**УДК: 004.85**

# ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ДЛЯ АНАЛИЗА СЕГМЕНТИРОВАННЫХ ОБЪЕКТОВ НА ЛЮМИНЕСЦЕНТНЫХ ИЗОБРАЖЕНИЯХ РАКОВЫХ КЛЕТОК

***Е.В. ЛИСИЦА, канд. физ.-мат. наук, доц. Н.Н. ЯЦКОВ,***
***канд. физ.-мат. наук, доц. В.В. СКАКУН, д-р физ.-мат. наук, проф. В.В. АПАНАСОВИЧ***
***(Белорусский государственный университет, Минск)***

*Исследованы методы классификации для анализа многоканальных изображений раковых клеток опухоли молочной железы. Каждый объект описывался 13 признаками, из которых 11 признаков описывали форму, 2 признака – цвет. Для тестирования использовался метод перекрестной проверки с контролем по отдельному объекту. Рассмотрены следующие методы: линейного и квадратичного дискриминантного анализа, наивный байесовский классификатор, многослойный персептрон, случайного леса, опорных векторов. Наилучшие результаты классификации получены для метода случайного леса, который показал точность классификации 0,97 при использовании всех признаков. Точность классификации этим методом на основе только признаков цвета – 0,96, а при классификации на основе только признаков формы – 0,92. Следующим по точности классификации является метод линейного дискриминантного анализа, обеспечивающий точность классификации 0,97 по всем признакам, 0,96 – по признакам цвета и 0,90 – по признакам формы. Наихудшие результаты получены для многослойного персептрона.*

***Ключевые слова:*** *машинное обучение, методы классификации, перекрестная проверка, дискриминантный анализ, байесовский классификатор, нейронные сети.*

Рак груди – одно из самых распространенных онкологических заболеваний в мире среди женщин [1]. В Республике Беларусь рак молочной железы занимает одну из ведущих позиций по диагностике и смертности среди женского населения [2]. В случае ранней диагностики (до пяти лет после возникновения опухоли) выживаемость возрастает с 56% до 86% [3]. Основной причиной, вызывающей рак, является неконтролируемый рост клеток в тканях груди [4]. Ненормальный рост клеток может быть как доброкачественным, так и злокачественным. Доброкачественные образования, в отличие от злокачественных, не распространяются в другие органы и части тела. Таким образом, необходима грамотная и своевременная диагностика в лечении опухолей [5]. Онкологические заболевания очень разнообразны, в первую очередь это обусловлено тем, что рак – это болезнь на клеточном уровне, а структура клетки варьируется от человека к человеку [7]. Всего лишь 5–10% раковых опухолей имеют наследственный характер, и около 90% раковых опухолей возникают в результате процессов старения [6].

Современные методы диагностики основаны на данных, получаемых при клиническом наблюдении, и проведении ряда тестов [8]. Для исследования онкологических заболеваний используются методы, основанные на мониторинге молекулярных процессов, происходящих в клетке. Одним из таких методов является люминесцентная микроскопия, которая получила широкое распространение при диагностике онкологических заболеваний. В этом методе интенсивность люминесцентного красителя определяет концентрацию определенных белков в клетке, которые в свою очередь отражают процессы, которые в ней происходят [9]. В результате проведения эксперимента получается несколько десятков изображений, каждое из которых содержит по несколько сотен клеток. Решается задача выделения (сегментации) клеток на изображении и последующего их анализа с целью классификации на здоровые и больные клетки, вычисления их относительной доли на изображении [10]. Известны автоматические методы и программные комплексы сегментации клеток [11], но их применение не всегда дает ожидаемые результаты вследствие большой сложности объектов, представленных на изображениях.

Информацию, полученную в результате сегментации, представляют в виде многомерного набора данных, по которому необходимо определить тип клетки. Для решения задачи классификации типа клетки можно использовать методы машинного обучения [12–14].

Машинное обучение получило широкое распространение в различных научных областях и относится к группе методов искусственного интеллекта. Эти методы используются для выделения информативных признаков и зависимостей между ними [15–17], идентификации, классификации, снижения размерности и прогнозирования [18; 19]. Среди наиболее распространенных методов машинного обучения можно выделить:

1) наивный байесовский классификатор (NB, от англ. Naïve Bayes). На практике NB используют для сравнения с более сложными методами классификации [20];

2) методы дискриминантного анализа (DA, от англ. Discriminant Analysis): линейный DA (LDA, от англ. Linear Discriminant Analysis) и квадратичный DA (QDA, от англ. Quadratic Discriminant Analysis). В методе LDA для оценки параметров распределений классов используется метод наименьших квадратов (lsqr) или спектральное разложение матрицы (eigen), в случае больших объемов данных возможно применение метода сингулярного разложения (svd). В случае QDA основное влияние на качество классификации оказывает параметр регуляризации [21];

3) метод опорных векторов (SVM, от англ. Support Vector Machine). Качество классификации для метода SVM в первую очередь определяется видом функции ядра: линейная функция задается через $C$ – штрафное значение за неверную классификацию и *tol* –точность критерия остановки; полиномиальная функция с параметрами степень $d$ и независимый член $r$; радиально-базисная функция зависит от параметра степени γ; сигмоидальная функция зависит от параметра степени γ и независимого члена *coef*0 [22; 23];

4) деревья решений (DT, от англ. Decision Trees). На практике чаще всего настраиваются следующие параметры DT: критерий разделения (*criterion*), может использоваться критерий Джини (*Gini*) или энтропия (*Entropy*); максимально разрешенная глубина $min_{samples}$; способ разбиения (*splitter*), который может быть случайным (*random*), или выбор наилучшего разбиения (*best*) [24];

5) случайный лес (RF, от англ. Random Forest). Для данного метода настраиваются два параметра: критерий построения деревьев (*criterion*), который может быть *Gini* или *Entropy*, количество деревьев (*n_estimators*) для классификации [24];

6) искусственные нейронные сети (ИНС). Основными параметрами для многослойного персептрона являются количество слоев *n_layer*, количество нейронов *neurons_number* в слое и обучающая функция $g$. На практике используется четыре вида функции $g$: линейная *identity*, которая подходит для решения только линейных задач, сигмоидальная функция *logistic*, функция гиперболического тангенса *tanh* и пороговый переход в нуле *relu* [25].

Часть методов классификации при их применении на практике неустойчивы к наличию выбросов в обучающей выборке, поэтому для отбора методов, устойчивых к выбросам данных, использовалась стандартизация для устранения неоднородности [13].

Рассмотренные методы используются при прогнозировании поведения рака груди, дифференциации между доброкачественными и злокачественными опухолями [20].

ИНС применяются для анализа рентгеновских изображений, получаемых с маммографа при ранней диагностике рака груди [27] и для оценки степени заболевания раком молочной железы [28]. Точность классификации с использованием нейронных сетей варьируется в пределах от 0,76 при исследовании рака легких [29] до 0,95 при исследовании рака желудка [30], когда использовался многослойный персептрон.

Методы машинного обучения также используются для оценки выживаемости при онкологическом заболевании, при этом ИНС и DT показывают точности 0,912 и 0,936 соответственно [31].

Наивный байесовский классификатор показал хорошие результаты классификации при решении задачи постановки диагноза при определении вида опухоли. При этом точность классификации опухолей составляла 0,95 [32]. Наряду с NB используются и другие методы классификации, основанные на применении теоремы Байеса. Например, методы DA (линейный и квадратичный) с точностью классификации от 77% до 79%, а также классификатор на основе гауссовских процессов с точностью классификации от 0,81 до 0,85 [33].

Метод SVM получил широкое распространение в исследовании генома рака, например, для решения задачи типизации раковых заболеваний, поиска новых биомаркеров рака и генов, отвечающих за появление раковых заболеваний. В проведенных исследованиях точность метода SVM варьировалась в пределах от 0,88 до 0,93 [34].

Использование ансамблевых методов, в которых производится усреднение по ансамблю слабых базовых классификаторов, не дает значительного улучшения в точности классификации, однако приводит к увеличению требовательности к вычислительным ресурсам, как это было показано в работе, посвященной исследованию методов DA, NB, SVM и их ансамблевых реализаций при анализе рака простаты [35]. Исключением является RF, представляющий собой ансамблевую реализацию метода DT. Методы RF используются для прогнозирования терапевтического эффекта при лечении онкологических заболеваний, при этом точность составляет 0,932 [36]. Также они используются для постановки диагноза рака простаты, где их точность составляет 0,83 [37].

Целью данной работы является сравнение различных методов классификации раковых клеток на трехканальных люминесцентных изображениях. Из каждой группы методов классификации, базируясь на результатах опубликованных исследований, были выбраны наилучшие: NB, линейный DA, квадратичный DA, DT, SVM, RF, MLP (от англ. Multi-layer perceptrone).

Рассмотрены девять случайно отобранных микрочипов из 187 микрочипов срезов тканей опухолей молочной железы. Экспертным путем установлены 6366 ядер на изображениях. Цель эксперимента – количественный анализ гетерогенности эстроген-рецептора при раке молочной железы.

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°C, 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 антител. 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. Белок маркируется цианиновым красителем Cy3 и регистрируется в зеленом цветовом канале изображения. Для маркировки всех ядер (ДНК) использован краситель DAPI и зарезервирован синий канал; красный канал изображения – для индикации ядер раковых клеток. Краситель Cy5 использован для маркировки белка эстроген-рецептора, который находится в ядрах раковых клеток. Соответственно, маркерами раковых клеток являются два красителя – Cy5 и Cy3. Размер изображений составляет 2048×2048 пикселей в каждом из трех каналов, разрешающая способность – 0,2 мкм/пиксель или 5 мкм [16; 17].

Выполнено сравнение методов классификации LDA, DDA, NB, DT, RF, MLP, LR. Сравнение произведено по признакам формы и цвета ядер, полученных в результате их сегментации на люминесцентных изображениях раковых клеток [16; 17]. Каждый объект (ядро) описывается 13 признаками, из которых два признака относятся к признакам цвета, а одиннадцать – к геометрическим признакам: 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, коэффициент плотности, третий Hu-момент, пятый Hu-момент, нормированный центральный момент, площадь выпуклой оболочки, верхний квантиль интенсивности в красном канале и математическое ожидание в зеленом. Признаки были отобраны при помощи пакета «EFS» языка программирования R. Анализ выполнялся на процессоре Intel(R) Xeon(R) CPU E5-2630 v2 @2.60 Ghz. Для анализа использовались методы, реализованные в библиотеке машинного обучения scikit-learn языка программирования Python. В качестве критерия использовалась точность классификации – относительная доля верно классифицированных объектов. Метод проверки – метод перекрестной проверки с контролем по отдельным объектам [40].

**Байесовские методы.** Для наивного байесовского классификатора получены следующие результаты: точность классификации по всем признакам составила 0,97, при использовании только признаков цвета – также 0,97, для признаков формы – 0,89. Как следует из полученных результатов, использование всех признаков одновременно показывает наилучшие результаты классификации, использование только признаков формы приводит к снижению качества кластеризации в среднем на 10%.

В таблице 1 показаны средние значения точности классификации методом LDA с различными параметрами, усредненные по девяти изображениям и рассчитанные как по всем признакам, так и по признакам формы и цвета в отдельности.

Таблица 1. – Результаты точности классификации для LDA

| Усечение | Метод оценки | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | svd | | | lsqr | | | eigen | | |
| | Признаки | | | | | | | | |
| | Все | Цвет | Форма | Все | Цвет | Форма | Все | Цвет | Форма |
| Нет | 0,97 | 0,96 | 0,90 | 0,97 | 0,96 | 0,90 | 0,75 | 0,93 | 0,75 |
| LW | – | – | – | 0,97 | 0,96 | 0,90 | 0,75 | 0,93 | 0,75 |

При использовании методов svd и lsqr результаты классификации совпадают. Временные затраты на обучение для методов оценки svd и lsqr без усечения одинаковы и составляют 0,003 с. При рассмотрении усечения временные затраты возрастают до 0,004 с. Для методов оценки lsqr и eigen усечение не влияет на качество классификации. Наихудшие результаты были получены для метода оценки eigen. Отличительной чертой LDA является то, что наилучшие результаты классификации (0,93) для него были получены при использовании только признаков цвета. Включение дополнительной информации от признаков формы только снижало точность классификации. Для LDA с методами оценки svd и lsqr характерно улучшение

точности классификации при использовании всех признаков по сравнению с использованием признаков цвета и формы в отдельности. При этом точность классификации с учетом только признаков формы для методов оценки svd и lsqr на 17% больше, чем при рассмотрении признаков формы при методе eigen. Таким образом, для дальнейшего анализа оптимальным для работы с LDA является метод оценки svd.

В таблице 2 показаны средние значения точности классификации метода QDA с различными значениями reg_param, усредненные по девяти изображениям по всем признакам, а также по признакам формы и цвета в отдельности.

Таблица 2. – Результаты точности классификации для QDA

| Признаки | reg_param | | | |
|---|---|---|---|---|
| | 0,1 | 1,0 | 10,0 | 100,0 |
| Все | 0,971 | 0,581 | 0,247 | 0,247 |
| Цвет | 0,967 | 0,910 | 0,247 | 0,247 |
| Форма | 0,891 | 0,581 | 0,247 | 0,247 |

По мере увеличения разброса в данных, что соответствует возрастанию значения reg_param, уменьшается точность классификации. При порядках от второго и более изменение параметра reg_param не приводит к изменению точности классификации. При значениях reg_param нулевого порядка совместное использование признаков формы и цвета позволяет улучшить качество классификации. Когда параметр reg_param принимает значения первого порядка, то наилучшие результаты классификации достигаются при использовании только признаков цвета. Точность классификации в этом случае ниже, чем при использовании значений reg_param нулевого порядка. Точность классификации, полученная методом QDA, сопоставима с методом LDA. Временные затраты на обучение обоих методов также показывают сопоставимые результаты.

**Метод опорных векторов.** Исследование метода SVM было проведено в следующем порядке: сначала SVM с линейным ядром, затем с полиномиальным, сигмоидальным и полиномиальным ядрами.

Для наших данных варьирование параметра *tol* в пределах от $10^{-4}$ до $10^{1}$ не влияет на точность классификации. Это можно объяснить тем, что отдельные признаки имеют большой разброс, в результате чего получаются большие значения ошибки классификации (таблица 3).

Таблица 3. – Результаты точности классификации для линейного SVM

| Признаки | *C* | | | | | |
|---|---|---|---|---|---|---|
| | 0,01 | 0,1 | 1 | 10 | 100 | 1000 |
| Все | 0,81 | 0,71 | 0,82 | 0,82 | 0,81 | 0,81 |
| Цвет | 0,97 | 0,97 | 0,91 | 0,88 | 0,55 | 0,64 |
| Форма | 0,65 | 0,60 | 0,49 | 0,52 | 0,52 | 0,52 |
| Нормированные данные | | | | | | |
| Все | 0,97 | 0,98 | 0,98 | 0,97 | 0,86 | 0,78 |
| Цвет | 0,96 | 0,97 | 0,97 | 0,97 | 0,94 | 0,91 |
| Форма | 0,90 | 0,91 | 0,91 | 0,87 | 0,73 | 0,64 |

Добавление нормировки увеличивает точность классификации для линейного метода опорных векторов при использовании как всех признаков, так и признаков формы и цвета в отдельности. Точность классификации значительно изменилась при использовании только признаков формы. Из этого можно сделать вывод, что признаки формы обладают высокой неоднородностью, а линейный метод опорных векторов является неустойчивым к выбросам. Поскольку наилучшие результаты были получены при штрафном параметре $C = 1$, то в дальнейших исследованиях для метода опорных векторов использовалось это значение.

Основным параметром, который оказал влияние на точность классификации для полиномиального случая, был параметр степени полинома. Наилучшие результаты были получены при использовании полинома первой степени. Как и следовало ожидать, они сопоставимы с результатами для линейного случая метода опорных векторов. Изменение степени полинома приводит к резкому ухудшению точности классификации. Так, для признаков формы, она уменьшается с 0,91 до 0,59. При использовании все признаков точность классификации падает с 0,98 до 0,76. Только для признаков цвета она остается сопоставимой со значениями, полученными при использовании полинома первой степени. В таблице 4 показаны результаты точности классификации при использовании полинома первой степени с различными значениями свободного коэффициента *r*.

Таблица 4. – Результаты точности классификации для SVM с полиномиальным ядром

| Признаки | *r* | | | | | |
|---|---|---|---|---|---|---|
| | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1 | 10 | 100 |
| Все | 0,89 | 0,80 | 0,78 | 0,82 | 0,66 | 0,80 |
| Цвет | 0,81 | 0,96 | 0,97 | 0,97 | 0,97 | 0,96 |
| Форма | 0,79 | 0,68 | 0,61 | 0,49 | 0,56 | 0,55 |
| Нормированные данные | | | | | | |
| Все | 0,94 | 0,97 | 0,98 | 0,98 | 0,97 | 0,91 |
| Цвет | 0,80 | 0,96 | 0,97 | 0,97 | 0,97 | 0,96 |
| Форма | 0,85 | 0,90 | 0,91 | 0,91 | 0,86 | 0,75 |

На ненормированных данных точность варьирования параметров привела к изменению точности классификации только при использовании признаков цвета.

Далее проводились исследования метода SVM с сигмоидальным ядром. Результаты точности классификации показаны в таблице 5 и на рисунке 1. Параметр *gamma* оказывает основное влияние на точность классификации. При его неверном выборе параметр *coef*0 не оказывает влияния на точность.

Таблица 5. – Результаты точности классификации для SVM с сигмоидальным ядром

| *coef*0 | *gamma* | | |
|---|---|---|---|
| | 0,001 | 0,01 | 0,1 |
| -100 | 0,75 | 0,75 | 0,79 |
| -10 | 0,75 | 0,82 | 0,58 |
| -1 | 0,97 | 0,52 | 0,75 |
| 0 | 0,947 | 0,53 | 0,75 |
| 1 | 0,927 | 0,61 | 0,75 |

На рисунке 1 показаны результаты точности классификации SVM с сигмоидальными ядром на нормированных данных. В отличие от исследования на ненормированных данных, где точность классификации была одинаковой и составляла 0,75 вне зависимости от изменения параметров *gamma* и *coef*0, наилучшая точность классификации была получена при $gamma = 0,01$ и $coef\,0 = 0$.

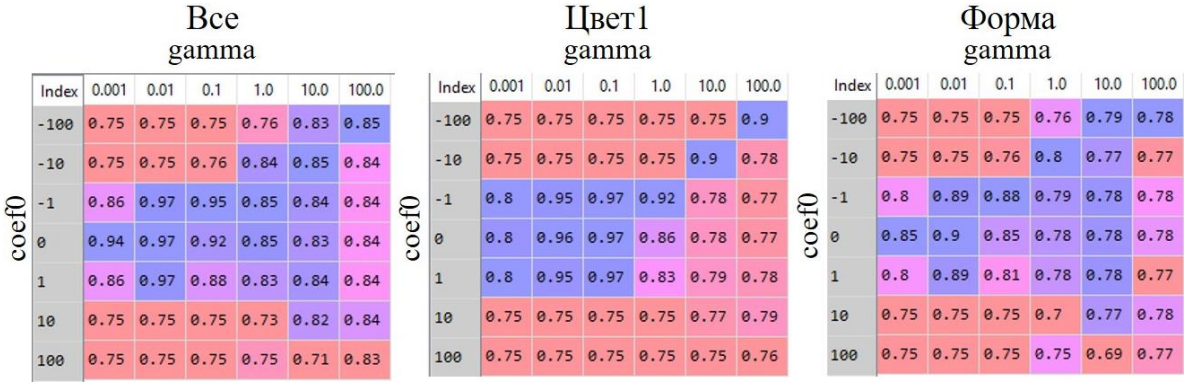


**Рисунок 1. – Результаты точности классификации для SVM с сигмоидальным ядром на нормированных данных**

Метод опорных векторов вне зависимости от вида используемого ядра является неустойчивым к наличию выбросов в выборке, а также чувствительным к выбору параметров метода, что затрудняет его использовании на практике.

**Решающие деревья.** В таблице 6 показаны результаты классификации решающими деревьями, построенные по алгоритму CART с варьированием двух параметров: *criterion* и *splitter*.

Таблица 6. – Результаты точности классификации для решающих деревьев

| Признаки | Все | | Цвет | | Форма | |
|---|---|---|---|---|---|---|
| *splitter* | *criterion* | | | | | |
| | *Gini* | *Entropy* | *Gini* | *Entropy* | *Gini* | *Entropy* |
| *random* | 0,961 | 0,960 | 0,956 | 0,954 | 0,883 | 0,875 |
| *best* | 0,965 | 0,961 | 0,955 | 0,956 | 0,887 | 0,876 |

Изменение критерия кластеризации *criterion* или способа разбиения *splitter* не оказывает значительного влияния на точность классификации. Изменение точности классификации не превосходило 0,01%. При использовании всех признаков были получены наилучшие результаты классификации, наихудшие – при использовании только признаков формы.

**Метод случайного леса.** Количество деревьев *n_estimators* при построении леса принимало значения *10*, *20*, *30*, *40*, *50*, *60*, *70*, *80*, *90*; использовались критерии построения деревьев Джинни и энтропии. При использовании всех признаков варьирование параметров метода не влияло на точность классификации, которая составляла 0,97. Аналогичные результаты были получены при отдельном рассмотрении признаков цвета (0,96) и формы (0,92). Ансамблевый метод классификации показал незначительное улучшение точности классификации по сравнению с решающими деревьями с 0,96 до 0,97 при использовании всех признаков. Однако он показал большую устойчивость при классификации по признакам формы, когда точность классификации была улучшена с 0,88 до 0,92.

**Многослойный персептрон.** Для исследования были рассмотрены однослойный и двухслойный персептроны с *1*, *5*, *10* и *20* нейронами в каждом слое. Использование нормированного набора данных позволяет улучшить точность классификации от 5 до 25%. Однако, поскольку нормировка устраняет неоднородность, это говорит о чувствительности метода к выбросам. Как и ожидалось, общей тенденцией является то, что двухслойная нейронная сеть показывает лучшие результаты классификации по сравнению с однослойным персептроном на всех наборах данных. Исключение составляет многослойный персептрон с передаточной функцией «пороговый переход в нуле», которая показала наилучшие результаты классификации для всех нейронных сетей. При использовании количества нейронов в сети большего, чем количество признаков, точность классификации улучшается только на ненормированном наборе данных. На нормированном наборе данных происходит либо ухудшение точности классификации, либо она остается неизменной. В таблице 7 показана точность классификации многослойным персептроном с передаточной функцией порогового перехода в нуле в зависимости от набора данных и количества нейронов в слое.

Таблица 7. – Результаты точности классификации для многослойной нейронной сети с передаточной функцией порогового перехода в нуле

| Архитектура нейронной сети | Количество нейронов | Тип признаков | | | | | |
|---|---|---|---|---|---|---|---|
| | | Все | Цвет | Форма | Все | Цвет | Форма |
| | | Ненормированные данные | | | Нормированные данные | | |
| | (1) | 0,80 | 0,87 | 0,75 | 0,96 | 0,96 | 0,90 |
| | (1, 1) | 0,92 | 0,96 | 0,83 | 0,97 | 0,97 | 0,91 |
| | (5) | 0,77 | 0,79 | 0,76 | 0,79 | 0,78 | 0,78 |
| | (5, 5) | 0,90 | 0,91 | 0,81 | 0,91 | 0,85 | 0,86 |
| | (10) | 0,78 | 0,89 | 0,77 | 0,95 | 0,95 | 0,89 |
| | (10, 10) | 0,88 | 0,97 | 0,82 | 0,97 | 0,97 | 0,91 |
| | (20) | 0,75 | 0,82 | 0,71 | 0,91 | 0,87 | 0,86 |
| | (20, 20) | 0,92 | 0,95 | 0,85 | 0,97 | 0,97 | 0,91 |

**Заключение.** Для большинства методов классификации точность увеличивается при использовании всех признаков (как формы, так и цвета). Все методы классификации показали значительное ухудшение точности при использовании только признаков формы или только цвета.

Метод опорных векторов и многослойный персептрон являются неустойчивыми к наличию выбросов в выборке, поэтому их использование на практике нецелесообразно.

Наихудшие результаты были получены для метода опорных векторов, этот метод является неустойчивым к наличию выбросов при обучении. При этом результаты классификации метода опорных векторов сильно зависят от выбранных параметров.

Оптимальным методом для решения поставленной задачи является метод случайного леса. Этот метод устойчив к наличию выбросов в обучающей выборке, т.к. в основу метода положено коллективное решение по наборам деревьев.

В группе байесовских методов при классификации по всем признакам и признакам цвета были получены результаты сопоставимые с методом случайного леса. Однако, при использовании только признаков формы точность классификации случайным лесом выше, чем у методов на основе теоремы Байеса.

## ЛИТЕРАТУРА

1. Программный пакет CELLDATAMINER для анализа люминесцентных изображений раковых клеток / Е.В. Лисица [и др.] // Информатика. – 2015. – № 4. – Стр. 73–84.

2. Третьяк, И.Ю. Клинико-морфологическая характеристика пациентов с отечноинфильтративной формой рака молочной железы / И.Ю. Третьяк, Ю.Е. Демидчик, С.А. Костюк // Актуальные вопросы диагностики и лечения злокачественных новообразований : материалы респ. науч.-практ. конф., посвященной 40-летию кафедры онкологии БГМУ, Минск, 21 нояб. 2014 / Белорус. гос. мед. ун-т ; под общ. ред. проф. А.В. Прохорова. – Минск, 2014. – С.102–106.
3. Paulin, F. and Santhakumaran, A. Extracting rules from feed forward neural networks for diagnosing breast cancer / F. Paulin, A. Santhakumaran // Artificial Intelligent Systems and Machine Learning. – 2009. – V. 1, № 4 – P. 143–146.
4. Global and Local Structure Preservation for Feature Selection / X. Liu [et al.] // Neural Networks and Learning Systems, IEEE Transactions. – 2013. – V. 99 – P. 1.
5. Hota, H. Diagnosis of breast cancer using intelligent techniques / H. Hota // Int J Emerg Sci Eng (IJESE). – 2013. –V. 1, № 3 – P. 45–53.
6. Reif, M. Efficient feature size reduction via predictive forward selection / M. Reif, F. Shafait // Pattern Recognition. – 2014. – V. 47, № 4 – P. 1664–1673.
7. Cancer worldwide // World cancer report 2014 / B. Stewart, C. P. Wild. – WHO, 2016. – P. 16–81.
8. Gayathri, B. Breast cancer diagnosis using machine learning algorithms –ASurvey / B. Gayathri, C. Sumathi, T. Santhanam // International Journal of Distributed and Parallel Systems. – 2013. – 4(3). – P. 105.
9. Digital quantitative measurements of gene expression / V. Mikkilineni [et al.] // Biotechnology and bioengineering. – 2004. – № 86, V. 2. – P. 117–124.
10. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls / O. Ronneberger [et al.] // Chromosome Res. – 2008. – № 16, V. 3. – P. 523–562.
11. Алгоритм автоматической сегментации границ ядер раковых клеток на трехканальных люминесцентных изображениях / Лисица Е.В. [и др.] // Журнал прикладной спектроскопии. – 2015. – № 82, Т. 4. – С. 598–607.
12. Stabenfeldt, S.E. Current trends in biomarker discovery and analysis tools for traumatic brain injury / S.E. Stabenfeldt, B.I. Martinez // Journal of Biological Engineering. – 2019. – № 13, V. 16. – P. 1–12.
13. Интеллектуальный анализ данных : пособие / Н.Н. Яцков. – Минск : БГУ, 2014. – 151 с.
14. MaGIC: a machine learning tool set and web application for monoallelic gene inference from chromatin / S. Vinogradova [et al.] // BMC Bioinformatics. – 2019. – № 20, V. 1. – P. 106–111.
15. Neumann, U. EFS: an ensemble feature selection tool implemented as R-package and web-application / U. Neumann, N. Genze, D. Heider // BioData Mining. – 2017. – № 10, V. 21.
16. Отбор характеристик распределения интенсивности в цветовых каналах на люминесцентных изображениях раковых клеток / Е. В. Лисица [и др.] // Журнал прикладной спектроскопии. – 2019. – Т. 86, № 3 – С. 394–400.
17. Отбор информативных геометрических признаков ядер клеток на люминесцентных изображениях раковых клеток / Е. В. Лисица [и др.] // Информатика. – 2019. – Т. 16, № 2 – С. 16–26.
18. Machine Learning-based Analysis of Rectal Cancer MRI Radiomics for Prediction of Metachronous Liver Metastasis / M. Liang [et al.] // Acad Radiol. – 2019. [Epub ahead of print].
19. Constructing Prediction Model of Lung Cancer Recurrence Risk Using Gene Function Clustering and Machine Learning / J. Zhong [et al.] / Comb Chem High Throughput Screen.
20. Tackling the poor assumptions of naive bayes text classifiers / J.D. Rennie [et al.] // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). – Washington DC, 2003. – P. 616–623.
21. Hastie, T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman. – Springer, 2009. – P. 106–119.
22. Guyon, I. Automatic Capacity Tuning of Very Large VC-dimension Classifiers / I. Guyon, B. Boser, V. Vapnik // Advances in neural information processing. – 1993.
23. Cortes, C. Support-vector networks / C. Cortes, V. Vapnik // Machine Learning. – 1995. – V. 20. – P. 273–297.
24. Classification and Regression Trees / L. Breiman [et al.] // Wadsworth, Belmont-CA. – 1984.
25. Scikit-learn: Machine Learning in Python / Pedregosa [et al.] // JMLR. – 2011. – V. 12, – P. 2825–2830.
26. Machine learning models in breast cancer survival prediction / M. Montazeria [et al.] // Technology and Health Care. – 2016.
27. Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer / M.M. Mehdy [et al.] // Comput Math Methods Med. – 2017.
28. Assessing Breast Cancer Risk with an Artificial Neural Network / M. Sepandi [et al.] // Comput Math Methods Med. – 2017.
29. Application of artificial neural network model combined with four biomarkers in auxiliary diagnosis of lung cancer / Xiaoran Duan [et al.] // Med Biol Eng Comput. – 2017. – № 55, V. 8. – P. 1239–1248.
30. Korhani Kangi, A. Predicting the Survival of Gastric Cancer Patients Using Artificial and Bayesian Neural Networks/ A. Korhani Kangi, A. Bahrampour // Asian Pac J Cancer Prev. – 2018. – № 19, V. 2. – P. 487–490.

31. Delen, D. Predicting breast cancer survivability: a comparison of three data mining methods / D. Delen, G. Walker, A. Kadam // Artificial intelligence in medicine. – 2005. – № 34, V. 2. – P. 113–127.
32. Jiang, W. A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system / W. Jiang [et al.] // Int J Cancer. – 2018. – № 42, V. 2. – P. 357–368.
33. Maniruzzaman, Md. Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm / Md. Maniruzzaman // Computer Methods and Programs in Biomedicine. – 2017. – V. 152. – P. 23–34.
34. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics / S. Huang [et al.] // Cancer Genomics Proteomics. – 2018. – № 15, V. 1. – P. 41–51.
35. Viswanath, S.E. Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: a multi-site study / S.E.Viswanath [et al.] // BMC Med Imaging. – 2019. – № 19, V. 1. – P. 22–34.
36. Predictors of the therapeutic effect of corticosteroids on radiation-induced optic neuropathy following nasopharyngeal carcinoma/ B. Zheng [and et.]// Support Care Cancer. – 2019.
37. Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen / L.H. Xiao [et al.] // Asian J Androl. – 2017. – № 19, V. 5. – P. 586–590.
38. Camp, R.L. Validation of tissue microarray technology in breast carcinoma / R.L. Camp, L.A. Charette, D.L. Rimm // Lab Invest. – 2000. – № 80, V. 12. – P. 1943–1949.
39. Zhang, H. The optimality of Naive Bayes / H. Zhang // Proc. FLAIRS. – 2004. – № 1, V. 2. – P. 1–6.
40. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection / R. Kohavi // Intl. Jnt. Conf. AI Montreal, Quebec, Canada. – 1995. – V. 2 – P. 1137–1143.

*Поступила 12.03.2020*

## CLASSIFICATION METHODS FOR THE ANALYSIS OF SEGMENTED OBJECTS ON FLUORESCENT IMAGES OF CANCER CELLS

***Y. LISITSA, M. YATSKOU, V. SKAKUN, V. APANASOVICH***

*The methods of classification to analyze the multi-channel fluorescent images of breast cancer were studied. Each object is described by 13 features, where 11 features are geometry characteristics and 2 features corresponds to color characteristics. The methods were studied onb the standardized and not-standardized data. The cross validation was used. The considered methods are linear and quadratic discriminant analysis, Naive Bayes, Support Vector Machines, Decision Trees, Random Forest, Neural network models. The most sufficient result were received for the Random Forest methods, where the accuracy of the classification is 0,97, when all features are used. If only color features are exploited, the accuracy is 0,96, and finally it is 0,92 for form features. The same results received the linear discriminant analysis, where the accuracy based on all features is 0,97, the accuracy received by color features is 0,96. Which is the same as for random forest classification. However for form features it is only 0,90. The most unsufficient results are obtained for Multi-layer Perceptron.*

***Keywords:*** *machine learning, classification methods, cross-validation, discriminant analysis, Bayesian classifier, neural networks.*