ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

ОБРАБОТКА ИЗОБРАЖЕНИЙ И РАСПОЗНАВАНИЕ ОБРАЗОВ

IMAGE PROCESSING AND PATTERN RECOGNITION

УДК 303.732.4

Поступила в редакцию 20.09.2018
Received 20.09.2018

Принята к публикации 15.11.2018
Accepted 15.11.2018

Отбор информативных геометрических признаков ядер клеток на люминесцентных изображениях раковых клеток

Е. В. Лисица✉, Н. Н. Яцков, В. В. Скакун, П. Д. Кривошеев, В. В. Апанасович

Белорусский государственный университет, Минск, Беларусь
✉E-mail: ylisitsa@gmail.com

Аннотация. Рассмотрены методы отбора информативных признаков для выделения геометрических признаков при описании ядер на люминесцентных изображениях раковых клеток. Выполнен обзор существующих геометрических признаков, который включает в себя как признаки формы, устойчивые к повороту и перемещению изображения, так и признаки расположения в пространстве. Для отбора наиболее информативных признаков использованы шесть методов: медианный, корреляционный с расчетом коэффициента корреляции по Пирсону, корреляционный с расчетом коэффициента корреляции по Спирмену, метод логистической регрессии, случайного леса с CART-деревьями и критерием Gini, случайного леса с CART-деревьями и критерием минимизации ошибки. В результате исследования из 59 признаков отобраны 11 наиболее информативных, выполнен анализ качества классификации с помощью метода случайного леса и рассчитаны временные затраты в зависимости от количества признаков для описания объектов. Для метода случайного леса использование 11 признаков является достаточным по точности классификации и позволяет снизить временные затраты в 2,3 раза.

Ключевые слова: корреляция, случайный лес, логическая регрессия, медианный метод, классификация

Для цитирования. Отбор информативных геометрических признаков ядер клеток на люминесцентных изображениях раковых клеток / Е. В. Лисица [и др.] // Информатика. – 2019. – Т. 16, № 2. – С. 7–17.

Selection of geometrical features of nuclei on fluorescent images of cancer cells

Yauheniya U. Lisitsa✉, Mikalai M. Yatskou, Victor V. Skakun, Pavel D. Kryvasheyeu, Vladimir V. Apanasovich

Belarusian State University, Minsk, Belarus
✉E-mail: ylisitsa@gmail.com

Abstract. The methods of geometric informative features selection of nuclei on fluorescent images of cancer cells are considered. During the survey, a review of existing geometric features was carried out, including both the signs of rotation resisted shape and displacement of the image, as well as signs of location in space. For the selection of characteristics, the methods were used: median, correlation with calculation of the Pearson correlation coefficient, correlation with calculation of the Spearman correlation coefficient, logistic regression model, random forest with CART trees and Gini criterion, random forest with CART trees and error minimization criterion. As a result of the investigation 11 characteristics were selected from 59 features, the quality of classification and time costs were calculated depending on the number of features for describing

the objects. The use of 11 features is sufficient for the accuracy of classification as it allows to reduce time costs in 2,3 times.

Key words: correlation, random forest, logistic regression, median, classification

For citation. Lisitsa Y. U., Yatskou M. M., Skakun V. V., Kryvasheyeu P. D., Apanasovich V. V. Selection of geometrical features of nuclei on fluorescent images of cancer cells. *Informatics*, 2019, vol. 16, no. 2, pp. 7–17 (in Russian).

Введение. Метод люминесцентной микроскопии получил широкое распространение в исследованиях, связанных с диагностикой раковых заболеваний [1]. Современные программные средства [2] используют автоматические методы сегментации изображений, что позволяет выявлять новые признаки у объектов, такие как эффективный радиус, центральные моменты, Hu-моменты и др., которые невозможно получить при визуальном анализе [3]. Однако использование чрезмерно большого количества признаков усложняет работу с данными. Одним из способов уменьшения вычислительных затрат, но при этом без потерь их информативности или с незначительными потерями, является отбор информативных признаков [4].

Методы отбора информативных признаков получили широкое распространение в предварительном анализе данных в медико-биологических исследованиях. В работе [5] рассмотрен mRMR-метод для отбора 23 белков из 187, которые затем использовались для классификации десяти типов рака [5]. Три различных метода: метод опорных векторов, случайного леса и классификатор на основе ближайшего центроида – использовались для отбора 10 информативных биомаркеров из 128 при классификации гормонального рецептор-положительного рака молочной железы [6]. Для отбора информативных признаков при классификации головной и шейной плоскоклеточной карцином были рассмотрены U-критерий Манна – Уитни, mRMR- и wrapper-методы [7]. Дисперсионный анализ и тест Стьюдента позволили отобрать информативные признаки для классификации раковых заболеваний на основе иммуноподписи [8]. Методы на основе критерия Стьюдента, ROC-кривой, энтропии, соотношения сигнал-шум и U-критерий Манна – Уитни использовались для отбора информативных признаков при диагностике лимфомы, рака простаты и лейкемии [9, 10]. Методы случайного леса показали достаточно точные результаты ранжирования признаков при медико-биологических исследованиях [11–15]. Их применение позволяет увеличить скорость обработки данных, снизить ошибку классификации и еще больше автоматизировать процесс обработки. Наиболее распространенными алгоритмами отбора информативных признаков являются методы, основанные на корреляции, логистической регрессии и случайном лесе. Какой именно из этих методов будет использоваться, зависит от решаемой задачи.

Целью настоящей работы является отбор информативных геометрических признаков ядер раковых и нераковых клеток на люминесцентных изображениях рака груди. Для достижения поставленной цели необходимо решить ряд задач: выполнить обзор существующих геометрических признаков для описания объектов; реализовать ранжирование признаков по мере важности различными методами отбора информативных признаков; оценить ошибку классификации по количеству признаков в наборе для описания объектов, на основе которой отобрать информативные признаки; определить наиболее устойчивый метод отбора информативных признаков.

Экспериментальные данные. В работе рассматриваются девять микрочипов срезов тканей опухолей молочной железы [16]. Экспериментальная часть исследований проведена по методикам [17, 18]. Изображения представляют собой популяции клеток, окрашенные в зеленые, синие и красные цвета (трехканальные люминесцентные сигналы в системе RGB). В цитоплазмах раковых клеток регистрируются процессы с участием белка цитокератина [17]. Белок маркируется цианиновым красителем Cy3 и регистрируется в зеленом цветовом канале изображения. Красный канал изображения зарезервирован для индикации ядер раковых клеток.

В ядрах раковых клеток находится белок эстроген-рецептор [19], для маркировки которого использован краситель [17, 18]. Для маркировки ядер применен краситель 4,6-диамидино-2-фенилиндол дигидрохлорид (DAPI) [17] и зарезервирован синий канал. Размер изображений – 2048×2048 пикселов в каждом из трех каналов, разрешающая способность 0,2 мкм/пиксел [20].

Перед сравнительным анализом алгоритмов клетки были размечены на раковые и нераковые экспертным путем. На рис. 1 показаны изображения, взятые для анализа.

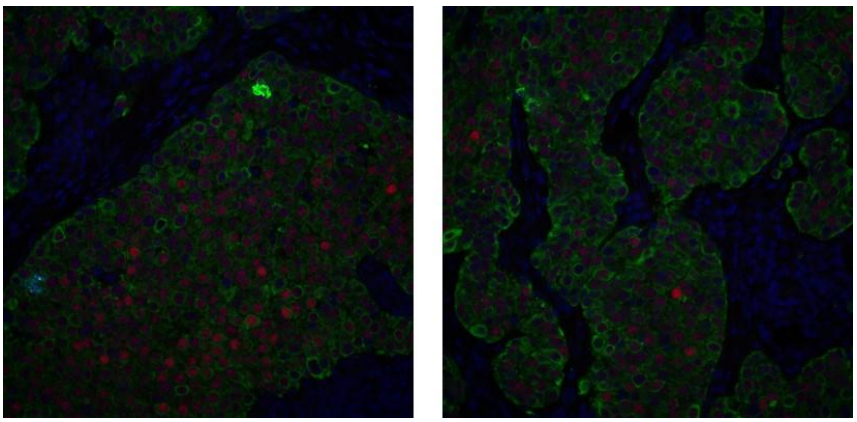


Рис. 1. Экспериментальные данные

Геометрические признаки объектов. Всего было рассмотрено 59 признаков [21–23]:

– площадь ($p1$) – количество пикселов, составляющих объект;

– минимальный ($p2$, $p3$) и максимальный ($p4$, $p5$) номера строк и столбца прямоугольника, ограничивающего области выбора;

– площадь ограничивающего прямоугольника ($p6$);

– координаты центра масс ($p7$ и $p8$);

– площадь выпуклой оболочки, описанной вокруг объекта ($p9$);

– эксцентриситет эллипса ($p10$), который имеет такие же центральные моменты инерции, что и объект;

– эквивалентный диаметр ($p11$) – диаметр круга с той же площадью, что и область выбора;

– экстент $p12 = p1/p6$;

– площадь объекта после заполнения его пустот ($p13$);

– матрица тензора инерции ($p14$, $p15$, $p16$, $p17$) и собственные значения матрицы тензора инерции ($p18$, $p19$);

– координаты центра масс ограничивающего прямоугольника ($p20$, $p21$);

– длины большой ($p22$) и малой ($p23$) осей эллипса, который имеет такие же нормированные вторые центральные моменты, что и объект;

– пространственные моменты до третьего порядка μ_{jk} ($p24 = \mu_{00}$, $p25 = \mu_{01}$, $p26 = \mu_{02}$, $p27 = \mu_{10}$, $p28 = \mu_{11}$, $p29 = \mu_{12}$, $p30 = \mu_{20}$, $p31 = \mu_{21}$, $p32 = \mu_{22}$), которые вычисляются как

$$\mu_{jk} = \sum_{(x,y)\in S} x^j y^k ,$$

где $j = 0, 1, 2$ и $k = 0, 1, 2$;

– центральные моменты $m\mu_{jk}$ ($p33 = m\mu_{00}$, $p34 = m\mu_{01}$, $p35 = m\mu_{02}$, $p36 = m\mu_{10}$, $p37 = m\mu_{11}$, $p38 = m\mu_{12}$, $p39 = m\mu_{20}$, $p40 = m\mu_{21}$, $p41 = m\mu_{22}$), рассчитанные по формуле

$$m\mu_{jk} = \sum_{(x,y)\in S} (x - p7)^j (y - p8)^k ,$$

где $j = 0, 1, 2$ и $k = 0, 1, 2$;

– Hu-моментные инварианты [24]

$$p42 = m\mu_{20} + m\mu_{02},$$

$$p43 = (m\mu_{20} - m\mu_{02})^2 + 4m\mu_{11}^2,$$

$$p45 = (m\mu_{30} + m\mu_{12})^2 + (m\mu_{30} + m\mu_{21})^2,$$

$$p46 = (m\mu_{30} - 3m\mu_{12})(m\mu_{30} + m\mu_{12})[(m\mu_{30} + m\mu_{12})^2 - 3(m\mu_{21} + m\mu_{03})^2] +$$
$$+(3m\mu_{21} - m\mu_{03})(m\mu_{03} + m\mu_{21})[3(m\mu_{12} + m\mu_{30})^2 - (m\mu_{30} + m\mu_{12})^2,$$

$$p47 = (m\mu_{20} - m\mu_{02})[(m\mu_{30} + m\mu_{12})^2 - (m\mu_{03} + m\mu_{21})^2] +$$
$$+4m\mu_{11}\ (m\mu_{30} + m\mu_{12})(m\mu_{03} + m\mu_{21}),$$

$$p48 = (3m\mu_{21} + m\mu_{03})(m\mu_{30} + m\mu_{12})[(m\mu_{12} + m\mu_{30})^2 - 3(m\mu_{03} + m\mu_{21})^2] -$$
$$-(m\mu_{30} - 3m\mu_{12})(m\mu_{03} + m\mu_{21})[3(m\mu_{30} + m\mu_{12})^2 - (m\mu_{21} + m\mu_{03})^2];$$

– нормированные центральные моменты $n\mu_{jk}$ ($p49 = n\mu_{01}$, $p50 = n\mu_{02}$, $p51 = n\mu_{10}$, $p52 = n\mu_{11}$, $p53 = n\mu_{12}$, $p54 = n\mu_{20}$, $p55 = n\mu_{21}$, $p56 = n\mu_{22}$), рассчитанные по формуле

$$n\mu_{jk} = m\mu_{jk} / \mu_{00}^{\frac{j+k}{2}+1},$$

где $j = 0, 1, 2$ и $k = 0, 1, 2$;

– ориентация $p57$, представляющая собой угол между осью X и главной осью эллипса, который имеет те же самые моменты, что и объект, и диапазон от $-pi/2$ до $pi/2$ в направлении против часовой стрелки;

– периметр $p58$;

– коэффициент плотности $p59 = p1/p9$.

Алгоритмы отбора информативных признаков

Для отбора информативных признаков был рассмотрен ряд методов [25]:

1. Медианный (Median), в основе которого лежит U-критерий Манна – Уитни [26]. Нулевая гипотеза H_0 заключается в следующем: если med_0 и med_1 – медианы двух выборок, то H_0: $\text{med}_0 = \text{med}_1$. Полученные p-значения используются в качестве уровня важности признаков. Следовательно, чем меньше p, тем более информативен признак. Для того чтобы ранжировать объекты по мере убывания их важности, применяется специальная нормировка [27].

2. Корреляционный. Корреляция между любыми двумя признаками может быть описана как количественная оценка степени статистической линейной зависимости между ними, которая может быть определена различными коэффициентами корреляции. В основе отбора лежит исключение тех признаков, которые имеют высокий коэффициент корреляции с исследуемым признаком, но при этом показывают низкую корреляцию при сравнении с другими признаками [27–30]. В работе рассмотрены два критерия: Пирсона, который используется для работы со случайными величинами с нормальным законом распределения, и Спирмена, который устойчив к наличию выбросов в данных и может применяться для случайных величин с различными законами распределения. Расчет корреляции по Пирсону и Спирмену задается формулами

$$r_{xy} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \sum_{j=1}^{n}(y_j - \bar{y})}},$$

$$\rho = 1 - 6\sum_{j=1}^{n} \frac{rk(x_j) - rk(y_j)}{n(n^2 - 1)},$$

где $X = (x_1, \ldots, x_n)$ и $Y = (y_1, \ldots, y_n)$ – наборы двух признаков, rk – ранговый набор.

3. Логистической регрессии (LR, logistic regression). В этом методе весовые β-коэффициенты логистической модели описывают важность признаков, однако для их использования необходимо предварительно выполнить нормировку признаков [31].

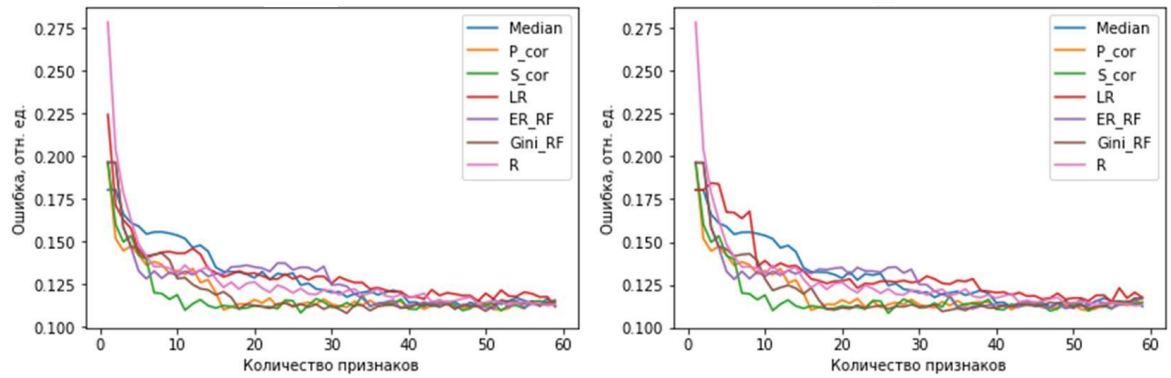
4. Случайного леса (RF, random forest) – метод машинного обучения, который представляет собой ансамбль деревьев. Прогноз получается как результат объединения ответов множества деревьев. При этом деревья обучаются независимо друг от друга и на разных подмножествах данных. В работе были рассмотрены только CART-деревья [32] с двумя критериями обучения: Gini [31] и ER (error-rate, скорость изменения ошибки) [25].

Оценка качества ранжирования признаков. Стандартным методом проверки качества ранжирования и отбора признаков является исследование изменения ошибки классификации (*er*) известным классификатором на основе выборок, упорядоченных методом отбора информативных признаков [33, 34]. Эта ошибка рассчитывается как доля неверно классифицированных объектов (*E*) от общего количества объектов (*T*), участвовавших в классификации: $er = E/T$.

Предварительная экспертная оценка деления клеток на раковые и нераковые позволяет вычислять ошибку классификации с тем или иным набором информативных признаков.

Описание эксперимента. Для оценки качества ранжирования признаков в роли эталонного алгоритма выступает случайный выбор признаков R (от англ. random), когда для заданного количества признаков производился их случайный выбор по равномерному закону распределения из общего набора признаков для описания объектов. Такая процедура повторялась 10 раз. Для метода P_cor коэффициент корреляции был установлен 0,9, для метода S_cor – 0,8. Количество деревьев для Gini_RF (CART-деревьев с критерием обучения Gini) и ER_RF (CART-деревьев с критерием обучения ER) было установлено равным 30. Поскольку некоторые методы, учитывающие расчет расстояний между объектами, являются чувствительными к выбросам, исследование проводилось на ненормированном и нормированном наборах данных. Для кластеризации использовался алгоритм случайного леса с критерием Джинни [35], количество деревьев равнялось 100. В качестве критерия было взято значение ошибки классификации. Исследования выполнялись с использованием библиотек scikit-learn и Pandas языка Python.

Результаты. На рис. 2 показаны результаты исследования методов отбора информативных признаков на ненормированных и нормированных данных. Как следует из полученных результатов, нормировка признаков не дает значимого преимущества при их ранжировании.



а) *б)*

Рис. 2. Изменение ошибки классификации в зависимости от количества признаков в ранжированном наборе: *а)* ненормированные данные; *б)* нормированные данные

Согласно полученным результатам устойчивыми к наличию выбросов являются все методы, кроме Gini_RF и LR. Ошибка классификации уменьшается быстрее на нормированном наборе данных для метода Gini_RF, в то время как для метода LR ошибка классификации убывает быстрее на ненормированном наборе. Худшие результаты показали три метода отбора информативных признаков:

1) для метода LR ошибка всегда больше, чем для R, за исключением случая, когда для описания используется только один признак;

2) ошибка классификации в методе Median убывает медленнее, чем в методе случайного выбора признаков R, в результате чего применение ранжирования по методу Median становится рациональным только в случае, когда для описания используется около 40 признаков;

3) метод ER_RF показывает удовлетворительную ошибку классификации в случае использования 30–40 информативных признаков, упорядоченных по степени важности согласно этому методу.

При использовании признаков, ранжированных методами P_cor и Gini_RF, результаты классификации сопоставимы. Ошибка классификации резко изменяется с 0,12 до 0,11, когда размер выборки признаков, ранжированных по методу P_cor, равняется 17, после этого значительных изменений не происходит. Для метода Gini_RF резкое изменение происходит чуть позже, когда количество признаков достигает значения 19, затем аналогично предыдущему методу ошибка классификации почти не изменяется и колеблется около значения 0,11.

Лучшие результаты были получены при ранжировании признаков методом S_cor. В этом случае на графике видно резкое изменение ошибки классификации до 0,110 по мере увеличения размера выборки до 11, затем ошибка классификации медленно возрастает до значения 0,112, вокруг которого в дальнейшем по мере увеличения размера обучающей выборки она и варьируется. Метод S_cor является нечувствительным к нормировке, и его можно применять для данных с выбросами.

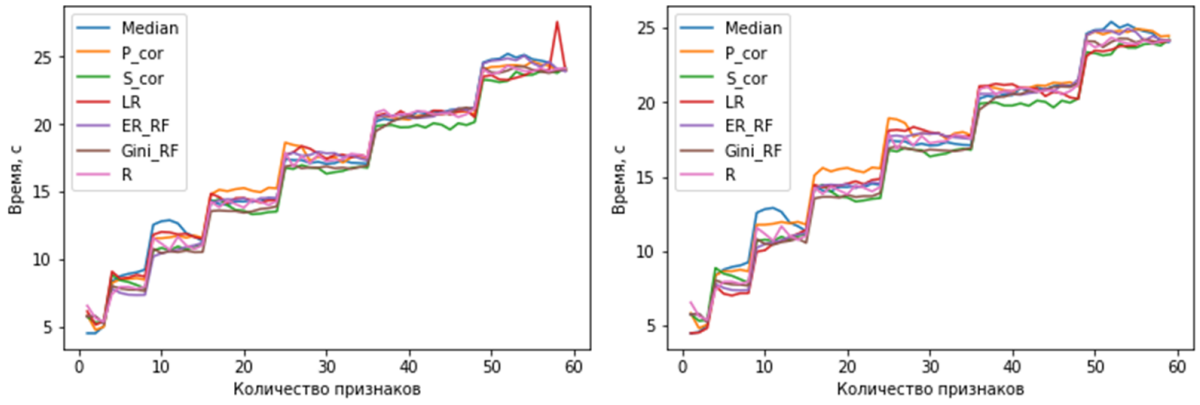
Таким образом, оптимальным набором для описания исследуемых изображений является набор из 11 признаков с их уровнями важности: $p19$ – 0,167, $p20$ – 0,131, $p21$ – 0,121, $p18$ – 0,109, $p42$ – 0,099, $p56$ – 0,087, $p59$ – 0,072, $p44$ – 0,067, $p46$ – 0,043, $p54$ – 0,034, $p5$ – 0,025. Использование этих 11 признаков позволяет получить наименьшую ошибку классификации. Из них только три признака были отобраны и другими методами: $p19$ был отобран методами ER_RF и Gini_RF, $p20$ – методом P_cor, а $p21$ – методом LR. Результаты сравнения уровней важности для отобранных признаков приведены в таблице.

Сравнение уровней важности для отобранных признаков шестью методами отбора значимых признаков (N – номер признака при ранжировании выбранным методом, V – уровень важности при ранжировании)

Признак	Median		P_cor		S_cor		LR		ER_RF		Gini_RF	
	N	*V*	*N*	*V*	*N*	*V*	*N*	*V*	*N*	*V*	*N*	*V*
*p*19	35	0,167	44	0,000	1	0,167	30	4E-10	1	0,167	2	0,163
*p*20	38	0,167	3	0,760	2	0,131	19	5E-07	29	0,021	39	0,021
*p*21	37	0,167	4	0,677	3	0,121	7	3E-05	28	0,023	49	0,017
*p*18	39	0,167	43	0,000	4	0,109	32	2E-10	24	0,025	44	0,019
*p*42	13	0,167	9	0,520	5	0,099	48	1E-16	21	0,032	26	0,029
*p*56	20	0,167	10	0,397	6	0,087	56	0E+00	31	0,020	28	0,028
*p*59	40	0,167	16	0,089	7	0,072	59	0E+00	30	0,021	12	0,054
*p*44	15	0,167	21	0,033	8	0,067	40	3E-14	32	0,019	14	0,041
*p*46	17	0,167	35	0,000	9	0,043	38	6E-14	40	0,012	27	0,028
*p*54	19	0,167	11	0,327	10	0,034	54	0E+00	39	0,012	51	0,017
*p*5	27	0,167	15	0,143	11	0,025	15	1E-06	36	0,016	20	0,035

Для классификации использовался метод случайного леса с критерием Джинни [34], количество деревьев равнялось 100. Как показано на рис. 3, общей особенностью зависимости времени обучения классификации методом случайного леса от количества признаков объектов является ступенчатое возрастание времени обучения по мере увеличения количества признаков. Резкие скачки изменения времени происходят при увеличении количества признаков для описания объектов до 4 (с 5,18 до 8,09 с), до 9 (с 8,00 до 11,10 с), до 16 (с 11,22 до 14,31 с), до 25 (с 14,46 до 17,69 с), до 36 (с 17,36 до 20,45 с), до 49 (с 20,97 до 24,04 с). Общей закономерностью поведения времени обучения внутри этих интервалов является его колебание вокруг

одного значения. Размеры интервалов, когда время обучения не изменяется, увеличиваются по мере возрастания количества признаков в наборе для описания объектов. Такое поведение временных затрат обусловлено способом построения деревьев в методе случайного леса: количество признаков, которое используется для обучения одного дерева, определяется как целая часть квадратного корня от общего количества признаков для описания объектов. Наименьшие временные затраты на обучение были при отборе признаков методом Gini_RF, в то время как наибольшие временные затраты наблюдались при отборе признаков методом P_cor. При использовании 11 признаков, отобранных методом S_cor, время обучения составляло 10,7 с, а время обучения случайного леса при использовании всех признаков составляло 24,2 с, что почти в два раза больше.



а) *б)*

Рис. 3. Зависимость времени обучения классификатора на основе метода случайного леса от количества признаков: *а)* ненормированные данные; *б)* нормированные данные

Заключение. В результате выполненной работы из 59 геометрических признаков формы были отобраны 11 наиболее информативных признаков (ошибка классификации 0,110) для описания ядер на люминесцентных изображениях раковых клеток с сохранением точности классификации объектов: второе собственное значение матрицы тензора инерции, координата центра масс по оси абсцисс, координата центра масс по оси ординат, первое собственное значение матрицы тензора инерции, первый Hu-момент, нормированный центральный момент $n\mu_{22}$, коэффициент плотности, третий Hu-момент, пятый Hu-момент, нормированный центральный момент $n\mu_{20}$, площадь выпуклой оболочки. Временные затраты в результате уменьшения количества признаков для описания объектов были сокращены в 2,3 раза.

Из шести рассмотренных методов отбора информативных признаков лучшие результаты показал метод на основе расчета корреляции по Спирмену, худшие результаты получены для медианного метода, а метод на основе логистической регрессии оказался наименее устойчивым к выбросам в обучающей выборке. Нормировка признаков не оказывает существенного влияния на эффективность работы алгоритмов отбора.

Временные затраты на обучение случайного леса ступенчато возрастают по мере увеличения количества признаков для описания. Ошибка классификации уменьшается по мере увеличения количества признаков, однако при верном их ранжировании можно наблюдать эффект переобучения, когда ошибка либо возрастает, либо остается неизменной по мере увеличения количества признаков в наборе для описания объектов.

Использование 11 признаков вместо 59 позволяет сократить временные затраты при реализации различных алгоритмов классификации и кластеризации, упростить визуализацию результатов. В дальнейшем отобранные признаки будут использоваться для автоматического анализа в программном пакете CellDataMiner.

Список использованных источников

1. Stewart, B. World Cancer Report 2014 / B. Stewart, C. P. Wild. – Geneva : WHO Press, 2015. – 512 p.

2. Программный пакет CellDataMiner для анализа люминесцентных изображений раковых клеток / Е. В. Лисица [и др.] // Информатика. – 2015. – № 4(48). – С. 73–84.

3. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls / O. Ronneberger [et al.] // Chromosome Research. – 2008. – No. 3. – P. 523–562.

4. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection / J. C. Ang [et al.] // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2016. – No. 5. – P. 971–989.

5. Classifying ten types of major cancers based on reverse phase protein array profiles / P. W. Zhang [et al.] // PLoS One. – 2015. – No. 5. – P. 3–7.

6. Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer / J. Sonntag [et al.] // Translational Proteomics. – 2014. – No. 2. – P. 52–59.

7. Kaddi, C. Models for predicting stage in head and neck squamous cell carcinoma using proteomic and transcriptomic data / C. Kaddi, M. D.Wang // IEEE J. of Biomedical and Health Informatics. – 2017. – No. 1. – P. 246–253.

8. Immunosignature system for diagnosis of cancer / P. Stafford [et al.] // Proc. of the National Academy of Sciences of the United States of America. – 2014. – No. 30. – P. 3072–3080.

9. Nguyen, T. Modified AHP for gene selection and cancer classification using type-2 fuzzy logic / T. Nguyen, S. Nahavandi // IEEE Transactions on Fuzzy Systems. – 2016. – No. 2. – P. 273–287.

10. Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification / T. Nguyen [et al.] // PloS One. – 2015. – No. 3.

11. 3-phosphoadenosine 5-phosphosulfate (paps) synthases, naturally fragile enzymes specifically stabilized by nucleotide binding / J. Van den Boom [et al.] // J. of Biological Chemistry. – 2012. – No. 21. – P. 17645–17655.

12. Insights into the classification of small GTPases / D. Heider [et al.] // Advances and Applications in Bioinformatics and Chemistry. – 2010. – No. 3. – P. 15–24.

13. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? / W. G. Touw [et al.] // Briefings in Bioinformatics. – 2013. – No. 3. – P. 315–326.

14. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers / J. N. Dybowski [et al.] // BioData Mining. – 2011. – No. 4. – P. 26–39.

15. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification / M. Riemenschneider [et al.] // BioData Mining. – 2016. – No. 9. – P. 10–16.

16. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas / H. Hoek [et al.] // Cancer Research. – 2004. – No. 15. – P. 5270–5282.

17. Quantitative analysis of estrogen receptor heterogeneity in breast cancer / G. G. Chung [et al.] // Laboratory Investigation. – 2007. – No. 7. – P. 662–669.

18. Camp, R. L. Automated subcellular localization and quantification of protein expression in tissue microarrays / R. L. Camp, G. G. Chung, D. L. Rimm // Nature Medicine. – 2002. – No. 11. – P. 1323–1327.

19. Quantifying estrogen and progesterone receptor expression in breast cancer by digital imaging / M. K. Szesze [et al.] // J. of Histochemistry and Cytochemistry. – 2005. – No. 6. – P. 753–762.

20. Имитационная модель трехканальных люминесцентных изображений популяций раковых клеток / Е. В. Лисица [и др.] // Журнал прикладной спектроскопии. – 2014. – № 6. – С. 907–913.

21. Burger, W. Principles of Digital Image Processing: Core Algorithms / W. Burger, M. Burge. – London : Springer-Verlag, 2009. – 332 p.

22. Jähne, B. Digital Image Processing. Iss. 5 / B. Jähne. – Berlin, Heidelberg : Springer, 2002. – 585 p.

23. Reiss, Th. H. Recognizing Planar Objects using Invariant Image Features / Th. H. Reiss. – Berlin : Springer, 1993. – 186 p.

24. Hu, M. K. Visual pattern recognition by moment invariants / M. K. Hu // IEEE Transactions on Information Theory. – 1962. – No. 8. – P. 179–187.

25. Neumann, U. EFS: an ensemble feature selection tool implemented as R-package and web-application / U. Neumann // BioData Mining. – 2017. – No. 10. – P. 21–30.

26. Bauer, D. F. Constructing confidence sets using rank statistics / D. F. Bauer // J. of the American Statistical Association. – 1972. – No. 67. – P. 687–690.

27. Yu, L. Efficient feature selection via analysis of relevance and redundancy / L. Yu // J. of Machine Learning Research. – 2004. – No. 5. – P. 1205–1224.

28. Suzuki, N. Developing landscape habitat models for rare amphibians with small geographic ranges: a case study of Siskiyou Mountains salamanders in the western USA / N. Suzuki, D. H. Olson, E. C. Reilly // J. of Machine Learning Research. – 2008. – No. 17. – P. 2197–2218.

29. Novel methods improve prediction of species distributions from occurrence data / J. Elith [et al.] // J. of Space and Time in Ecology. – 2006. – No. 29. – P. 129–151.

30. Yu, L. Efficient feature selection via analysis of relevance and redundancy / L. Yu, H. Liu // J. of Machine Learning Research. – 2004. – No. 5. – P. 1205–1224.

31. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach / U. Neumann [et al.] // BioData Mining. – 2016. – No. 9. – P. 36–50.

32. Breiman, L. Random forests / L. Breiman // Machine Learning. – 2001. – No. 5. – P. 5–32.

33. Feature selection based on quality of information / J. Liu [et al.] // Neurocomputing. – 2017. – No. 225. – P. 11–22.

34. Measures of rule quality for feature selection in text categorization / E. Montañés [et al.] // Advances in Intelligent Data Analysis V. – 2003. – No. 225. – P. 589–598.

35. Scikit-learn: Machine learning in Python / F. Pedregosa [et al.] // J. of Machine Learning Research. – 2011. – No. 12. – P. 2825–2830.

References

1. Stewart B., Wild C. P. *World Cancer Report 2014*. Geneva, WHO Press, 2015, 512 p.

2. Lisitsa Y. U., Yatskou M. M., Apanasovich V. V., Apanasovich T. V. Programmnyj paket CellDataMiner dlja analiza ljuminescentnyh izobrazhenij rakovyh kletok [The software package CellDataMiner for data mining of fluorescent images of cancer cells]. *Informatics*, 2015, no. 4(48), pp. 73–84 (in Russian).

3. Ronneberger O., Baddeley D., Scheipl F., Verveer P. J., Burkhardt H., …, Joffe B. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls. *Chromosome Research*, 2008, no. 3, pp. 523–562.

4. Ang J. C., Mirzal A., Haron H., Hamed H. N. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, no. 5, pp. 971–989.

5. Zhang P. W., Chen L., Huang T., Zhang N., Kong X.Y., Cai Y. D. Classifying ten types of major cancers based on reverse phase protein array profiles. *PLoS One*, 2015, no. 5, pp. 3–7.

6. Sonntag J., Bender C., Soons Z., Heyde S. von der, König R., …, Korf U. Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer. *Translational Proteomics*, 2014, no. 2, pp. 52–59.

7. Kaddi C., Wang M. D. Models for predicting stage in head and neck squamous cell carcinoma using proteomic and transcriptomic data. *IEEE Journal of Biomedical and Health Informatics*, 2017, no. 1, pp. 246–253.

8. Stafford P., Cichacz Z., Woodbury N. W., Johnston S. A. Immunosignature system for diagnosis of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, no. 30, pp. 3072–3080.

9. Nguyen T., Nahavandi S. Modified AHP for gene selection and cancer classification using type-2 fuzzy logic. *IEEE Transactions on Fuzzy Systems*, 2016, no. 2, pp. 273–287.

10. Nguyen T., Khosravi A., Creighton D., Nahavandi S. Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. *PloS One*, 2015, no. 3.

11. Boom J. Van den, Heider D., Martin S. R., Pastore A., Mueller J. W. 3-phosphoadenosine 5-phosphosulfate (paps) synthases, naturally fragile enzymes specifically stabilized by nucleotide binding. *Journal of Biological Chemistry*, 2012, no. 21, pp. 17645–17655.

12. Heider D., Hauke S., Pyka M., Kessler D. Insights into the classification of small GTPases. *Advances and Applications in Bioinformatics and Chemistry*, 2010, no. 3, pp. 15–24.

13. Touw W. G., Bayjanov J. R., Overmars L., Backus L., Boekhorst J., Wels M., Hijum van S. A. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, 2013, no. 3, pp. 315–326.

14. Dybowski J. N., Riemenschneider M., Hauke S., Pyka M., Verheyen J., Hoffmann D., Heider D. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Mining*, 2011, no. 4, pp. 26–39.

15. Riemenschneider M., Senge R., Neumann U., Hüllermeier E., Heider D. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Mining*, 2016, no. 9, pp. 10–16.

16. Hoek H., Rimm D. L., Williams K. R., Zhao H., Ariyan S., …, Halaban R. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Research*, 2004, no. 15, pp. 5270–5282.

17. Chung G. G., Zerkowski M. P., Ghosh S., Camp R. L., Rimm D. L. Quantitative analysis of estrogen receptor heterogeneity in breast cancer. *Laboratory Investigation*, 2007, no. 7, pp. 662–669.

18. Camp R. L., Chung G. G., Rimm D. L. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*, 2002, no. 11, pp. 1323–1327.

19. Szesze M. K., Crisman C. L., Crow L., McMullen S., Major J. M., …, Wasserman L. M. Quantifying estrogen and progesterone receptor expression in breast cancer by digital imaging. *Journal of Histochemistry and Cytochemistry*, 2005, no. 6, pp. 753–762.

20. Lisitsa Y. U., Yatskou M. M., Apanasovich V. V., Apanasovich T. V., Shitik M. M. Imitacionnaja model' trehkanal'nyh ljuminescentnyh izobrazhenij populjacij rakovyh kletok [Simulation model for three-channel luminescent images of cancer cell populations]. Zhurnal prikladnoj spektroskopii [*Journal of Applied Spectroscopy*], 2014, no. 6, pp. 907–913 (in Russian).

21. Burger W., Burge M. *Principles of Digital Image Processing: Core Algorithms*. London, Springer-Verlag, 2009, 332 p.

22. Jähne B. *Digital Image Processing*. Iss. 5. Berlin, Heidelberg, Springer, 2002, 585 p.

23. Reiss Th. H. *Recognizing Planar Objects using Invariant Image Features*. Berlin, Springer, 1993, 186 p.

24. Hu M. K. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 1962, no. 8, pp. 179–187.

25. Neumann U. EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData Mining*, 2017, no. 10, pp. 21–30.

26. Bauer D. F. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 1972, no. 67, pp. 687–690.

27. Yu L. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 2004, no. 5, pp. 1205–1224.

28. Suzuki N., Olson D. H., Reilly E. C. Developing landscape habitat models for rare amphibians with small geographic ranges: a case study of Siskiyou Mountains salamanders in the western USA. *Journal of Machine Learning Research*, 2008, no. 17, pp. 2197–2218.

29. Elith J., Graham C. H., Anderson R. P., Dudík M., Ferrier S., …, Zimmermann N. E. Novel methods improve prediction of species distributions from occurrence data. *Journal of Space and Time in Ecology*, 2006, no. 29, pp. 129–151.

30. Yu L., Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 2004, no. 5, pp. 1205–1224.

31. Neumann U., Riemenschneider M., Sowa J.-P., Baars T., Kälsch J., Canbay A., Heider D. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining*, 2016, no. 9, pp. 36–50.

32. Breiman L. Random forests. *Machine Learning*, 2001, no. 5, pp. 5–32.

33. Liu J., Lin Y., Lin M., Wu S., Zhang J. Feature selection based on quality of information. *Neurocomputing*, 2017, no. 225, pp. 11–22.

34. Montañés E., Fernández J., Díaz I., Combarro E. F., Ranilla J. Measures of rule quality for feature selection in text categorization. *Advances in Intelligent Data Analysis V*, 2003, no. 225, pp. 589–598.

35. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., …, Duchesnay É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, no. 12, pp. 2825–2830.

Информация об авторах

Лисица Евгения Владимировна, научный сотрудник, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: ylisitsa@gmail.com

Яцков Николай Николаевич, кандидат физико-математических наук, доцент, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: yatskou@bsu.by

Information about the authors

Yauheniya U. Lisitsa, Researcher, the Faculty of Radiophysics and Computer Technologies of the Belarusian State University, Minsk, Belarus.
E-mail: ylisitsa@gmail.com

Mikalai M. Yatskou, Cand. Sci. (Phys.-Math.), Assoc. Prof., the Faculty of Radiophysics and Computer Technologies of the Belarusian State University, Minsk, Belarus.
E-mail: yatskou@bsu.by

Скакун Виктор Васильевич, кандидат физико-математических наук, доцент, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: skakun@bsu.by

Кривошеев Павел Дмитриевич, студент, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: kryvasheyeu.pavel@gmail.com

Апанасович Владимир Владимирович, доктор физико-математических наук, профессор, первый проректор, Институт информационных технологий и бизнес-администрирования, Минск, Беларусь.
E-mail: apanasovichv@gmail.com

Victor V. Skakun, Cand. Sci. (Phys.-Math.), Assoc. Prof., the Faculty of Radiophysics and Computer Technologies of the Belarusian State University, Minsk, Belarus.
E-mail: skakun@bsu.by

Pavel D. Kryvasheyeu, Student, the Faculty of Radiophysics and Computer Technologies of the Belarusian State University, Minsk, Belarus.
E-mail: kryvasheyeu.pavel@gmail.com

Vladimir V. Apanasovich, Dr. Sci. (Phys.-Math.), Professor, First Vice-Rector, Institute of IT & Business Administration, Minsk, Belarus.
Email: apanasovichv@gmail.com