213.
14. Boareto M., Caticha N. t-Test at the Probe Level: An Alternative Method to Identify Statistically Significant Genes for Microarray Data. *Microarrays.* 2014. V. 3. № 4. P. 340-351. doi: 10.3390/microarrays3040340.
15. Fox R., Dimmic M. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics.* 2006. V. 7. doi: 10.1186/1471-2105-7-126.
16. Shukla A., Tripathi D. Identification of potential biomarkers on microarray data using distributed gene selection approach. *Mathematical Biosciences.* 2019. V. 315. doi: 10.1016/j.mbs.2019.108230.
17. Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A., Benitez J., Herrera F. A review of microarray datasets and applied feature selection methods. *Information Sciences.* 2014. V. 282. P. 111–135. doi: 10.1016/j.ins.2014.05.042.
18. Aboudi N., Benhlima L. Review on wrapper feature selection approaches. In: *2016 International Conference on Engineering & MIS (ICEMIS).* IEEE, 2016. P. 1–5. doi: 10.1109/ICEMIS.2016.7745366.
19. Sanz H., Valim C., Vegas E., Oller J., Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics.* 2018. V. 19. doi: 10.1186/s12859-018-2451-4.
20. Li Z., Xie W., Liu T. Efficient feature selection and classification for microarray data. *PLoS ONE*. 2018. V. 13. № 8. doi: 10.1371/journal.pone.0202167.
21. Anaissi A., Kennedy P., Goyal M. Feature selection of imbalanced gene expression microarray data. In: *2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).* IEEE, 2011. P. 73–78. doi: 10.1109/SNPD.2011.12.
22. Kang C., Huo Y., Xin L., Tian B., Yu B. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of theoretical biology*. 2019. V. 463. P. 77–91. doi: 10.1016/j.jtbi.2018.12.010.
23. Chuang L., Yang C., Wu K., Yang C. A hybrid feature selection method for DNA microarray data. *Computers in Biology and Medicine*. 2011. V. 41. № 4. P. 228–237. doi: 10.1016/j.compbiomed.2011.02.004.
24. Huijuan L., Junying C., Ke Y., Qun J., Yu X., Zhigang G. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing.* 2017. V. 256. P. 56–62. doi: 10.1016/j.neucom.2016.07.080.
25. Shukla A., Singh P., Vardhan V. A hybrid gene selection method for microarray recognition. *Biocybernetics and Biomedical Engineering*. 2018. V. 38. № 4. P. 975–991. doi: 10.1016/j.bbe.2018.08.004.
26. Sun Y., Lu C., Li X. The Cross-Entropy Based Multi-Filter Ensemble Method for Gene Selection. *Genes*. 2018. V. 9. № 5. doi: 10.3390/genes9050258.
27. Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition*. 2012. V. 45. № 1. P. 531–539. doi: 10.1016/j.patcog.2011.06.006.
28. Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. Data classification using an ensemble of filters. *Neurocomputing*. 2014. V. 135. P. 13–20. doi: 10.1016/j.neucom.2013.03.067.

29. Strimbu K., Tavel J.A. What are Biomarkers? *Current Opinion in HIV and AIDS*. 2010. V. 192. № 3. P. 214–216. doi: 10.1097/COH.0b013e32833ed177.
30. Дронов С.В., Петухова Р.В. Один вид связи между номинальной и бинарной переменными. *Известия АлтГУ*. 2010. Т. 65. № 1/2. С. 34–36.
31. Дронов С.В., Бойко И.Ю. Метод оценки степени связи бинарного и номинального показателей. *Прикладная дискретная математика*. 2015. Т. 30. № 4. С. 109–119. doi: 10.17223/20710410/30/11.
32. Анисимов Д.С., Подлесных С.В., Колосова Е.А., Щербаков Д.Н., Петрова В.Д., Джонстон С.А., Лазарев А.Ф., Оскорбин Н.М., Шаповал А.И., Рязанов М.А. Анализ многомерных данных пептидных микрочипов с использованием метода проекции на латентные структуры. *Математическая биология и биоинформатика*. 2017. Т. 12. № 2. С. 435–445. doi: 10.17537/2017.12.435.
33. Gravier E. A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*. 2010. V. 49. № 12. P. 1125–1134. doi: 10.1002/gcc.20820.
34. Student. The probable error of a mean. *Biometrika*. 1908. V. 6. № 1. P. 1–25. doi: 10.2307/2331554.
35. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*. 1947. № 18. P. 50–60. doi: 10.1214/aoms/1177730491.
36. Анисимов Д.С., Рязанов М.А., Шаповал А.И. Применение метода проекции на латентные структуры в задачах классификации на примере данных пептидных микрочипов. В: *Сборник трудов всероссийской конференции по математике "МАК-2016" (Барнаул, 29 июня – 1 июля 2016 г.).* Барнаул: Изд-во АлтГУ, 2016. С. 92.
37. Эсбенсен К. *Анализ многомерных данных. Избранные главы*. Барнаул: Изд-во Алт. ун-та. 2003. 157 с.
38. Cox D.R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society*. 1958. V. 20. № 2. P. 215–242. doi: 10.1111/j.2517-6161.1958.tb00292.x.
39. Вапник В.Н. *Восстановление зависимостей по эмпирическим данным*. М.: Наука, 1979. 448 с.
40. Cover T.M., Hart P.E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967. V. 13. № 1. P. 21–27. doi: 10.1109/TIT.1967.1053964.
41. Breiman L. Random Forests. *Machine Learning*. 2001. 45. № 1. P. 5–32.
42. Boser B.E., Guyon I.M., Vapnik V.N. A Training Algorithm for Optimal Margin Classifiers. In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory – COLT'92* (Pittsburgh. 27–29 July 1992). New York, 1992. P. 144–152. doi: 10.1145/130385.130401.
43. *Hyperopt: Distributed Asynchronous Hyper-parameter Optimization*. URL: https://github.com/hyperopt/hyperopt (дата обращения: 16.04.2019).
44. Youden W.J. Index for rating diagnostic tests. *Cancer*. 1950. V. 3. № 1. P. 32–35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Рукопись поступила в редакцию 18.07.2019. Переработанный вариант поступил 15.01.2020.
Дата опубликования 30.01.2020.

# Approach to The Selection of Significant Features in Solving Biomedical Problems of Binary Classification of Microarray Data

**Boyko I.Y., Anisimov D.S., Smolyakova L.L., Ryazanov M.A.**

*Altay State University, Barnaul, Russian Federation*

***Abstract.*** In modern biomedical research aimed at finding methods for early diagnosis of cancer, microarrays containing certain biological information about patients are used. Based on these data, patients are assigned to one of two classes, corresponding to the presence and absence of some diagnosis. When solving this problem, one of the steps that have a decisive influence on the quality of classification is the significant features selection. This paper proposes a criterion for the selection of significant features, based on the ledge-coefficient of correlation. The ledge-coefficient was previously used to estimate the degree of interrelation of numerical and binary features. For two sets of microarray data, comparative examples of their binary classification are presented using three feature selection algorithms, three dimensionality reduction methods, six classification models. The use of the ledge-criterion for feature selection made it possible to obtain a classification quality comparable to the results of using common methods of feature selection, such as *t*-test and *U*-test. For the data set of the peptide microarrays considered in the paper, the effectiveness of applying the projection method to latent structures had previously been identified. The use of this method in combination with the significant features' selection using the ledge-criterion made it possible to obtain a higher classification quality measure.

***Key words:*** *feature selection, ledge-coefficient, binary classification, microarrays, ROC-curve, projection to latent structures.*