ОТБОР ХАРАКТЕРИСТИК РАСПРЕДЕЛЕНИЯ ИНТЕНСИВНОСТИ В ЦВЕТОВЫХ КАНАЛАХ НА ЛЮМИНЕСЦЕНТНЫХ ИЗОБРАЖЕНИЯХ РАКОВЫХ КЛЕТОК

Е. В. Лисица *, В. В. Скакун, В. В. Апанасович

УДК 51-76;615.471;004.89

Белорусский государственный университет,
220030, Минск, просп. Независимости, 4, Беларусь; e-mail: lisitsa@bsu.by

(Поступила 16 ноября 2018)

Рассмотрены методы (корреляционный, логистической регрессии, медианный и случайного леса) для отбора информативных характеристик распределения интенсивности флуоресценции ядер на многоканальных люминесцентных изображениях раковых клеток. Исходные данные представляют собой трехканальные изображения, зарегистрированные в цветовой системе RGB (Red-Green-Blue). Рассмотрено 39 стандартных характеристик распределений по 13 на каждый цветовой канал. Использование шести признаков дает точность классификации, сопоставимую с точностью при 39 признаках. При незначительном изменении точности классификации (на 0.005) достаточно всего два признака. Предлагается использовать приведенные данные иммуногистохимического анализа биомаркеров в клетках рака молочной железы при анализе люминесцентных изображений при обработке результатов в онкоцитологии.

Ключевые слова: *описательная статистика, люминесцентная микроскопия, отбор признаков, раковые клетки, онкоцитология.*

The different methods (correlation, logistic regression, median and random forest methods) for the selection of informative characteristics of the intensity distribution of fluorescent nuclei on multichannel luminescent images of cancer cells are considered. The input data are the three-channel RGB images. In total, 39 standard characteristics of distributions are studied, including 13 characteristics per each color channel. It is established that the use of 6 features permits to achieve the same classification accuracy as for using 39 features. Moreover, one can use only two features with an insignificant increase in the classification accuracy (by 0.005). It is proposed to use the data of the immunohistochemical analysis of biomarkers in breast cancer cells during the analysis of luminescent images when processing the results in oncocytology.

Keywords: *descriptive statistics, fluorescence microscopy, cancer cells, feature selection, oncocytology.*

Введение. Особенность онкологических заболеваний — их большая вариабельность от пациента к пациенту, поскольку болезнь начинается на клеточном уровне, а структура клетки уникальна для каждого человека. Для анализа внутриклеточных молекулярных процессов широкое применение получили методы люминесцентной микроскопии [1]. Результаты таких исследований могут дать возможность для индивидуальной целенаправленной лекарственной терапии.

Интенсивность люминесценции красителя отражает активность определенных процессов, происходящих в клетке [2]. В каждом изображении обычно находится несколько сотен объектов, при этом ядро может характеризоваться несколькими десятками параметров свечения, которые можно рассматривать как его признаки [3]. Такие объемы данных в силу больших вычислительных затрат не позволяют в клинической практике широко использовать методы машинного обучения. Анализ одного изображения может достигать нескольких часов. Отбор наиболее информативных признаков — один из широко распространенных этапов предварительной обработки данных в алгоритмах машинного обучения. Методы отбора информативных признаков позволяют снизить размерность исходных данных за счет ранжирования признаков по степени их важности. Это позволяет исключить из анали-

---

SELECTION OF INTENSITY DISTRIBUTION CHARACTERISTICS IN THE COLOR CHANNELS OF FLUORESCENT IMAGES OF CANCER CELLS

Y. U. Lisitsa *, V. V. Skakun, V. V. Apanasovich (*Belarusian State University, 4 Nezavisimosti Prosp., Minsk, 220030, Belarus; e-mail: lisitsa@bsu.by*)

за неинформативные признаки [4], что приведет к сокращению временных затрат и снижению избыточности данных, повысив точность классификации.

На сегодняшний день существует большое количество методов отбора признаков, многие из которых успешно применяются в диагностике раковых заболеваний. Среди них выделяется метод инкрементной выборки (IFS, incremental feature selection), который используется для сокращения количества признаков в задаче классификации онкологических заболеваний по типам [5]. При диагностике рака молочной железы применяют метод обратно-фазового белкового чипа (reversep hase protein array) для отбора биомаркеров рака [6]. При прогнозировании ранней стадии рака в плоскоклеточной карциноме головы и шеи с помощью протеомных и транскриптомических данных используются такие методы снижения размерности, как методы минимальной избыточности и метод перестановок [7]. Для оценки качества отбора информативных признаков часто применяется дисперсионный анализ и критерий Стьюдента [8]. В работах [9, 10] отбор информативных признаков проводился на основе интеграции пяти подходов отбора на основе фильтров (критерий Стьюдента, ROC-кривые, медианный метод, метод на основе энтропии и соотношения сигнал/шум) и посредством метода анализа иерархии (AHP, analytic hierarchy process). Также существуют методы для отбора информативных признаков на основе скрытых марковских моделей [11], метод *k*-ближайших соседей [12], опорных векторов [13], случайного леса (RF, random forest) [14]. Однако использование методов отбора признаков не всегда является надежным и устойчивым, например, при наличии большого количества признаков или недостатка объема выборки для отбора информативных признаков. Поэтому на практике для отбора признаков применяются несколько методов, что позволяет ранжировать признаки и одновременно учитывать специфические недостатки каждого метода.

Цель настоящей работы — выявить информативные характеристики интенсивности на люминесцентных изображениях раковых клеток с помощью методов отбора информативных признаков. Шесть различных методов отбора признаков использованы для исследования 39 характеристик распределения интенсивности в ядрах по многоканальным изображениям раковых клеток. Далее ядра называем объектами.

Материалы и методы. Рассмотрены девять случайно отобранных микрочипов из 187 микрочипов срезов тканей опухолей молочной железы [15]. Экспертным путем установлены 6366 ядер на изображениях. Цель эксперимента — количественный анализ гетерогенности эстроген-рецептора при раке молочной железы [16].

Предварительно изготовленные парафинизированные препараты ткани подвергались депарафинированию и извлечению антигенов путем варки под давлением. Препараты инкубировались с 0.3 % бычьего сывороточного альбумина в 0.1 М трис-буфере (BSA/TBS) в течение 30 мин при комнатной температуре. Далее препараты инкубировались с первичными антителами, разведенными в BSA/TBS, в течение 1 ч при комнатной температуре или в течение ночи при +4 °С, троекратно промывались в течение 5 мин раствором BSA/TBS, содержащим 0.05 % Tween-20. Соответствующие вторичные антитела, растворенные в BSA/TBS, добавлялись на 1 ч при комнатной температуре. Они представляли собой антитела, конъюгированные с флуорофором (Amersham, Piscataway, и Molecular Probes, Eugene, США) и/или конъюгированные декстрановым остовом, несущим пероксидазу хрена (HRP) (Envision, DAKO, Carpinteria, США). Вместе со вторичными антителами в растворе присутствовал краситель 4,6-диамидино-2-фенилиндол дигидрохлорид (DAPI) для визуализации ядер. Затем срезы повторно подвергались троекратной отмывке BSA/TBS, содержащим 0.05 % Tween-20. Для автоматического анализа препараты инкубировались с флуоресцентным хромагеном (цианин-5-тирамид, NEN LifeScience, Products, США), в результате чего молекулы цианина ковалентно сшивались за счет активности HRP и накапливались в непосредственной близости от мест связывания меченых вторичных антитител. Препараты, предназначенные для автоматического анализа, покрывались средой, предотвращающей выцветание (гельватол с 0.6 % *н*-пропилгаллатом). Изображения микрочипов получены с помощью платформы Deltavision и программного обеспечения Soft Worx 2.5 (Applied Precision, США) с использованием камеры с водяным охлаждением Photometrics серии 300 и линз ×10 Nikon Super-Fluor на флуоресцентном микроскопе Nikon TE 200.

Изображения представляют собой популяции клеток, маркированные тремя красителями и сохраненные в RGB-формате. В противоположность здоровым клеткам в цитоплазме раковых клеток появляется белок цитокератин [16]. Белок маркируется цианиновым красителем Cy3 и регистрируется в зеленом цветовом канале изображения. Для маркировки всех ядер (ДНК) использован краситель DAPI [16] и зарезервирован синий канал; красный канал изображения — для индикации ядер рако-

вых клеток. Краситель Cy5 [16, 17] использован для маркировки белка эстроген-рецептора [18], который находится в ядрах раковых клеток. Соответственно, маркерами раковых клеток являются два красителя — Cy5 и Cy3. Размер изображений 2048×2048 пикселей в каждом из трех каналов, разрешающая способность 0.2 мкм/пиксель, или 5 мкм [19]. Экспериментальная часть исследований выполнена по методикам [16, 17], полный протокол проведения эксперимента — в [16].

В качестве характеристик распределения интенсивности люминесценции в цветовых каналах *R*, *G*, *B* использованы мода (*Mo*), медиана (*Me*) и среднее значение (*M*). Эти характеристики широко применяются в описательной статистике [20]. Для того чтобы оценить величину, на которую характеристики отдельных объектов отличаются от их центральной тенденции, рассчитаны максимальное (max) и минимальное (min) значения, нижний ($p25$) и верхний ($p75$) квантили, межквартильный интервал (*pd*), дисперсия (*D*) и стандартное отклонение (*std*). Стандартная ошибка (*se*) характеризует стандартное отклонение выборочного среднего, коэффициент эксцесса (*kur*) — остроту пика в распределении интенсивности, коэффициент асимметрии (*skew*) — асимметрию распределения интенсивности. Коэффициент вариации (*var*) рассчитывается как отношение *std*/*M* [20]. Таким образом, для описания одного объекта использованы 39 характеристик распределения интенсивности в трех цветовых каналах.

Поскольку некоторые методы отбора информативных признаков чувствительны к наличию выбросов данных, для устранения этого к исходным данным может быть применена нормировка:

$$\overline{p}_i = (p_i - M(p))/std(p),$$

где p_i — признак i-го объекта.

В настоящей работе использованы следующие методы отбора признаков [4].

Медианный. Сравниваются объекты, относящиеся к двум классам, на основе *U*-критерия Манна—Уитни: чем меньше критерий, тем больше информативный признак. Для удобства делается нормировка, которая ранжирует важность признаков по убыванию.

Корреляционный. Из группы признаков, коррелирующих между собой, остается только тот, у которого наибольшее значение корреляции с признаком, отвечающим за принадлежность к классу; важность остальных признаков становится равной нулю; расчет корреляции может проводиться по Пирсону или Спирмену.

Метод логистической регрессии. Весовые коэффициенты характеризуют важность признаков.

Метод случайного леса RF (четыре модификации). В основе метода лежит использование ансамбля решающих деревьев. Параметры каждого дерева задаются случайным образом, например, для каждого дерева произвольно задается набор признаков из общего множества признаков. Деревья решений по их назначению можно разделить на две группы: CART, в основе которых заложен CART-алгоритм, и CF-деревья, использующие условное деление — специальную непараметрическую проверку для разделения дерева. В зависимости от способа выбора признаков при обучении различают такие виды деревьев, как метод на основе изменения ошибки (ER, error rate), которые сравнивают изменение ошибки при использовании различных наборов для обучения деревьев. Вторая группа методов использует критерий Джинни (Gini), третья группа рассчитывает площадь под ROC-кривой (receiver operating characteristic — рабочая характеристика приемника) по площади под кривой AUC (area under ROC). В настоящей работе рассмотрены только CART-деревья с ER- и Gini-критериями.

Рассмотрены восемь методов отбора информативных признаков: медианный (Median), корреляционный с расчетом корреляции по Пирсону (*P_cor*), корреляционный с расчетом корреляции по Спирмену (*S_cor*), метод логистической регрессии (*LR*), *ER_RF* — CART-деревья на основе изменения ошибки, Gini_*RF* — CART-деревья с Джинни-критерием, *ER_CF* — CF-деревья на основе изменения ошибки, *AUC_CF* — CF-деревья на основе изменения ошибки с оценкой ошибки по AUC.

Проанализированы девять изображений. Получены 6366 объектов, для описания которых использовано 39 характеристик интенсивности. Для корреляционных методов установлено пороговое значение 0.8, количество итераций для обучения каждого дерева 100. Взяты два набора данных — нормированный и ненормированный. Для того чтобы оценить качество отбора признаков, необходимо изучить, как изменяется ошибка классификации при использовании некоторого стандартного метода, например RF, по мере увеличения количества признаков. Для оценки качества ранжирования признаков в качестве эталонного алгоритма рассмотрен случайный выбор признаков *R* (random), когда для заданного количества признаков проводится их случайный выбор по равномерному закону распреления из общего набора признаков для описания объектов. Качество кластеризации изучено с помощью алгоритма RF с критерием Джинни [21], количество деревьев 10.

Результаты и их обсуждение. Получим оценки важности признаков на ненормированном и нормированном наборах данных. На рис. 1 показаны оценки важности признаков, полученные различными методами отбора на ненормированном наборе данных. Наихудшие результаты показал медианный метод отбора признаков. В этом случае важность всех признаков одинакова — 0.167, кроме *skewR* (важность 0.158), который и являлся самым неинформативным. Таким образом, медианный фильтр — неэффективный метод при большом наборе признаков и не рекомендуется к применению.

Метод *P_cor* отобрал 14 характеристик цвета: *seB* — 0.167, *p*75*G* — 0.158, max*R* — 0.150, min*R* — 0.141, *MoB* — 0.140, *varR* — 0.129, min*B* — 0.091, *skewB* — 0.083, *skewG* — 0.081, *varB* — 0.071, *kurR* — 0.056, *varG* — 0.046, *DG* — 0.043, *kurB* — 0.033. Первая группа (*seB*, *p*75*G*, max*R*, min*R*, *MoB*, *varR*) содержит три характеристики красного канала, две — зеленого и один — синего. Вторая группа обладает значительно меньшей важностью, чем первая.

Метод *S_cor* отобрал 11 характеристик: *std R* — 0.167, *p*75*G* — 0.165, *seB* — 0.158, min*R* — 0.149, *p*25*B* — 0.120, *skewG* — 0.093, *skewB* — 0.074, *varG* — 0.065, *kurR* — 0.060, *varB* — 0.058, *kurB* — 0.038. Первая группа содержит четыре характеристики, из которых две относятся к красному каналу, и по одной к синему и зеленому. Вторая группа обладает меньшей важностью по сравнению с первой.

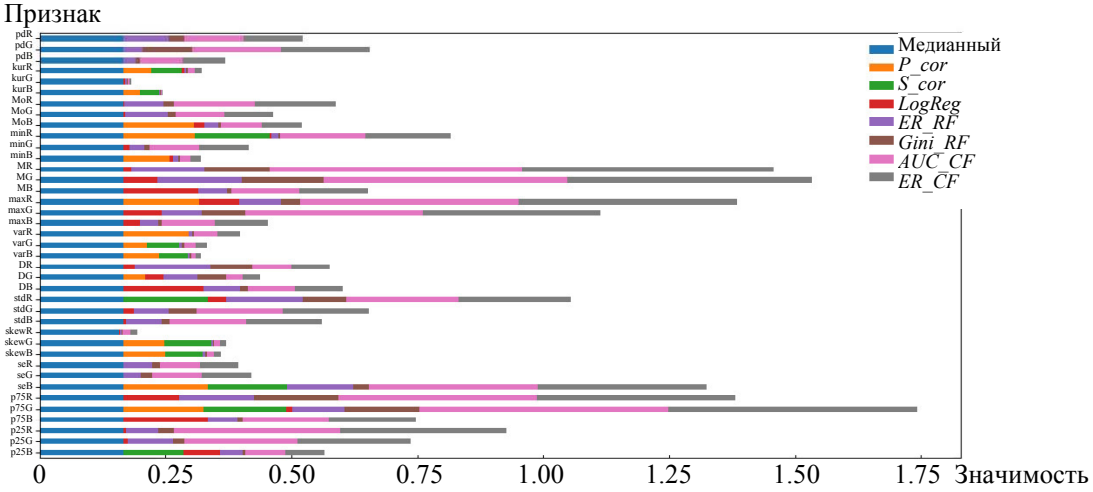

Рис. 1. Важность признаков цвета на ненормированном наборе данных

Метод *LR* установил важность для 36 характеристик, которая варьируется от 0.167 до $2.14 \cdot 10^{-6}$. Десять наиболее значимых из них: *p*75*B* — 0.167, *DB* — 0.159, *MB* — 0.149, *p*75*R* — 0.110, max*R* — 0.080, max*G* — 0.075, *p*25*B* — 0.072, *MG* — 0.067, *stdR* — 0.038, *DG* — 0.035. В первую группу входят три характеристики распределения интенсивности по синему каналу (*p*75*B*, *DB*, *MB*). Интересно, что этот краситель отвечает за окрашивание ядер как здоровых, так и раковых клеток. Это свидетельствует о том, что накопление красителя происходит по-разному в ядрах здоровых и больных клеток. Вторая группа признаков (*p*75*R*, max*R*, max*G*, *p*25*B*, *MG*) включает в себя две характеристики по красному и зеленым каналам и одну синего цвета.

Метод *ER_RF* позволяет определить для каждого признака его важность. Десять наиболее информативных признаков: *MG* — 0.167, *stdR* — 0.152, *DR* — 0.151, *p*75*R* — 0.150, *MR* — 0.144, *seB* — 0.131, *p*75*G* — 0.103, *p*25*G* — 0.089, *pdR* — 0.089, *MoG* — 0.084. Необходимо отметить, что важность первого признака *MG* значительно выше, чем остальных, несмотря на то что этот признак используется для описания интенсивности только в зеленом канале (отвечает за окрашивание цитоплазмы раковых клеток). Это показывает, что данный краситель накапливается как в цитоплазме, так и в ядрах здоровых и раковых клеток, но с разными законами распределения. Вторая группа (*stdR*, *DR*, *p*75*R*, *MR*, *seB*) содержит только признаки по красной компоненте, которые изначально используются для маркировки раковых клеток. Далее в значимости признаков не происходит существенных изменений.

Аналогично методу *ER_RF* метод *Gini-RF* для каждого признака определяет его важность, однако в этом методе она убывает быстрее, чем в *ER_RF*. Десять самых информативных признаков, отобранных методом: *p*75*R* — 0.167, *MG* — 0.163, *p*75*G* — 0.150, *MR* — 0.131, *pdG* — 0.100, *stdR* — 0.086, max*G* — 0.086, *DR* — 0.083, *DG* — 0.056, *stdG* — 0.054. В первую группу попали два признака

(*p*75*R* и *MG*), при этом их значимости сопоставимы. Признак *MG* получил почти такую же значимость, как и в методе *ER_RF*. Для второй группы (*p*75*G*, *MR*, *pdG*) характерно изменение значимости признаков на 0.014—0.030. Метод *Gini-RF* поставил низкие значимости признакам синего канала. Самый высоких у них — *seB*, ему соответствует значение 0.032.

Метод *AUC_CF* аналогично другим методам на основе случайного леса определил важность каждого признака. Десять наиболее информативных признаков, отобранных этим методом: *MR* — 0.500, *p*75*G* — 0.494, *MG* — 0.485, max*R* — 0.434, *p*75*R* — 0.394, max*G* — 0.353, *seB* — 0.335, *p*25*R* — 0.330, *p*25*G* — 0.225, *stdR* — 0.223. Резкое изменение значимости признаков начинается с *p*25*G*. Согласно этому методу, признаки по зеленой и красной компонентам имеют сопоставимые значимости.

Метод *ER_CF* показал значения, схожие с *AUC_CF*. Максимальное расхождение ≤3 %.

Как следует из полученных результатов, ранжирование разными методами различается значительно. Поэтому для оценки качества работы методов построены оценки временных затрат и ошибки классификации по мере увеличения наборов признаков согласно ранжированию их исследуемыми методами (рис. 2). Как и ожидалось, время обучения случайного леса зависит от количества признаков, используемых для описания объектов (рис. 2, *а*). Для случайного леса из 10 деревьев можно выделить шесть групп признаков по размерности: до двух, 3—7, 8—14, 15—23, 24—34, 35 и более. Если для каждого признака посчитать количество раз, которые он был отобран в первой десятке по значимости, то получится: 4 — *seB*, *stdR*, *MG*, *p*75*G*, max*R*; 3 — *p*25*B*, *MR*, min*R*, max*G*, *p*75*R*; 2 — *MB*, *MoB*, *skewG*, min*B*, *skewB*, *p*25*G*, *varB*; 1 — *p*75*B*, *DB*, *DR*, *varR*, *varG*, *p*25*R*, min*G*, *kurR*, *pdR*, *DG*, *MoG*. В результате только пять признаков отобраны большинством методов как наиболее значимые. Такое разбиение на группы по размерности обусловлено не качеством отбора признаков, а особенностью метода случайного леса, так как при построении обучающей выборки для одного решающего дерева ее размер равен корню от общего количества признаков, используемых для описания.

В табл. 1 показана суммарная важность признаков по всем методам кроме медианного (не информативен) и *ER_CF* (практически совпадает с *AUC_CF*). Согласно суммарной значимости, наиболее информативны признаки *p*75*G* и *MG*. Интересно, что оба признака являются оценками распределения интенсивности красителя цитоплазмы в ядрах. Вторая группа признаков по значимости: *p*75*R*, *seB*, *stdR*, *MR*, max*R*. В этой группе находятся преимущественно признаки, отвечающие за распределение интенсивности онкомаркера в ядрах. Меньшая информативность этих признаков по сравнению с признаками, посчитанными по зеленой компоненте, возможно, обусловлена тем, что красный краситель встречается только в 70 % раковых клеток, в то время как зеленый используется для 100 % раковых клеток. Из полученных результатов можно сделать вывод, что часть красителя цитоплазмы накапливается в ядрах.

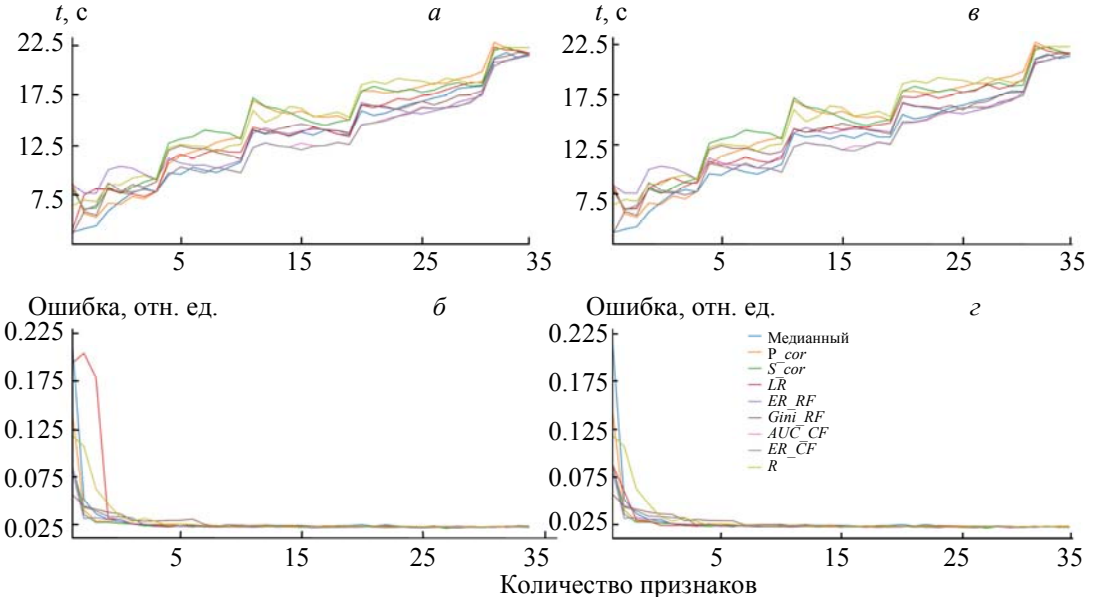

Рис. 2. Оценка временных затрат (*а*, *в*) и точности (*б*, *г*) классификации по признакам, отобранным методами на ненормированном (*а*, *б*) и нормированном наборах данных (*в*, *г*)

Т а б л и ц а 1. Важность признаков цвета на ненормированном наборе данных

Признак	Важность	Признак	Важность	Признак	Важность
p75G	1.249	*DB*	0.507	max*B*	0.348
MG	1.048	*DR*	0.500	*skewB*	0.346
seB	0.989	*p25B*	0.488	*seG*	0.322
p75R	0.987	*stdG*	0.482	*seR*	0.318
MR	0.957	*pdG*	0.480	min*G*	0.317
max*R*	0.951	*MoB*	0.441	*varB*	0.311
stdR	0.832	*MoR*	0.427	*varG*	0.310
max*G*	0.761	*stdB*	0.409	*kurR*	0.308
min*R*	0.647	*pdR*	0.405	min*B*	0.299
p25R	0.596	*DG*	0.404	*pdB*	0.284
p75B	0.574	*MoG*	0.367	*kurB*	0.243
MB	0.516	*skewG*	0.358	*skewR*	0.180
p25G	0.512	*varR*	0.352	*kurG*	0.178

Согласно изменению ошибки, при обучении случайного леса имеет место резкий скачок в изменении ошибки для четырех методов (медианный, *P_cor*, *S_cor*, *ER_RF*) при использовании двух признаков по сравнению с одним признаком. Для пяти методов (*ER_RF*, *S_cor*, медианный, *LR*, *P_cor*) ошибка перестает резко изменяться при шести и более признаках. Для метода *LR* характерно увеличение ошибки при малом количестве признаков: когда признаков больше восьми, его результаты сопоставимы с другими методами (рис. 2, *б*). При 2—5 признаках, отобранных методом *ER_RF*, почти не изменяется качество классификации случайным лесом. Для описания объектов достаточно шести признаков вместо 39, в этом случае ошибка классификации 0.025 соответствует ошибке классификации при использовании всего набора признаков. В случае двух признаков (*p75R* и *MG*) самая малая ошибка (0.03) у метода *ER_RF*.

Проведем отбор наиболее информативных признаков после нормировки набора данных. При работе с нормированным набором данных все методы кроме *LR* (результаты приведены ниже) показали схожие результаты в отборе признаков (рис. 3). Временные затраты и ошибка классификации аналогичны полученным при работе с ненормированным набором данных.

Метод *LR* из 39 информативных признаков отобрал 36, при этом их важность отличается на порядки (от 0.167 для *stdR* до 0.00028 для *varG*). Десять наиболее информативных признаков согласно этому методу: *stdR* — 0.167, *p75R* — 0.143, *MB* — 0.133, *p75B* — 0.129, *MG* — 0.129, *stdG* — 0.091, *p25B* — 0.073, max*R* — 0.065, max*G* — 0.053, min*G* — 0.043. У 21 признака значимость превосходит 0.0167. В отличие от результатов, полученных на ненормированном наборе данных, здесь важны признаки, которые рассчитаны в красном канале.

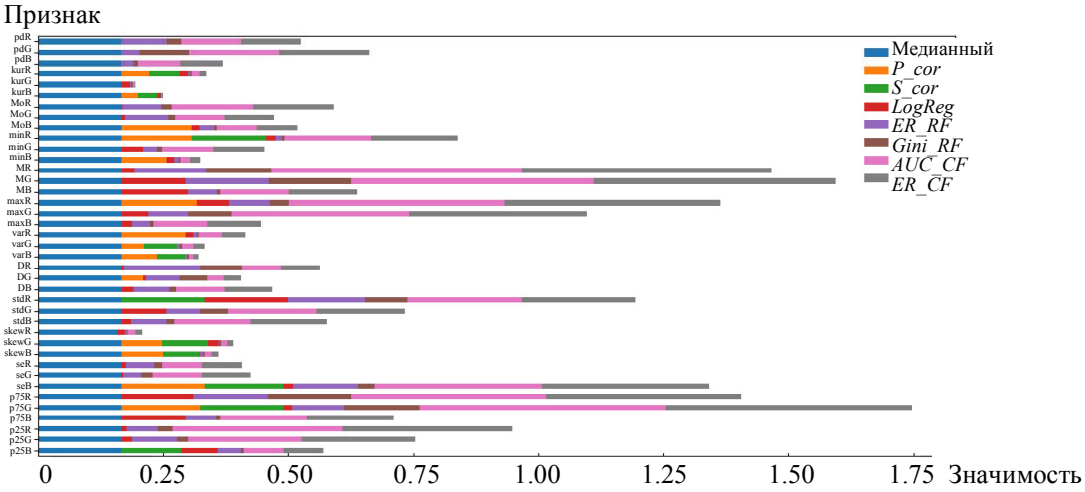

Рис. 3. Важность признаков цвета на нормированном наборе данных

Заключение. Показано, что для описания всех объектов достаточно шести найденных наиболее информативных признаков, при этом сохраняется точность классификации 0.025. При использовании всего двух наиболее информативных признаков (верхний квантиль интенсивности в красном канале и математическое ожидание в зеленом) ошибка классификации 0.03, что достаточно для уверенной классификации объектов на изображениях. Временные затраты в результате уменьшения количества признаков для описания объектов сокращены в три раза.

Из шести рассмотренных методов отбора информативных признаков наилучшие результаты показал метод случайного леса с CART-деревьями на основе изменения ошибки, наихудшие получены для медианного метода, а метод логистической регрессии оказался наименее устойчивым к выбросам в обучающей выборке. Временные затраты на обучение случайного леса ступенчато возрастают по мере увеличения количества признаков для описания объектов.

Отобранные признаки на практике можно применять для описания ядер на люминесцентных изображениях в задачах классификации и кластеризации без использования всего набора признаков распределений интенсивности красителей в ядрах и цитоплазме с сохранением точности классификации. Разработанный методический подход позволяет исследовать популяцию гетерогенных клеток, различающихся выраженностью того или иного признака, а также ранжировать признаки по степени значимости, что может найти применение в научных исследованиях в области цитологии.

[1] **B. Stewart, C. P. Wild.** World Cancer Report 2014, Geneva, Switzerland, World Health Organization, International Agency for Research on Cancer, WHO Press (2015) 16—81
[2] **V. Mikkilineni, R. D. Mitra, J. Merritt, J. R. DiTonno, G. M. Church, B. Ogunnaike, J. S. Edwards.** Biotechn. Bioengin., **86**, N 2 (2004) 117—124
[3] **O. Ronneberger, D. Baddeley, F. Scheipl, P. J. Verveer, H. Burkhardt, C. Cremer, L. Fahrmeir, T. Cremer, B. Joffe.** Chromosome Res., **16**, N 3 (2008) 523—562
[4] **U. Neumann, N. Genze, D. Heider.** BioData Mining, **10** (2017) 10:21
[5] **P.-W. Zhang, L. Chen, T. Huang, N. Zhang, X.-Y. Kong, Y.-D. Cai.** PloS One, **10**, N 3 (2015) e0123147; doi: 10.1371/journal.pone.0123147
[6] **J. Sonntag, C. Bender, Z. Soons, S. von der Heyde, R. Konig, S. Wiemann, H.-P. Sinn, A. Schneeweis, T. Beibarth, U. Korf.** Translat. Proteom., **2** (2014) 52—59
[7] **D. K. Chanchala, D. W. May.** IEEE J. Biomed. Health Inform., **21**, N 1 (2017) 246—253
[8] **P. Stafford, Z. Cichacz, N. W. Woodbury, S. A. Johnston.** Proc. Nat. Acad. Sci., **111**, N 30 (2014) E3072—E3080
[9] **T. Nguyen, S. Nahavandi.** IEEE Transact. Fuzzy System., **24**, N 2 (2016) 273—287
[10] **T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi.** PloS One, **10**, N 3 (2015) e0120364
[11] **S. R. Eddy.** Current Opinion in Struct. Biol., **6**, N 3 (1996) 361—365
[12] **G. Ritter, H. Woodruff, S. Lowry, T. Isenhour.** IEEE Transact. Inform. Theory, **21**, N 6 (1975) 665—669
[13] **M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, B. Scholkopf.** IEEE Intell. System. Appl., **13**, N 4 (1998) 18—28
[14] **W. Pang, H. Jiang, S. Li.** Biomed. Res. Int. (**2017**) 1—14
[15] **H. M. Kluger, R. Halaban, D. L. Rimm.** Cancer Res., **64**, N 23 (2004) 8767—8772
[16] **G. G. Chung, M. P. Zerkowski, S. Ghosh, R. L. Camp, D. L. Rimm.** Lab. Invest., **87**, N 7 (2007) 662—669
[17] **R. L. Camp, G. G. Chung, D. L. Rimm.** Nature Med., **8**, N 11 (2002) 1323—1327
[18] **M. K. Szeszel, C. L. Crisman, L. Crow, S. McMullen, J. M. Major, L. Natarajan, A. Saquib, J. R. Feramisco, L. M. Wasserman. J. Histochem.** Cytochem., **53**, N 6 (2005) 753—762
[19] **Е. В. Лисица, Н. Н. Яцков, В. В. Апанасович, Т. В. Апанасович, М. М. Шитик.** Журн. прикл. спектр., **81**, № 6 (2014) 907—913 [**Y. V. Lisitsa, M. M. Yatskou, V. V. Apanasovich, T. V. Apanasovich, M. M. Shytsik.** J. Appl. Spectr., **81** (2014) 996—1003]
[20] **T. G. Nick.** Methods Mol. Biol., **404** (2007) 33—52
[21] **F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay.** J. Mach. Learn. Res., **12** (2011) 2825—2830