УДК 519.254+004

МОДЕЛИ И МЕТОДЫ ЭКОНОМИЧНОГО РАНЖИРОВАНИЯ ДИАГНОСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ ПРИ ИХ АППРОКСИМАЦИИ РАСПРЕДЕЛЕНИЕМ ЛАПЛАСА

Статья поступила в редакцию 06.07.2017, в окончательном варианте – 10.10.2017.

Колесин Игорь Дмитриевич, Санкт-Петербургский государственный университет, 198504, Российская Федерация, г. Санкт-Петербург, Петергоф, Университетский проспект, 35,
доктор физико-математических наук, профессор, e-mail: kolesin_id@mail.ru, https://elibrary.ru/author_profile.asp?id=822730
Трояножко Ольга Александровна, Санкт-Петербургский государственный университет, 198504, Российская Федерация, г. Санкт-Петербург, Петергоф, Университетский проспект, 35,
аспирант, e-mail: med_otpor@mail.ru

Предложен экономичный метод ранжирования диагностических показателей (ДП) для классификации опухолей по двум группам («доброкачественная» или «злокачественная»). Метод апробирован с использованием базы данных, сформированной Висконсинским университетом в США. Реализованы два этапа метода: сначала выявляются наиболее информативные ДП из всех доступных; затем – проводится классификация. На первом этапе выбор и оценка степени информативности ДП выполняется с использованием коэффициента перекрытия. Он используется для измерения меры сходства между двумя функциями распределения или двумя выборками, представленными этими распределениями. При этом, чем меньше площадь перекрытия функций плотности распределения для различных видов объектов, тем более информативен ДП. При аппроксимации мы использовали классическое распределение Лапласа. Наличие для него простой аналитической формы первообразной является преимуществом перед иными распределениями, не имеющими простой первообразной в явном виде. В результате ранжирования мы получаем упорядоченный по уменьшению степени информативности список ДП на основе соответствующих им значений коэффициентов перекрытия. На практике важна экономичность не только алгоритма, но и программной разработки на его основе – особенно при массовых проведениях исследований. Исходя из этого, нами при сравнении учитывалась временн***а***я сложность и экономичность алгоритма ранжирования. Отметим, что рост вычислительных мощностей ЭВМ может снижать затраты времени на проведение расчетов, но не улучшает качества распознавания объектов. На втором (заключительном) этапе был использован алгоритм, основанный на применении дискретной функции ошибок. Сравнительный анализ подтвердил, что точность диагностики вида опухолей оказалась не хуже точности, полученной другими методами, причем при меньшей сложности алгоритма ранжирования. В результате точность диагностики по трем ДП, найденным данным методом ранжирования, удалось повысить до 96,31 % (для совокупности объектов, представленных в указанной выше базе данных Висконсинского университета). Предложенный метод ранжирования может быть использован на практике как один из вспомогательных – для ранней экспресс-диагностики вида опухолей при массовых обследованиях.

Ключевые слова: высокотехнологичная диагностика, рак молочной железы, вычислительные алгоритмы, диагностика, информатизация, качество медицинских услуг, коэффициент перекрытия, медицинские информационные системы, ранжирование показателей, распределение Лапласа, сложность алгоритма ранжирования, телемедицинские технологии

MODELS AND METHODS OF EFFICIENT DIAGNOSTIC FEATURE RANKING VIA LAPLACE APPROXIMATION

The article has been received by editorial board 06.07.2017, in the final version – 10.10.2017.

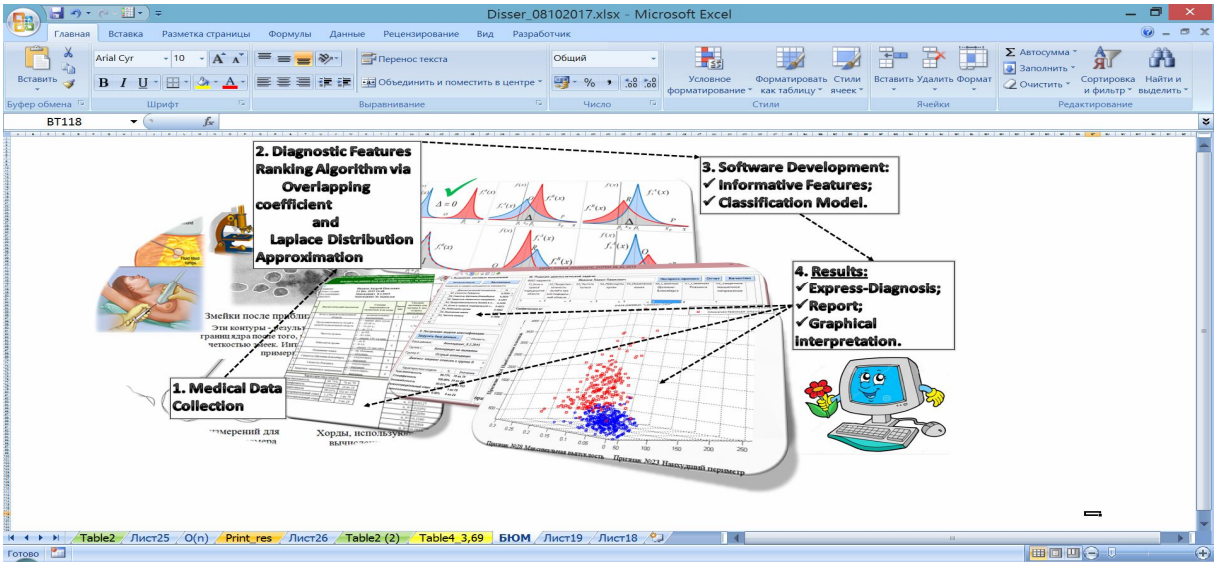
Kolesin Igor D., Saint Petersburg State University, 35 Universitetskiy Ave., Saint Petersburg, Peterhof, 198504, Russian Federation,
Doct. Sci. (Physics and Mathematics), Professor, e-mail: kolesin_id@mail.ru, https://elibrary.ru/author_profile.asp?id=822730
Troyanozhko Olga A., Saint Petersburg State University, 35 Universitetskiy Ave., Saint Petersburg, Peterhof, 198504, Russian Federation,
postgraduate, e-mail: med_otpor@mail.ru

An economical method of diagnostic features (DF) ranking for classification of tumors in two groups (“benign” or “malignant”) based on data from the University of Wisconsin is presented. Two stages of method are implemented: first, the most informative DF from all available are identified; then – the classification is carried out. At the first stage selection and

assessment of DFinformativeness degree is made using the overlapping coefficient (OVL). OVL is used to measure the similarity between two distribution functions or two samples represented by these distributions. At the same time, the smaller overlapping of the distribution density functions for different types of objects is, the more informative is the DF. While approximating, we used the classical Laplace distribution. The presence of a simple analytic form of the antiderivative for it is an advantage over other distributions that do not have a simple primitive in explicit form. As a result of ranking, we havegot a list of indicators ordered by decreasing degree of informativeness based on the corresponding OVL values. In practice, it is more important not only to economize the algorithm, but also to develop software – especially for large-scale research. Based on this, the comparison took into account the time complexity and profitability of the ranking algorithm. We should note, that increasing ECM computing capacity can reduce time spent on calculations, but does not improve the quality of object recognition. At the second, final stage, an algorithm based on the use of a discrete error function was used. Comparative analysis confirmed that the accuracy of diagnosis of the type of tumors was notworse than the accuracy obtained by other methods, but ranking algorithm was less complex. As a result, we managed to increase accuracy of diagnosis for the three DF found by this ranking method to 96,31 %. The proposed method of ranking can be used in practice as one of the auxiliary ones for early rapid diagnosis in mass surveys.

Keywords: hi-tech diagnostics, breast cancer,computational algorithms, diagnostics, informatization, quality of medical services, overlapping coefficient, medical information systems, feature ranking, Laplace distribution, ranking algorithm complexity, telemedical technologies

Graphical annotation (Графическая аннотация)



Введение. Одна из первоочередных задач врача-онколога заключается в выявлении наличия патологии и постановке диагноза на основании доступных результатов инструментальной диагностики, а также сведений, собранных в процессе анамнеза пациентов.

Как правило, количество информации, на основании которой врачом-онкологом делается заключение, достаточно велико. Как следствие количество диагностических показателей (ДП) также оказывается чрезмерно большим для оценки и анализа. Предварительный этап обследований включает использование рентгенографических методов или УЗИ, что позволяет получить соответствующие изображения. В случаях, подозрительных на наличие новообразований, дополнительно используются тонкоигольные аспирационные пункции с последующими патогистологическими исследования срезов тканей – для оценки количества делящихся клеток. Технология получения микрофотографий и то, что при этом является ДП, описано ниже в разделе «Экспериментальный материал и методика исследований».

По результатам анализа характеристик ядер клеток формируется набор ДП и их значений в виде матрицы. Для удобства работы с базой данной (БД) о ДП ставится задача понижения размерности этой матрицы. Это необходимо для выбора самых существенных ДП, имеющих наибольшую дифференцирующую способность в отношении оценки категории оцениваемых объектов. Из практики известно, что можно успешно разделять множества, используя лишь некоторое количество из всего множества имеющихся ДП. Такое понижение размерности матрицы (уменьшение количества анализируемых ДП) облегчает и ускоряет процесс классификации.

Поэтому целью настоящей статьи является разработка методов ранжирования для повышения эффективности медицинской диагностики на примере оценки «доброкачественности» или «злокачественности» опухолей молочной железы. При этом в работе акцент делается на экономичность ранжирования ДП, с использованием коэффициента перекрытия подграфиков плотностей распределения при их аппроксимации распределением Лапласа.

Общая характеристика проблематики ранжирования. На данный момент в литературе представлены три главные парадигмы ранжирования для выбора ДП.

(1) Метод фильтрации (от англ. «filter method») основан на использовании меры релевантности между показателем и группой, к которой относится пациент (здоровый/больной, доброкачественная/злокачественная опухоль) [4, 17, 29].

(2) Метод обвития (от англ. «wrapper method»), где набор показателей тесно привязан к используемому алгоритму [20, 25].

(3) Встроенный метод (от англ. «embedded method»), использующий возвратный алгоритм удаления показателей [14, 30].

В вычислительном отношении методы фильтрации обычно оказываются самыми быстрыми из трех рассматриваемых методов отбора ДП, т.к. они не требуют использования алгоритма обучения для оценивания показателей. Также в методах фильтрации отсутствует зависимость от множества дополнительных параметров, подбор которых часто является искусством и плохо алгоритмизуется.

Из последних работ по разработке моделей, методов и комплексов программ в сфере здравоохранения, предназначенных для решения диагностических задач, аналогичных по тематике данной статье, можно выделить, например, работы [1, 2, 9].

В данной статье исследуется алгоритм ранжирования ДП на основе распределения Лапласа с приложением его к БД по раку молочной железы, сформированной в Висконсинском университете США. Недостатки существующих методов ранжирования среди ближайших аналогов трудно выразить в явной форме, т.к. сравнений лишь по трудозатратам или по затраченному машинному времени недостаточно. При этом стоит разделять понятия экономичности подхода к ранжированию и экономичность алгоритма ранжирования. Экономичность алгоритма состоит в меньшем числе операций, а экономичность подхода – в меньшем числе «образов», дающих представление о выполнении ранжирования. В данной работе используются простые геометрические образы (фигуры, площади, перекрытие), а результат получается хороший. Введем дополнительные термины: «операционная емкость алгоритма ранжирования» – число операций перед присвоением ранга; «иллюстративная емкость алгоритма ранжирования» – число используемых образов. Эти термины можно (удобно) применить при сравнении алгоритмов. С использованием данных терминов, можно утверждать, что операционная емкость алгоритма ранжирования при аппроксимации значений рассматриваемых в данной статье ДП распределением Лапласа **лучше**, чем при аппроксимации распределением Гаусса – вследствие простой первообразной. Также иллюстративная емкость алгоритма ранжирования при аппроксимации распределениями Лапласа и Гаусса **лучше**, чем при полном переборе ДП.

Обзор результатов использования распределения Лапласа в прикладных задачах. С конца 1970-х гг. стали чаще публиковаться работы, в которых для обработки демографических, медицинских, экономических и других данных использовался первый закон Лапласа вместо распределения Гаусса [22]. Связь с χ^2 – распределением также увеличивает частоту использования распределения Лапласа. В работе [3] показано возникновение распределения Лапласа вместо предельного нормального закона. В статье [6] описаны свойства двухкомпонентного многомерного распределения Лапласа, предложено его применение к задаче распознавания речевых сигналов, показана его связь с многомерным нормальным распределением. Некоторые прикладные задачи, решенные с помощью распределения Лапласа: прогнозирование появления ураганов и смерчей в Японии [24]; моделирование финансовых потоков – как в личных доходах, так и на фондовых биржах [21,26]; распознавание речи и аудио сигналов [16]. Обратимся к предыстории ранжирования показателей на основе коэффициента перекрытия (далее – «Δ»). По определению, Δ – это относительная величина, равная площади перекрытия графиков плотностей распределения [14]. Величина Δ является мерой сходства между распределениями и принимает значения $[0;1]$ [18, 19]. Если $\Delta = 0$, то перекрытие отсутствует, если $\Delta = 1$, то распределения равны. Таким образом, из минимального Δ следует максимальная информативность показателя, а из максимального Δ – минимальная информативность, что показано на рисунке 1.

Метод на основе использования Δ получил широкое применение в области распознавания и классификации образов. Приведем некоторые примеры использования Δ: в экспериментах по молекулярной классификации опухолей для определения значимых подтипов рака [28]; при оценке параметров в смеси нормальных распределений [23]; для описания различий в поведенческих характеристиках у диких синиц в рамках индивидуального и группового поведения [12]; для проведения тестов на симметрию, используя оценку плотности ядра и расстояние Кульбака – Лейблера [27] и т.д.

Исходя из вышеизложенного, можно сделать вывод, что ранжирование ДП с помощью коэффициента перекрытия Δ на основе распределения Лапласа является одним из результативных подходов в задачах классификации.

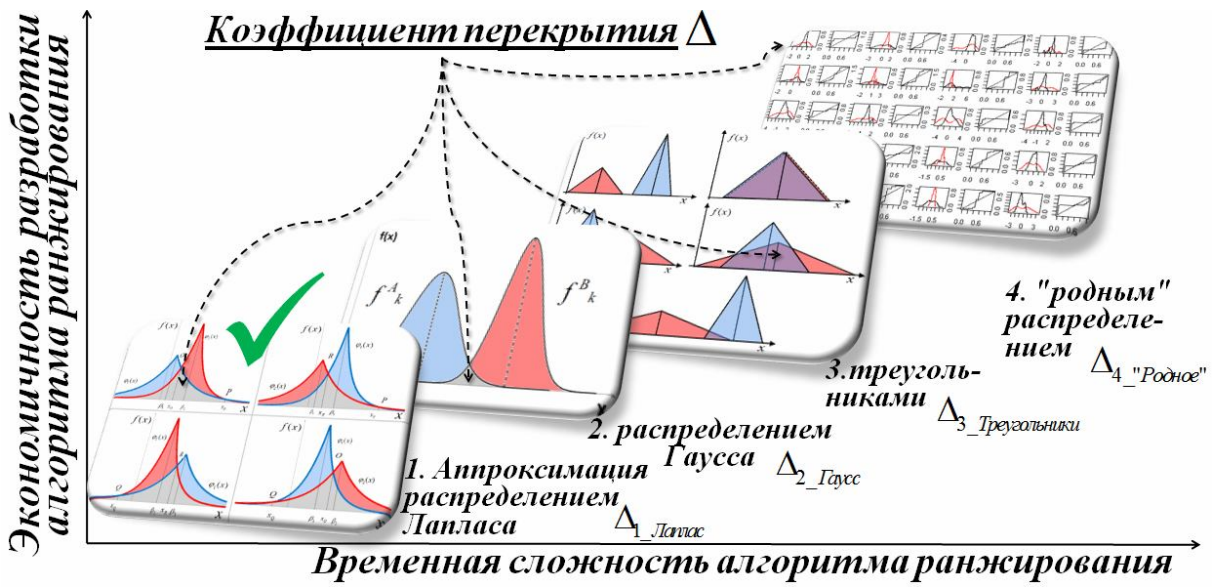


Рисунок 1 – Графическое представление коэффициента перекрытия Δ

Экспериментальный материал, использованный при исследовании. Работоспособность и эффективность метода ранжирования ДП на основе распределения Лапласа исследовалась на данных пациентов с различными патологиями. В данной статье приведены эксперименты с БД по раку молочной железы РМЖ-569, находящейся в открытом доступе репозитория Висконсинского университета (University of Wisconsin) и доступной для скачивания по ссылке [13]. БД содержит данные о 569 пациентах, имеющих опухоль, и о 30 непрерывных по времени ДП у этих пациентов. Указанные ДП в БД РМЖ-569 были получены в результате инвазивного метода диагностики – тонкоигольной аспирационной пункции. По результатам анализа характеристик ядер клеток при патогистологическом исследовании проводилось разделение пациентов по двум группам: доброкачественная опухоль и злокачественная. По аналогии со способом диагностирования злокачественности опухолей по формам теней на рентгеновских снимках, где характеристики контуров новообразований для специалистов-рентгенологов очень информативны, в исследуемой БД РМЖ-569 исследуются характеристики контуров ядер клеток. Исследование микрофотографий проводилось методом «змеек» (от англ. “snakes”), в котором анализируется не исходная клетка, а её контур. Ниже на рисунке 2 приведен образец микрофотографий с указанием на ней некоторых характеристик клеток (гладкость, симметрия, вогнутость, фрактальный размер).

Пункции делались, т.к. у пациентов были серьезные подозрения на рак. Сами препараты после пункций получались срезанием тонких «ломтиков» тканей на микротоме, а потом прокрашиванием (например, гематоксилин-эозином). В рассматриваемой БД отражена не доля делящихся клеток, а оценки характеристик ядер клеток.

Оцифрованные снимки (микрофотографии) представлены тремя подгруппами ДП, характеризующими ядра клеток.

1 – Размер, выраженный радиусом и площадью.

2 – Форма ядер клеток, выраженная в виде совокупности характеристик: гладкость, степенью вогнутости, плотность, доля вогнутых участков контура, симметрией и фрактальностью.

3 – Текстура, полученная разницей интенсивности черно-белой шкалы в компонентах пикселя [31, 32].

Снимки для анализа в указанной выше базе данных были сгенерированы цветной видео камерой JVC TK-1070U, прикрепленной сверху микроскопа Olympus. Слайд проектировался на камеру с объективом 63х и линзой 2,5х. Снимок считывался программой ComputerEyes/RT (Digital Vision, Inc., Dedham MA 02026) как targa файл 512×480. Конечный снимок сохранялся в памяти как двумерный массив, где каждому элементу изображения (пикселю) соответствовала величина между 0 и 255, отражающая световую интенсивность в данной точке. При обработке микрофотографий были выбраны области с наименьшим перекрытием ядер клеток.

При ранжировании ДП авторами настоящей статьи используется идея включения в число показателей граничных и средних значений с разбиением на специфические подгруппы. Погрупповой подход подразумевает разбиение ДП на группы. Такой подход имитирует консилиум врачей разных специальностей, где каждую выделенную группу ДП оценивает врач-специалист в соответствующей предметной области. Приведем схему работы погруппового подхода при выделении информативных показателей. Она состоит в объединении специфических показателей в меньшие группы – с выделением наиболее информативных ДП в каждой группе, что показано на рисунке 3.

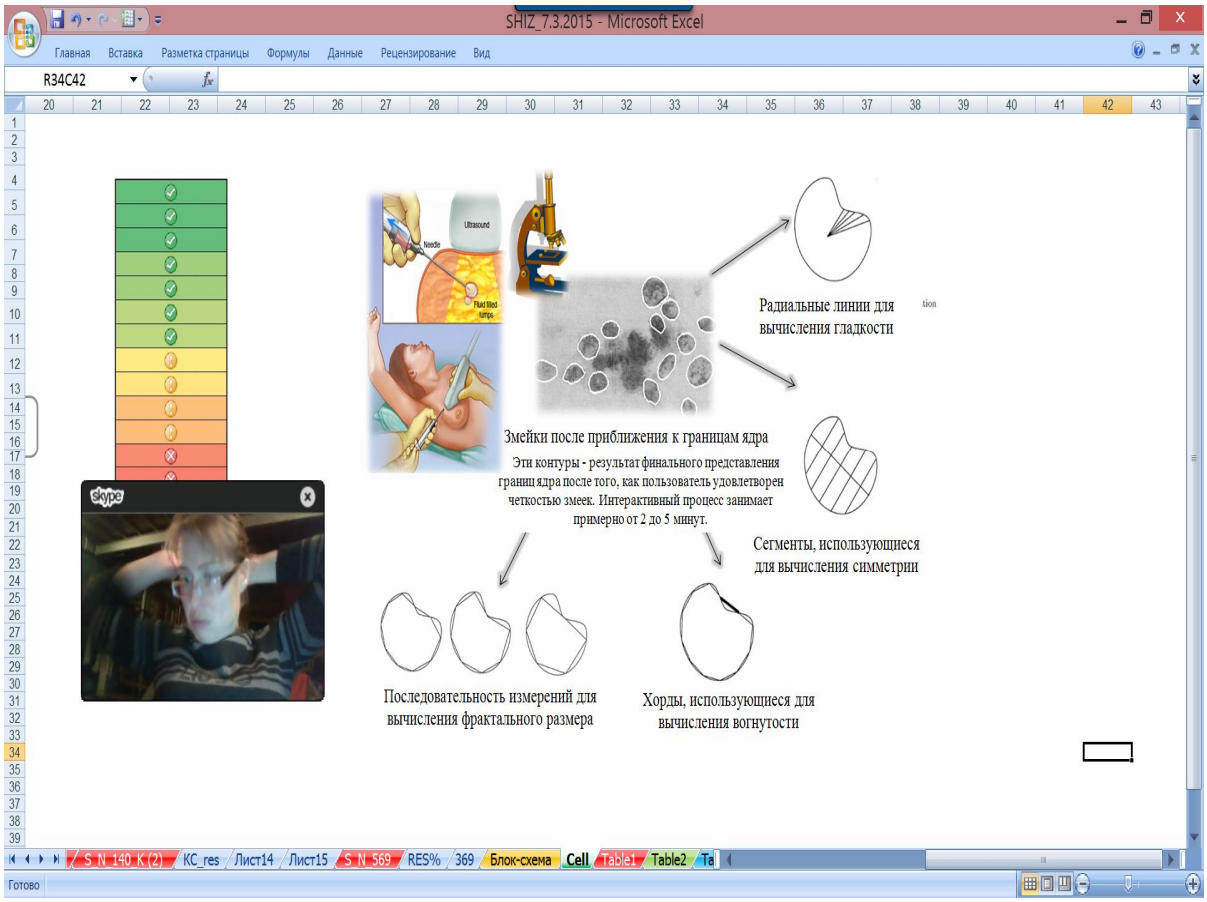


Рисунок 2 – Системное представление объекта исследования при патологии опухоли молочной железы: часть исходной информации представлена в виде элементов, соответствующих измеряемым физиологическим показателям; степень искажения формы ядра клетки является исходной информацией для работы описываемых алгоритмов [31]

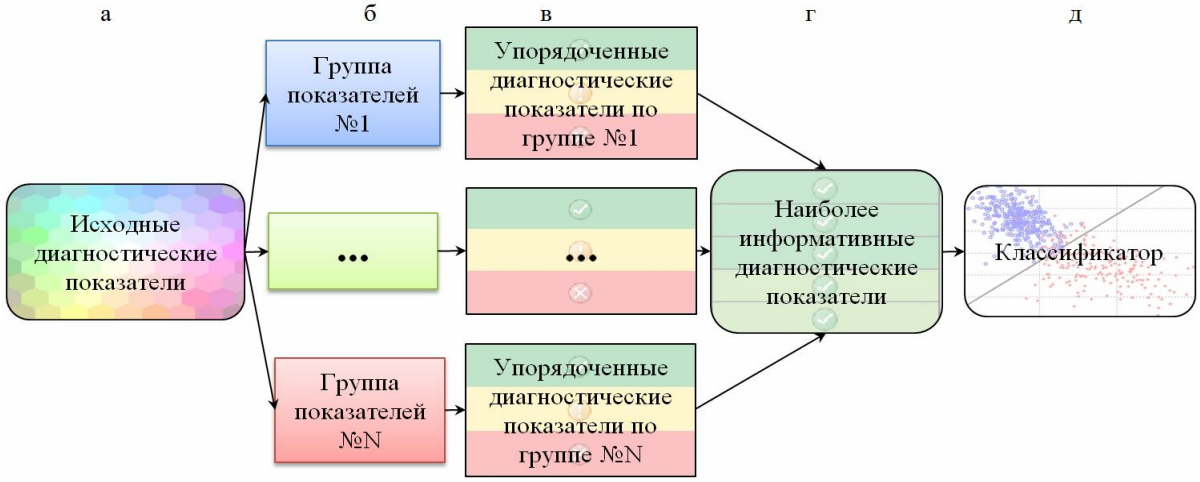


Рисунок 3 – Схема обработки медицинских данных (**а** – блок формирования множества ДП (при необходимости дополнение исходной группы показателей группой нормальных и граничных значений), **б** – блок разбиения групп на специфические подгруппы, **в** – блок ранжирования ДП, **г** – блок выделения информативных показателей, **д** – блок классификации)

Преимущество погруппового подхода заключается в поиске ДП в каждой отдельной группе, а не по всему списку показателей в целом. Это дает более качественную выборку показателей и как следствие, лучшее качество диагностики.

Использованный метод обработки данных. Пусть в пространстве m ДП заданы два множества точек: $X_A = \{a_i \in R^m | i = 1:n_A\}$ и $X_B = \{b_j \in R^m | j = 1:n_B\}$, из которых n_A соотносятся с объектами класса A, а n_B – с объектами класса B ($n_A + n_B = n$). Предположим, все множество точек состоит из объединения $x \in X$, $X = X_A \bigcup X_B$ множества точек $a_i = (a_{i1}, a_{i2}, ..., a_{im})$ и $b_j = (b_{j1}, b_{j2}, ..., b_{jm})$. Необходимо найти достаточно простой критерий разделения точек множеств X_A и X_B, т.е. указать правило идентификации точек множества X.

Рассмотрим матрицу X размерностью $n*m$, составленную из n строк (пациентов) m столбцов (ДП). Тогда каждый столбец $a_1 = (a_{11}, a_{21}, ..., a_{n_A 1})', a_2 = (a_{12}, a_{22}, ..., a_{n_A 2})', ..., a_m = (a_{1m}, a_{2m}, ..., a_{n_A m})'$ представляет собой численные значения $k-го$ ДП $(k = 1:m)$ из множества X_A. Соответственно, каждый столбец $b_1 = (b_{11}, b_{21}, ..., b_{n_B 1})', b_2 = (b_{12}, b_{22}, ..., b_{n_B 2})', ..., b_m = (b_{1m}, b_{2m}, ..., b_{n_B m})'$ представляет собой численные значения $k-го$ ДП из множества X_B (см. рис. 4).

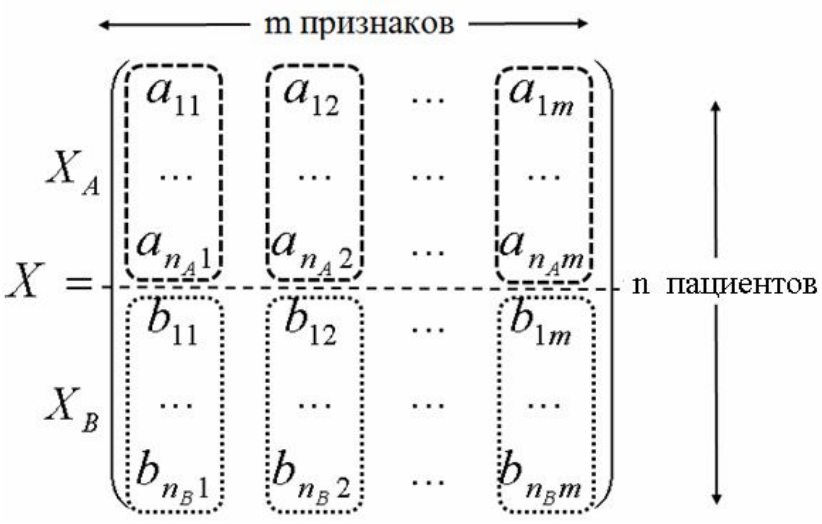


Рисунок 4 – Исходная матрица $n*m$ пациентов и ДП

Случайные величины a_{ik} и b_{jk}, соответствующие значению ДП для каждого из пациентов класса X_A и X_B, соответственно, аппроксимируются распределением Лапласа. По стандартным формулам математической статистики вычисляются сначала математическое ожидание, потом дисперсия величин a_{ik} и b_{jk}. На основе полученных данных строятся функции распределения этих величин $f^A{}_k$ и $f^B{}_k$.

Критерием значимости ДП (его «информативностью») служит описанный выше коэффициент перекрытия (Δ) или площадь перекрытия подграфиков плотностей распределения β. В дискретном случае Δ вычисляется аналогично: $\Delta = \sum_X \min[f^A(x), f^B(x)]$. Наименьшее значение площади S_k соответствует наиболее важному ДП (коэффициент перекрытия Δ помечен серым цветом на рисунке 1).

Иными словами, принцип обработки двух выборок из групп (доброкачественная/злокачественная) состоит в вычислении математического ожидания и дисперсии случайных величин для дальнейшей аппроксимации распределением Лапласа. На основе построенных графиков функций легко находятся площади перекрытия, а по ним определяется значимость (информативность) ДП.

Преимущество распределения Лапласа заключается в его аналитической форме, первообразная которого выражается формулой: $F(x) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(x - \beta)(1 - e^{-\frac{\sqrt{2}}{\sigma}|x-\beta|})$, где β – математическое ожидание, σ – среднеквадратичное отклонение. Упорядочивая полученный ряд из Δ по возрастанию, выбираем наиболее информативные показатели. По отобранным показателям будем строить оптимальную гиперплоскость $D(\alpha,x)=\alpha_0+\alpha_1 x_1+...+\alpha_m x_m$, разделяющую « A » и « B ».

Напомним, что врачами используются различные критерии качества классификации.

1. Чувствительность – это относительная частота отнесения истинно больного к группе больных.
2. Специфичность – это относительная частота отнесения истинно здорового к группе здоровых.
3. Безошибочность – это относительная частота принятия безошибочных решений, как по отношению к истинно больным, так и истинно здоровым.
4. Ложноотрицательные ответы (ошибки первого рода) – это относительная частота отнесения истинно больных к группе здоровых.
5. Ложноположительные ответы (ошибки второго рода) – это относительная частота отнесения истинно здоровых к группе больных.

Коллегиально было принято решение рассматривать безошибочность как критерий качества классификации. В соответствии с этим, будем считать, что гиперплоскость разделяет пациентов с ошибкой r, равной отношению числа неверно распознанных объектов $(\overline{n}_A+\overline{n}_B)$ к общему числу объектов N: $r=(\overline{n}_A+\overline{n}_B)/N$. Количество неверно классифицированных пациентов по выделенным в ходе ранжирования ДП должно быть минимальным.

Будем поворачивать гиперплоскость, минимизируя число неверно распознанных объектов. Для этого введем функцию ошибок:

$$r(\alpha)=\frac{n_A^k+n_B^k}{n},$$

где n_A^k, n_B^k – число неверно распознанных объектов на k-ом шаге приближения к минимуму функции $r(\alpha)$. Требуется построить последовательность векторов $\alpha^0,\alpha^1,\ldots$ такую, что

$$r(\alpha^k)\to\min_k, \quad k=0,1,\ldots\ r(\alpha^k)=\frac{n_A^k+n_B^k}{n}$$

Перейдем от пространства измеряемых показателей «x» к пространству весовых коэффициентов линейной дискриминантной функции $D(\alpha,x)$ и будем отмечать каждый новый набор ее коэффициентов точкой α и значением $r(\alpha)$. Соединяя точки α с одинаковым значением $r(\alpha)$, получим в двумерном случае линию равного уровня $r(\alpha)=const$, а в многомерном – некоторую поверхность. Требуется построить семейство поверхностей равного уровня $r(\alpha)$, различающихся на величину δ, (либо $k\delta$, где k – целое). Здесь δ – минимальная ошибка, обусловленная дискретностью функции $r(\alpha,x)$ и числом объектов n:

$$\delta=\frac{1}{n}100\,(\%)$$

и разработать алгоритм движения (перемещения) к минимуму, используя эту информацию для модификации известного метода оврагов [7].

Программное обеспечение. На основе представленных выше алгоритмов авторами статьи было разработано программное обеспечение (ПО) в пакете прикладных программ MATLAB и на языке C# для ранжирования ДП в медицинских БД и построения классификатора по наиболее информативным показателям. Код программы насчитывает 2346 строк. Детальное описание этого ПО представлено в статье [11]. Вычисления проводились на ПЭВМ Acer Aspire V3-571G с процессором Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz с ОЗУ 8,00ГБ на 64-разрядной операционной системе, процессором x64.

Результаты исследования и их обсуждение. В результате первого этапа работы программы (ранжирования) требуется отразить численные значения площадей перекрытия Δ. На основании меры сходства между двумя выборками a_{ik} и b_{jk}, $k=\overline{1,N}$, строятся функция распределения Лапласа $f_k^A(x)$ и $f_k^B(x)$ (принципы получения выборок и их характеристики описаны выше в разделе «Методы исследования»); выделяются характерные случаи перекрытия графиков функций плотности распределения; вычисляются значения Δ для каждого случая.

В таблице 1 представлены названия ДП и сравнены соответствующие им значения коэффициентов перекрытия, аппроксимированных распределением Гаусса $\Delta_{Гаусс}$ и Лапласа $\Delta_{Лаплас}$ – в виде упорядоченного по уменьшению степени информативности списка показателей, причем разделенных по трем подгруппам. В двух крайних столбцах стрелочками  «вниз» помечены коэффициенты перекрытия с наименьшими значениями, соответствующие наиболее информативным показателям; стрелочками "вверх" – неинформативные.

Таблица 1 – **Результаты ранжирования ДП для БД РМЖ-569**

$\Delta_{Гаусс}$	**Показатели формы ядра клетки**		$\Delta_{Лаплас}$
0,1863	Max_Вогн_участки_конт._28	Max_Вогн_участки_конт._28	0,1544
0,1958	Вогн_участки_контура_8	Вогн_участки_контура_8	0,1602
0,3147	Степень_вогнутости_7	Степень_вогнутости_7	0,2440
0,3712	Max_Степень_вогнутости_27	Max_Степень_вогнутости_27	0,2837
0,4289	Max_Плотность_26	Max_Плотность_26	0,3355
0,4369	Плотность_6	Плотность_6	0,3393
0,5842	Max_Симметрия_29	Max_Симметрия_29	0,4918
0,6329	Max_Гладкость_25	Max_Гладкость_25	0,5098
0,6422	Ст.откл._Выгнутость_18	Ст.откл._Выгнутость_18	0,5185
0,6794	Наихудший_Фракт_размер_30	Гладкость_5	0,5671
0,6877	Гладкость_5	Наихудший_Фракт_размер_30	0,5925
0,7134	Ст.откл._Степ_вогнут_17	Симметрия_9	0,6021
0,7176	Симметрия_9	Ст.откл._Степ_вогнут_17	0,6379
0,7505	Ст.откл._Плотность_16	Ст.откл._Плотность_16	0,6415
0,8158	Ст.откл._Фракт_размер_20	Ст.откл._Фракт_размер_20	0,8351
0,8265	Ст.откл._Симметрия_19	Ст.откл._Симметрия_19	0,8672
0,9397	Ст.откл._Гладкость_15	Ст.откл._Гладкость_15	0,9050
0,9439	Фракт_размер_10	Фракт_размер_10	0,9557
	Показатели размера ядра клетки		
0,1947	Max_Размер_24	Max_Размер_24	0,1554
0,1960	Max_Радиус_21	Max_Радиус_21	0,1604
0,2031	Ст.откл._Размер_14	Ст.откл._Размер_14	0,1812
0,2557	Размер_4	Размер_4	0,2009
0,2690	Радиус_1	Радиус_1	0,2112
0,3569	Ст.откл._Радиус_11	Ст.откл._Радиус_11	0,2924
	Показатели текстуры ядра клетки		
0,5947	Max_Текстура_22	Max_Текстура_22	0,4712
0,6338	Текстура_2	Текстура_2	0,5102
0,9037	Ст.откл._Текстура_12	Ст.откл._Текстура_12	0,9262
	Объединенный показатель размера и формы ядра клетки		
0,1869	Max_Периметр_23	Max_Периметр_23	0,1542
0,2507	Периметр_3	Периметр_3	0,1983
0,3453	Ст.откл._Периметр_13	Ст.откл._Периметр_13	0,2869

Для всех исходных ДП также были подсчитаны средние значения, максимальные (наихудшие) значения, стандартное отклонение. Для каждого пациента анализировалось по 10–20 клеток (выбиравшихся случайным образом из числа клеток, не имевших перекрытий с другими клетками на микрофотографиях) – для вычисления среднего значения и средней квадратической ошибки каждой переменной. А также три наибольшие или наихудшие значения были усреднены для определения наихудшего среднего значения.

В таблице 1 ДП представлены уже в упорядоченном виде по степени уменьшения информативности. Так, например, ДП *радиус* представлен тремя величинами: *Радиус_1* (среднее значение), *Ст.откл._Радиус_11* (средняя квадратическая ошибка), *Max_Радиус_21* (наибольшее).

Исходя из того, что коэффициент перекрытия $\Delta = \int_{R^n} \min[f^A(x), f^B(x)]\,dx$ является мерой сходства между двумя распределениями (это и есть формула, по которой считается «информативность»" ДП), следует считать, что он может варьироваться от 0 до 1. Если Δ=0, тогда перекрытие отсутствует (непересекающиеся распределения, свидетельствующие о наибольшей дифференцирующей способности ДП); если Δ = 1, то распределения идентичны – это, наоборот, свидетельствует об отсутствии дифференцирующей способности.

В таблице 1 четыре столбца. Два крайних – это численные значения коэффициента перекрытия по каждому ДП. Мы видим, что величина Δ для ДП «наибольшая доля вогнутых участков контура» x_{28} при ранжировании с аппроксимацией распределением Гаусса, равна 0.1863, а при аппроксимации распределением Лапласа – равна 0.1544. Это говорит о высокой дифференцирующей способности ДП x_{28}. В то же время Δ для показателя «средний фрактальный размер» x_{10} равны, соответственно, 0.9439 и 0.9557 – это свидетельствует о низкой дифференцирующей способности ДП x_{10}. Исходя из приведенных значений формируется выборка с минимальным Δ. На основе этой выборки строится экспресс-прогноз.

Основываясь на результатах расчетов, подробно описанных в статье [10], авторами был также исследован вопрос о влиянии вида распределения на точность ранжирования; проведено сравнение результатов ранжирования посредством коэффициента перекрытия при аппроксимации различными распределениями.

Результаты анализа данных, представленных в таблице 1 демонстрируют изменение коэффициента перекрытия от наименьшего значения $\Delta = 0.1542$ (соответствующему наиболее информативному показателю «максимальный периметр» x_{23}) до наибольшего значения $\Delta = 0.9557$ (соответствующему наименее информативному показателю «фрактальный размер» x_{10}). Приведенный пример наглядно иллюстрирует экономичность ранжирования ДП на основе коэффициента перекрытия при аппроксимации распределением Лапласа. По результатам видно, что информативные показатели совпадают при аппроксимации распределением Гаусса и Лапласа. Однако алгоритм $\Delta_{1\_Лаплас}$ более экономичен по времени разработки алгоритма, его программирования, а также времени выполнения вычислений – ввиду простой аналитической формы первообразной (по сравнению с $\Delta_{2\_Гаусс}$).

В результате работы программы с помощью алгоритма $\Delta_{1\_Лаплас}$ были выделены наиболее информативные показатели: наихудший показатель текстуры x_{22}, наибольшая площадь x_{24} и наибольшая доля вогнутых участков контура x_{28}, что показано на рисунке 5.

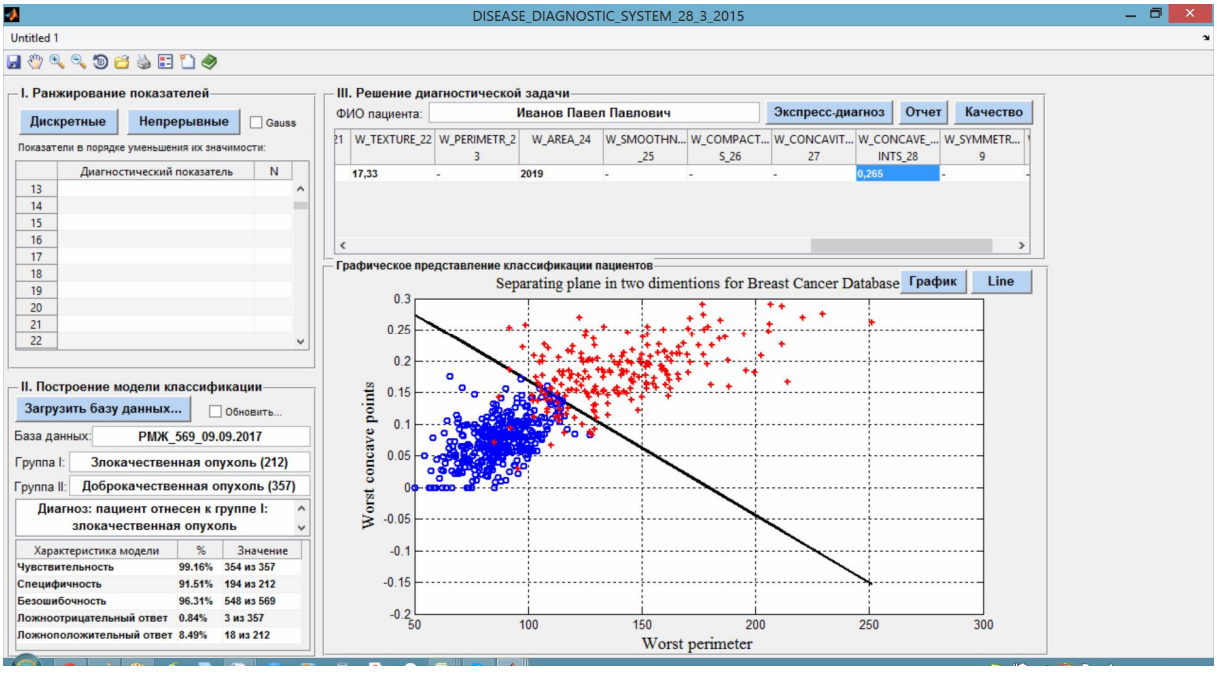


Рисунок 5 – Результаты экспресс-диагностики по наиболее информативным ДП

Выбор критерия по конкретным вариантам для отдельных объектов является важным вопросом в методах фильтрации, но в то же время не существует выраженного аналитически жесткого критерия отбора. В общем случае либо выбирается заданное число ДП, либо поиск продолжается до тех пор, пока мера оценки для всех оставшихся невыбранных показателей не перейдет пороговое значение. В нашем случае были выбраны ДП – «лидеры» с наименьшим коэффициентом перекрытия в каждой группе (в таблице 1 эти показатели выделены синей заливкой различной плотности). Были выбраны именно эти три показателя, т.к. соотношение «количество ДП/качество диагностики» оказалось оптимальным при

трех ДП. Если уменьшать количество ДП, то качество, соответственно, падает. Если увеличивать до четырех, пяти и т.д., то качество возрастает, но не до 100 %. 96,31 % по трём ДП – это достаточно хороший результат, исходя из того, что процент правильно распознанных пациентов по всем ДП составил 96,5 %. Потеря качества классификации при переходе от использования всех ДП к трём наиболее информативным составляет всего 0,19 %, тогда как число показателей сокращается с 30 до 3, т.е. в 10 раз. Следовательно, при незначительной потере качества классификации исходная размерность пространства показателей существенно уменьшается, это и позволяет ограничиваться первыми тремя показателями.

В разработанной авторами программе при диагностике нового пациента мы вводим лишь эти три показателя, а не все. Сразу после нажатия кнопки «Экспресс-прогноз» [11] результат будет представлен в строке «Безошибочность»: 548 из 569 пациентов (96,31 %) классифицированы верно (см. рис. 6).

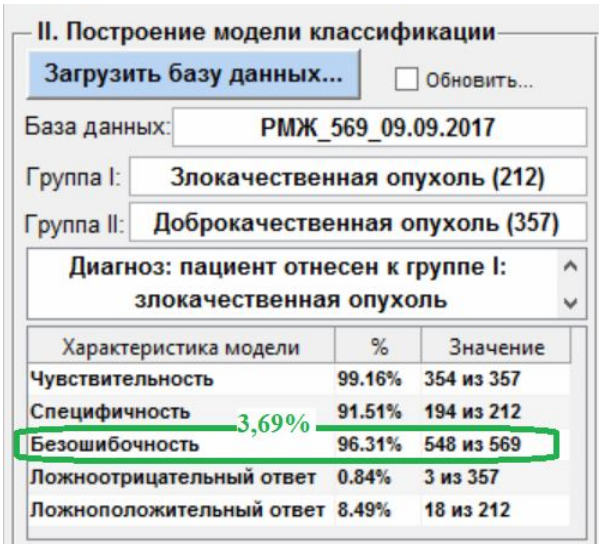


Рисунок 6 – Результаты диагностики по информативным показателям

В итоге процент ошибочно классифицированных снимков снижен до $r = 3.69\%$ – уровня качества диагностики, полученного другими, более сложными методами.

Сравнение результатов при разных подходах к ранжированию. Сравним результаты применения предложенного метода по точности классификации с другими методами, результаты работы которых доступны для тестируемой БД РМЖ-569 (см. табл. 2). При сравнении подходов к ранжированию видно, что качество классификации выше при применении погруппового подхода (3,69 %), нежели смешанного (5,27 %), изначально предложенного в работе [8]. Стоит заметить, что авторы используемой в статье БД – Dr. William H. Wolberg, W. Nick Street & Olvi L. Mangasarian [31] (США) использовали метод полного перебора и получили идентичное качество диагностики с ошибкой классификации равной 3,69 %.

Таблица 2 – **Сравнение результатов точности классификации при пяти разных подходах к ранжированию**

Метод ранжирования показателей	Набор диагностических показателей	Процент ошибок
О.Л. Мангасарян, полный перебор	Max_Размер_24, Max_Гладкость_25, Текстура_2	3.69%
Лаплас, погрупповой подход	Max_Вогн_участки_конт._28, Max_Размер_24, Max_Текстура_22	3.69%
Гаусс, погрупповой подход	Max_Вогн_участки_конт._28, Max_Размер_24, Max_Текстура_22	3.69%
Лаплас, смешанный подход	Max_Периметр_23, Max_Размер_24, Max_Вогн_участки_конт._28,	5.27%
Гаусс, смешанный подход	Max_Периметр_23, Max_Размер_24, Max_Вогн_участки_конт._28,	5.27%

Сравнительный анализ трудозатрат при использовании Δ с аппроксимацией распределением Лапласа. При количестве ДП n и количестве пациентов N временная сложность алгоритмов ранжиро-

вания методом фильтрации определяется как $T = O(nN)$. Предложенный в данной статье алгоритм более экономичен, сопоставление его с аналогами в графической форме показано выше на рисунке 1. Исходя из того, что трудозатраты обычно связывают с деятельностью людей, мы фокусируем внимание на экономичности алгоритма на этапах его разработки и программирования. При этом применение аппроксимации распределением Лапласа обеспечивает лучшее удобство и выигрыш по времени вычислений по сравнению с распределением Гаусса.

На рисунке 1 также приводится графическое сопоставление экономичности и временной сложности алгоритмов ранжирования с использованием Δ, первые три из которых уже апробированы на БД по онкологии, и их результаты работы доступны для сравнения [8, 32]. В связи с этим возникает вопрос о выборе среди них наименее трудозатратного. Покажем, что предложенный авторами метод (подход) является именно таковым. Временн***а***я сложность и экономичность разработки алгоритма ранжирования с помощью коэффициента перекрытия при аппроксимации распределением Лапласа $\Delta_{1\_Лаплас}$, Гаусса $\Delta_{2\_Гаусс}$, треугольниками $\Delta_{3\_Треугольники}$ и лучшим $\Delta_{4\_Лучшее}$ минимальны при $\Delta_{1\_Лаплас}$: $T_{1\_Лаплас} < T_{2\_Гаусс} < T_{3\_Треугольники} < T_{4\_"Лучшее"}$. Ввиду отсутствия аналитической формулы, показывающей нарастание трудозатрат аналогов по сравнению с предложенным методом, приведем эвристическое обоснование экономичности алгоритма $\Delta_{1\_Лаплас}$.

1. Простая аналитическая форма первообразной распределения Лапласа, зависящая лишь от двух параметров – масштаба σ и сдвига β – является преимуществом перед аппроксимацией распределением Гаусса, первообразная в явном виде у которого отсутствует. Экономичность разработки, включающая в себя количество машинных операций при вычислениях по программе, проще благодаря наличию первообразной.

2. При аппроксимации треугольниками мы вынуждены вычислять значения их высот и оснований, выбирая их таким образом, чтобы площади самих треугольников равнялись единице. Метод аппроксимации треугольниками $\Delta_{3\_Треугольники}$ в результате дает относительно точные значения, но, все же, последовательность показателей в ранжировании нарушается [8].

3. Можно возразить, что для получения более точной аппроксимации необходимо воспользоваться программным обеспечением [11]. Оно в результате обработки выборок a_i и b_j вычисляет наилучшую функцию плотности распределений и подбирает многочисленные параметры. Это смотрится как более точный метод, и может стать предметом дальнейшего исследования. Однако аппроксимация распределением Лапласа имеет меньшую сложность вычислений и более высокую скорость программирования алгоритма. При этом качество решения нашей задачи (ранжирования ДП методом фильтрации через сравнение Δ) не ухудшается.

4. Более того, в результатах сравнительного анализа в Табл. 2 показано, что при использовании алгоритма ранжирования с помощью коэффициента перекрытия при аппроксимации распределением Лапласа $\Delta_{1\_Лаплас}$ мы получаем то же качество диагностики по трем наиболее информативным показателям, что и при полном переборе, которым воспользовались наши американские коллеги-эксперты в области обработки БД по онкологии (процент правильно классифицированных составил 96,31%; ошибка $r = 100\% - 96,31\% = 3,69\%$ [31]). Полный перебор наиболее прост для программирования и в результате дает наилучший набор ДП. Однако он не экономичен с точки зрения объемов вычислений.

5. Если бы алгоритм $\Delta_{1\_Лаплас}$ действительно был бы неприемлем для ранжирования ДП, то мы бы не получили идентичные результаты по сравнению с алгоритмом $\Delta_{2\_Гаусс}$; не получили бы процент ошибки $r = 3,69\%$, соответствующий полному перебору. Следовательно, использование алгоритма $\Delta_{1\_Лаплас}$ является рациональным для ранжирования.

Заключение. 1. Предложенный алгоритм ранжирования ДП на основе коэффициента перекрытия при аппроксимации распределением Лапласа $\Delta_{1\_Лаплас}$ дал результат, идентичный по сравнению с другими методами, отраженными выше в таблице 2. Такой результат достигается с меньшими трудозатратами, т.к. степень сложности реализованного алгоритма $\Delta_{1\_Лаплас}$ ниже, чем его функциональных аналогов.

2. Описанные алгоритмы легли в основу программного обеспечения для ранжирования ДП и построения классификатора, которое было разработано авторами. При этом был значительно упрощен алгоритм ранжирования – за счет аналитической формы первообразной функции распределения Лапласа.

3. Также алгоритм $\Delta_{l\_Лаплас}$ может заинтересовать разработчиков «медицинских» приложений для любых смартфонов (в телефонах нет таких ОС), поддерживающих работу на платформах Windows или Android.

4. Итак, основное достоинство предложенного метода ранжирования – экономичность, простота реализации и идентичное с аналогами качество диагностики. Предложенный метод ранжирования может быть простым и удобным решением для постановки экспресс-диагноза по малому набору диагностических показателей.

Список литературы

1. Брумштейн Ю. М. Анализ и управление энергобезопасностью деятельности медицинских учреждений / Ю. М. Брумштейн, Д. А. Захаров, И. А. Дюдиков // Прикаспийский журнал: управление и высокие технологии. – 2015. – №1. – С. 44-58 (http://hi-tech.asu.edu.ru/?articleId=844).

2. Брумштейн Ю. М. Системный анализ направлений и особенностей информатизации сферы здравоохранения России / Ю. М. Брумштейн, Е. В. Скляренко, А. С. Мальвина // Прикаспийский журнал: управление и высокие технологии. – 2013. – №4(24). – С. 73-86 (http://hi-tech.asu.edu.ru/files/4(24)/73-86.pdf).

3. Бенинг В. Е. Об использовании распределения Стьюдента в задачах теории вероятностей и математической статистики / В. Е. Бенинг, В. Ю. Королев // Теория вероятностей и ее применения. – 2004. – № 3. – С. 417-435.

4. Габидулин Э. М. Лекции по теории информации / Э. М. Габидулин, Н. И. Пилипчук. – М.:МФТИ, 2007. – 214 с.

5. Горбунова А. А. Пат. 2013615968 Российская Федерация, Статистический анализ интервальных наблюдений одномерных непрерывных случайных величин "Интервальная статистка 5.1" / Горбунова А. А., Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н., Рогожников А. П., Чимитова Е. В.; заявитель и патентообладатель НГТУ. – № 2013612140; опубл. 21.03.2013. – Реестр программ для ЭВМ.

6. Золотухин И. В. Двухкомпонентное многомерное распределение Лапласа / И. В. Золотухин // Вестник новгородского государственного университета. – 2010. – № 68. – С. 60–64.

7. Калиткин, Н.Н. Численные методы / Н.Н. Калиткин. – М.: Наука, 1978. – 512 с.

8. Кокорина А. В. Оптимизационный подход в задачах математической диагностики: дис. канд. физ.-мат. наук: защищена 24.11.04 / Кокорина Анастасия Владимировна. – М.: РГБ, 2005. – 118 с.

9. Мальвина А. С. Автоматизация, диспетчеризация и информатизация высокотехнологичных медучреждений как средство повышения эффективности их работы / А. С. Мальвина, Ю. М. Брумштейн, Е. В. Скляренко, А. Б. Кузьмина // Прикаспийский журнал: управление и высокие технологии. – 2014. – №1(25). – С. 122-138 (http://hi-tech.asu.edu.ru/?articleId=785).

10. Трояножко О. А. Анализ влияния выбора вида распределения на точность ранжирования диагностических показателей / О. А. Трояножко, И. Д. Колесин // Научно-технический вестник Поволжья. –2013. – №6. – С. 457–461.

11. Трояножко О. А. Двухэтапная экспертная система диагностики заболеваний / О. А. Трояножко, И. Д. Колесин // Журнал "Известия Юго-Западного государственного университета" Серия Управление, вычис-лительная техника, информатика. Медицинское приборостроение. – 2017. – №1(22) – С. 82-88.

12. Araya-Ajoy Y. G., Characterizing behavioural 'characters': an evolutionary framework / Y. G. Araya-Ajoy, N. J. Dingemanse // Proceedings of the Royal Society B-Biological Sciences. – 2014. – Vol. 281(1776). doi:10.1098/rspb.2013.2645

13. Bache K. U. CI Machine Learning Repository / K. Bache, M. Lichman. – Irvine, CA: University of California, School of Information and Computer Science, 2013. – Режим доступа: http://archive.ics.uci.edu/ml, свободный.

14. Barnhill S. Gene selection for cancer classification using support vector machines / S. Barnhill, I. Guyon, V. Vapnik // Machine learning, 2002. – Vol. 46. –pp. 389–422.

15. Bradley E. L. Overlapping Coefficient. / E. L. Bradley // Encyclopedia of statistical Sciences S. Kotz and N. L. Johnson (Eds.), Wiley: New York, 1985. – vol. 6. – pp. 546-547.

16. Chen B. Speech Enhancement Using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling / B. Chen, P. Loizou // Proc IEEE ICASSP, 2005. – pp. 1097-1100.

17. Duch W. Filter methods / W. Duch // Feature Extraction: Foundations and Applications Studies in Fuzziness & Soft Computing, 2006, Springer. – pp. 89–117.

18. Inman H. F. Behavior and Properties of the Overlapping Coefficient as a Measure of Agreement Between Distributions / H. F. Inman. – University of Alabama in Birmingham, School of Joint Health Sciences, 1984. – 416 pp.

19. Inman H. F. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities / H. F. Inman, E. L. Bradley // Communications in Statistics - Theory and Methods, 1989. – Vol. 18. – pp. 3851-3874.

20. John G. H. Wrappers for feature subset selection / G. H. John and R. Kohavi // Artifcial intelligence, 1997. – 97. – pp. 273–324.

21. Kozubowski T. J. Asymmetric Laplace laws and modeling financial data / T. J. Kozubowski, K. Podgorski // Math. Comput. Modelling. – 2001. – Vol.34. – pp. 1003–1021.

22. Kotz S. The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering and finance / S. Kotz, T. J. Kozubowski, K. Podgorski. Birkhäuser Boston, USA, 2001. – 349 pp.

23. Lourens S. Bias in Estimation of a Mixture of Normal Distributions. / S. Lourens, Y. Zhang, J. D. Long, J. S. Paulsen // J Biomet Biostat. – 2013. – Vol. 4:179. doi: 10.4172/2155-6180.1000179

24. Okubo T. On the distribution of extreme winds expected in Japan / T. Okubo, N. Narita. National Bureau of Standards Special Publication, 1980. – 12 pp.

25. Pengyi Y. Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning / Y. Pengyi, W. Liu, B. Z. Bing // Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, 2013. – Vol. 7818. – pp. 544–555.

26. Rachev S. Stable Paretian Models in Finance / S. Rachev, S. Mittnik. Wiley, 2000. – 874 pp.

27. Samawi H. A Test of Symmetry Based on the Kernel Kullback-Leibler Information with Application to Base Deficit Data / H. Samawi, R. L. Vogel // Biom Biostat Int J . – 2016. – Vol. 3(2). (DOI:10.15406/bbij.2016.03.00060)

28. Silva-Fortes C. Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups / C. Silva-Fortes, M. A. Turkman, L. Sousa // BMC Bioinformatics . – 2012 . – 13(147). DOI: 10.1186/1471-2105-13-147

29. Sun Y. Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications / Y. Sun // IEEE Trans. Pattern Anal. Mach. Intell., 2007. – Vol. 29(6). – pp.1035–1051.

30. Tibshirani R. Regression shrinkage and selection via the lasso / R. Tibshirani // Journal of the Royal Statistical Society Series B (Methodologi-cal), 1996. – Vol.58. – pp. 267–288.

31. Wolberg W. H. Breast cytology diagnosis via digital image analysis. / W. H. Wolberg, W. N. Street, O. L. Mangasarian // Analytical and Quantitative Cytology and Histology, 15(6):396–404, 1993.

32. Wolberg W. H. Computerized Diagnosis of Breast Fine-Needle Aspirates / W. H. Wolberg, W. N. Street, O. L. Mangasarian // The Breast J. – 1997 .– Vol. 3. – pp. 77–80.

References

1. Brumshteyn, Yu. M., Zakharov D.A., Dudikov I.A. Analiz i upravlenije energobezopasnost'ju deyatelnostju meditsinskikh uchrezhdenij [Power safety analysis and management of medical institution activity]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Control and High Technologies], 2015, no.1, pp. 44-58 (http://hi-tech.asu.edu.ru/?articleId=844).

2. Brumshteyn, Yu. M., Sklyarenko Ye. V., Malvina A. S., Aksenova Yu. Yu., Kuzmina A. B. Sistemnyy analiz napravleniy i osobennostey informatizatsii sfery zdravookhraneniya Rossii [The system analysis of trends and characteristics of healthcare informatization of Russia]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Control and High Technologies], 2013, no.4(24), pp. 73-86 (http://hi-tech.asu.edu.ru/files/4(24)/73-86.pdf).

3. Bening B. E., Korolev V. Yu. Ob ispolzovanii raspredeleniya Studenta v zadachakh teorii veroyatnostey i matematicheskoy statistiki [About application of Student distribution in probability theory and mathematical statistics]. *Teoriya veroyatnostey i ee primeneniya* [Probability theory and its applications], 2004, no.3, pp. 417-435

4. Gabidulin E. M., Pilipchuk N. I. Lekcii po teorii informatsii [Lectures about information theory]. M.: MFTI, 2007, 214 p.

5. Gorbunova A. A. Pat. 2013615968 Russian Federation, Statistichsky analiz intervalnyh nabludeniy odnomernyh nepreruvnykh sluchajnykh velichin "Intervalnaya statistika 5.1" [Statistical analysis of interval observations of one-dimensional continuous random variables], Lemeshko B. Yu., Lemeshko S. B., Postovalov S. N., Rogozhnikov A. P., Chimitova E. V., MGTU, no. 2013612140, publ. 21.03.2013, Register of computer programs.

6. Zolotukhin I. V. Dvukhkomponentnoe mnogomernoe raspredelenie Laplasa [Two-component multidimensional Laplace distribution]. *Vestnik novgorodskogo gosudarstvennogo universiteta* [Bulletin of novgorod state university], 2010, no.68, pp. 60-64.

7. Kalitkin, N.N. Numerical methods / N.N. Kalitkin. – M.: Nauka, 1978. – 512 p.

8. Kokorina A. V. Optimizatsionnyy podhod v zadachakh matematicheskoy diagnostiki: dissertation [Optimization approach in problems of mathematical diagnostics]. M.:RGB, 2005, 118 p.

9. Malvina, A. S., Brumshteyn Yu. M., Sklyarenko E. V., Kuzmina A. B. Avtomatizatsiya, dispetcherizatsiya vysokotekhnologichnykh meduchrezhdenij kak sredstvo povysheniya effiktivnosti ikh raboty [Automation and computerization of scheduling high-tech medical facilities as a means of improving their performance]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Control and High Technologies], 2014, no.1(25), pp. 122-138 (http://hi-tech.asu.edu.ru/?articleId=785).

10. Troyanozhko O. A., Kolesin I. D. Analiz vliyaniya vibora vida raspredeleniya na tochnost' ranzhirovaniya diagnosticheskikh pokazateley [Analysis of influence of distribution type on the accuracy of features ranking] *Nauchno-tekhnicheskiy vestnik Povolzhja* [Technical scientific bulletin of *Povolzhje*], 2013, no.6, pp. 457–461.

11. Troyanozhko O. A., Kolesin I. D., Dvukhetapnaya expertnaya sistema diagnostiki zabolevaniy [Two-stage medical expert system]. *Journal "Izvestiya Ugo-Zapadnogo gosudarstvennogo universiteta" seriya Upravlenie, vichislitelnaya technika, informatika. Meditsinskoe priborostroenie.* [Journal "Bulletin of South-East State University" Series Management, computing, informatics. Medical instrument-making], 2014, no. 1(22), pp. 82-88.

12. Araya-Ajoy, Y. G., Characterizing behavioural 'characters': an evolutionary framework / Y. G. Araya-Ajoy, N. J. Dingemanse // Proceedings of the Royal Society B-Biological Sciences. – 2014. – Vol. 281(1776). doi:10.1098/rspb.2013.2645

13. Bache, K. UCI Machine Learning Repository / K. Bache, M. Lichman. – Irvine, CA: University of California, School of Information and Computer Science, 2013. – Режим доступа: http://archive.ics.uci.edu/ml, свободный.

14. Barnhill, S. Gene selection for cancer classification using support vector machines / S. Barnhill, I. Guyon, V. Vapnik // Machine learning, 2002. – Vol. 46. –pp. 389–422.

15. Bradley, E. L. Overlapping Coefficient. / E. L. Bradley // Encyclopedia of statistical Sciences S. Kotz and N. L. Johnson (Eds.), Wiley: New York, 1985. – vol. 6. – pp. 546-547.

16. Chen, B. Speech Enhancement Using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling / B. Chen, P. Loizou // Proc IEEE ICASSP, 2005. – pp. 1097-1100.

17. Duch, W. Filter methods / W. Duch // Feature Extraction: Foundations and Applications Studies in Fuzziness & Soft Computing, 2006, Springer. – pp. 89–117.

18. Inman, H. F. Behavior and Properties of the Overlapping Coefficient as a Measure of Agreement Between Distributions / H. F. Inman. – University of Alabama in Birmingham, School of Joint Health Sciences, 1984. – 416 pp.

19. Inman, H. F. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities / H. F. Inman, E. L. Bradley // Communications in Statistics - Theory and Methods, 1989. – Vol. 18. – pp. 3851-3874.

20. John, G. H. Wrappers for feature subset selection / G. H. John and R. Kohavi // Artifcial intelligence, 1997. – 97. – pp. 273–324.

21. Kozubowski, T. J. Asymmetric Laplace laws and modeling financial data / T. J. Kozubowski, K. Podgorski // Math. Comput. Modelling. – 2001. – Vol.34. – pp. 1003–1021.

22. Kotz, S. The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering and finance / S. Kotz, T. J. Kozubowski, K. Podgorski. Birkhäuser Boston, USA, 2001. – 349 pp.

23. Lourens, S. Bias in Estimation of a Mixture of Normal Distributions. / S. Lourens, Y. Zhang, J. D. Long, J. S. Paulsen // J Biomet Biostat. – 2013. – Vol. 4:179. doi: 10.4172/2155-6180.1000179

24. Okubo, T. On the distribution of extreme winds expected in Japan / T. Okubo, N. Narita. National Bureau of Standards Special Publication, 1980. – 12 pp.

25. Pengyi, Y. Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning / Y. Pengyi, W. Liu, B. Z. Bing // Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, 2013. – Vol. 7818. – pp. 544–555.

26. Rachev, S. Stable Paretian Models in Finance / S. Rachev, S. Mittnik. Wiley, 2000. – 874 pp.

27. Samawi, H. A Test of Symmetry Based on the Kernel Kullback-Leibler Information with Application to Base Deficit Data / H. Samawi, R. L. Vogel // Biom Biostat Int J . – 2016. – Vol. 3(2). (DOI:10.15406/bbij.2016.03.00060)

28. Silva-Fortes, C. Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups / C. Silva-Fortes, M. A. Turkman, L. Sousa // BMC Bioinformatics . – 2012 . – 13(147). DOI: 10.1186/1471-2105-13-147

29. Sun, Y. Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications / Y. Sun // IEEE Trans. Pattern Anal. Mach. Intell., 2007. – Vol. 29(6). – pp.1035–1051.

30. Tibshirani, R. Regression shrinkage and selection via the lasso / R. Tibshirani // Journal of the Royal Statistical Society Series B (Methodologi-cal), 1996. – Vol.58. – pp. 267–288.

31. Wolberg, W. H. Breast cytology diagnosis via digital image analysis. / W. H. Wolberg, W. N. Street, O. L. Mangasarian // Analytical and Quantitative Cytology and Histology, 15(6):396–404, 1993.

32. Wolberg, W. H. Computerized Diagnosis of Breast Fine-Needle Aspirates / W. H. Wolberg, W. N. Street, O. L. Mangasarian // The Breast J. – 1997 .– Vol. 3. – pp. 77–80.

РЕДАКЦИОННЫЙ КОММЕНТАРИЙ К СТАТЬЕ

Статья посвящена актуальной теме, содержит оригинальный материал, текст неплохо структурирован.

Однако по работе целесообразно сделать ряд замечаний.

1. Название статьи выглядит шире, чем ее содержание. Уже в начале введения авторы значительно ограничивают рассматриваемую предметную область, в отношении которой будут применяться рассматриваемые подходы (методы). Это разделение опухолей молочной железы на «доброкачественные» и «злокачественные» – на основе обработки изображений ядер клеток на микрофотографиях срезов кусочков тканей, взятых при пункциях. Как следствие, рассматриваемый авторами набор диагностических показателей (ДП) для опухолей молочной железы является достаточно специфичным и относится только к ядрам клеток. Поэтому диагностику классов опухолей в данной статье приходится рассматривать как иллюстративный пример для предлагаемых авторами подходов. Для этого класса задач показано, что предложенные подходы эффективны. Однако, насколько обосновано распространение этих подходов на другие классы задач (с совершенно другими наборами ДП) – этот вопрос требует, видимо, дополнительного анализа (изучения), проверки на экспериментальном материале.

2. Ту же конкретную задачу (анализ классов новообразований в молочной железе) можно решать и другими методами. Например, по результатам анализа формы опухоли, полученной при проведении рентгеновской компьютерной томографии (это не требует проведения пункций); с использованием микрофотографий препаратов – по долям клеток, находящихся в стадии деления (алгоритмы автоматического обнаружения таких клеток на изображениях сейчас активно разрабатываются) и др.

3. Вопросами определения релевантности информативных признаков для медицинских интеллектуальных систем поддержки принятия диагностических решений занимались многие отечественные исследователи и довольно давно. В самой статье ничего не сказано ни про разработанные ими методы, ни про самих исследователей. В списке литературы имеется только одна работа отечественных авторов, которая имеет весьма косвенное отношение к выявлению информативности признаков для систем поддержки принятия решений в медицине.

4. Сама идея определения релевантности информативных признаков по пересечению гистограмм классов не совсем корректна. Вспомним классический пример, когда в декартовых координатах рисуют прямоугольник, делят его диагональю на две области – класса. При этом гистограммы признаков (координат точек, представляющих оцениваемые объекты, в областях-классах) практически полностью перекрываются, а разделение объектов (с помощью указанной выше диагонали) равно 100 %.

5. Авторами в явной форме ничего не сказано об используемом методе классификации (разделения объектов на классы). Есть ссылка на источник № 11. Но в нем описывается только внешняя оболочка программного обеспечения и примеры результатов его работы. Можно догадаться, что речь в статье идет о дискриминантном анализе. Однако возможны и другие методы построения разделяющих гипер-плоскостей для разделения объектов на классы.

УДК [007.2+004.94]:616

МОДЕЛИРОВАНИЕ МОРФОЛОГИЧЕСКИХ ОБРАЗОВАНИЙ НА РЕНТГЕНОГРАММАХ ГРУДНОЙ КЛЕТКИ В ИНТЕЛЛЕКТУАЛЬНЫХ ДИАГНОСТИЧЕСКИХ СИСТЕМАХ МЕДИЦИНСКОГО НАЗНАЧЕНИЯ[1]

Статья поступила в редакцию 26.08.2017, в окончательном варианте – 11.10.2017.

Кудрявцев Павел Сергеевич, Юго-Западный государственный университет, 305004, Российская Федерация, г. Курск, ул. Челюскинцев, 19Б

аспирант, e-mail: 79pavel97@mail.ru

Кузьмин Александр Алексеевич, Юго-Западный государственный университет, 305004, Российская Федерация, г. Курск, ул. Челюскинцев, 19Б

кандидат технических наук, доцент, ORCID https://orcid.org/0000-0001-7980-0673 SCOPUS https://www.scopus.com/authid/detail.uri?authorId=36142241500, ResearcherID http://www.researcherid.com/rid/F-8405-2013, e-mail: ku3bmin@gmail.com, https://elibrary.ru/author_profile.asp?id=616342

Савинов Денис Юрьевич, Юго-Западный государственный университет, 305004, Российская Федерация, г. Курск, ул. Челюскинцев, 19Б

аспирант, e-mail: marina-savinova-93@mail.ru

Филист Сергей Алексеевич, Юго-Западный государственный университет, 305004, Российская Федерация, г. Курск, ул. Челюскинцев, 19Б

доктор технических наук, профессор, ORCID https://orcid.org/0000-0003-1358-671X, SCOPUS https://www.scopus.com/authid/detail.uri?authorId=6603139063, ResearcherID http://www.researcherid.com/rid/O-4610-2015, e-mail: SFilist@gmail.com, https://elibrary.ru/author_profile.asp?id=251980

Шаталова Ольга Владимировна, Юго-Западный государственный университет, 305004, Российская Федерация, г. Курск, ул. Челюскинцев, 19Б

кандидат технических наук, доцент, ORCID https://orcid.org/0000-0002-0901-9272, SCOPUS https://www.scopus.com/authid/detail.uri?authorId=24477712800, ResearcherID http://www.researcherid.com/rid/C-3687-2015, e-mail: shatolg@mail.ru, https://elibrary.ru/author_profile.asp?id=673680

Дифференциальная диагностика онкологии и пневмонии по изображению на рентгенограмме грудной клетки (ИРГК) является сложной задачей. Для ее решения необходимы репрезентативные обучающие выборки, полученные по ИРГК пациентов с этими заболеваниями, которые затем используются в классификаторах ИРГК. Для этого необходимо селектировать ИРГК по типу морфологических образований (МО) с определенной дислокацией или с сочетанными патологиями. Это весьма сложный и трудоемкий процесс. Поэтому предложено моделировать морфологические образования необходимые для формирования обучающих выборок для настройки нейронных сетей, предназначенных для классификации рентгеновских снимков. Согласно предложенному методу построения моделей МО, осуществлялся статистический анализ спектров Уолша в многомасштабных окнах. Идея формирования модели МО, связанного с нозологией ω_ℓ, состоит в следующем. На текущем изображении ИРГК выделяется область (прямоугольная) *L1×L2,* в которой строится модель МО заданного класса. Затем осуществляется «подгонка» спектра каждого окна M_k, образованного вокруг текущего пикселя к эталонному окну. Учитывая принятую структуру окна *M*: 16x16; 32x32 и 64x64 пикселя, вокруг каждого пикселя ИРГК образуются окна трех типов. Заполнение области *L1×L2* пикселями, соответствующими выбранной модели, ведется с самого большого по размеру окна M_3. Определяем спектральные коэффициенты в этом окне для текущего пикселя *m* окна *L1×L2* и минимизируем евклидово расстояние между текущим спектром и эталоном. При достижении удовлетворительной «подгонки» спектров, переключаемся на окно $k = 2$ и также оптимизируем спектральное соотношение. После этого переходим к k=1. Эта процедура может быть выполнена в цикле, пока функционал, характеризующий качество «подгонки», не станет приемлемым для всех *k.* В результате проведенных исследований предложен метод моделирования морфологических образований на рентгенограммах грудной клетки. Метод позволяет формировать обучающие выборки для классификаторов рентгеновских снимков по заданной патологии.

Ключевые слова: рентгеновский снимок, модель морфологического образования, спектр Уолша, пиксель, классификация изображения, окно анализа, алгоритм построения модели, показатели качества сегментации изображений

[1]Исследования выполнены при финансовой поддержке РФФИ в рамках научного проекта № 16-07-00164 а.