

Визуальная аналитика в цифровой биомедицине на примере качества диагностики рака молочной железы

О.А. Трояножко, И.Д. Колесин

Аннотация—В работе рассмотрена оптимизация классификация больных раком молочной железы. Решается вопрос отнесения опухоли к доброкачественной или злокачественной. Работая в рамках стандартных подходов медицинской диагностики, авторами было проведено два этапа: упорядочивание по степени значимости диагностических показателей, а потом – непосредственное проведение диагностики по выделенным показателям. Диагностические показатели – гистохимия пунктатов опухоли молочной железы – получены с помощью тонкоигольной аспирационной пункции. Далее происходит анализ изображений ядер клеток на микрофотографиях, полученных из срезов кусочков ткани. Цифровая обработка снимков ядер клеток даёт возможность определить тип ядра клетки. Ввиду того, что полное отделение двух множеств невозможно даже по полному набору диагностических показателей, решается задача о минимизации количества ошибочно диагностированных пациентов.

Для регистрации количества неверно диагностированных пациентов использована дискретная функция ошибок. Основываясь на исследовании геометрических особенностей поведения дискретной функции ошибок в пространстве параметров линейной дискриминантной функции авторами разработан и запрограммирован алгоритм движения к минимуму, визуально показывающий в 2D и 3D динамику изменения значений функции ошибок. В итоге было показано, что для базы данных рака молочной железы РМЖ-569 достаточно надежную диагностику с точностью 94,73% можно провести, ограничиваясь двумя диагностическим показателям.

Ключевые слова—автоматизированные системы, биомедицина, диагностика, научная визуализация, точность классификации, оптимизация.

I. ВВЕДЕНИЕ

Данная статья посвящена визуальной аналитике качества диагностики на примере распространенной патологии - рака молочной железы. По оценкам ВОЗ, рак молочной железы является самым распространенным раковым заболеванием среди женщин во всем мире – на него приходится 16% всех случаев заболевания раком среди женщин [1].

Классификация опухоли в молочной железе решается и другими не требующими проведения пункций методами. Например, по результатам анализа формы опухоли, полученной при проведении

маммографии, рентгеновской компьютерной томографии. Сегодня также динамично развиваются алгоритмы автоматического нахождения на изображениях микрофотографий препаратов долей клеток, находящихся в стадии деления.

Методы визуальной аналитики, предложенные авторами в данной работе, продемонстрированы на классификации опухолей при анализе изображений ядер клеток на микрофотографиях, полученных из срезов кусочков ткани. Рассматриваемые методы подтвердили свою высокую эффективность для этого класса задач. Используемый в исследовании набор диагностических показателей может быть задействован в диагностике ядер клеток аналогичной патологии. Для решения вопроса обоснованности генерализации этих подходов на другие классы задач с отличными диагностическими показателями планируется вспомогательный анализ с тестированием на соответствующих базах данных.

Вопросами разработки систем поддержки принятия решений в медицине с предварительным анализом информативности диагностических признаков занималась целая плеяда отечественных исследователей: Ю.И. Журавлев [2,3], А.Я. Червоненкис [4], В.Н. Вапник [5,6,7], Н.Г. Загоруйко [8], А.Д. Букин [9], А.А. Генкин [10], В.Ф. Демьянов [11], В.А. Дюк [12,13] и др. Отыскание оптимальной гиперплоскости с наилучшим качеством разделения двух классов лежит в основе проблемы классификации данных. Ряд приемов решения этой задачи сводится к построению гиперплоскости $d(x, \alpha) = (x, \alpha)$ (здесь x – вектор диагностических показателей; α – вектор коэффициентов линейной функции), секущей два множества с наименьшим числом неверно распознанных объектов.

На сегодняшний момент разработано достаточное количество детерминистских подходов в теории распознавания образов:

- искусственные нейронные сети с алгоритмом обучения персептронов;
- метод опорных векторов для случая линейной разделимости и линейной неразделимости классов;
- кластерный анализ с разными алгоритмами расстановки центров кластеров;

- классификация с помощью функций расстояния между векторами-образами и классами;
- метод потенциальных функций, позволяющий с помощью несложного алгоритма оценивать вес объектов обучающей выборки;
- методы дискриминантного анализа с нахождением линейной и обобщенной решающей функции.

Несмотря на моду нейросетей в задачах распознавания образов, подходы с линейной классификацией являются простыми в реализации и эффективными. В данном исследовании используется модификация дискриминантного анализа с поиском линейной решающей функции для разделения объектов на классы.

Конечной целью исследования является разработка математической модели классификации с последующим созданием программного комплекса, включающего визуальный прием оптимизации и предназначенного для повышения эффективности медицинской экспресс-диагностики.

Исходя из цели исследования были решены следующие задачи:

В рамках общепринятого подхода проведена минимизация признакового пространства для последующей диагностики не по всему набору признаков, а по наиболее информативным из них.

Решена задача отыскания оптимальной разделяющей функции с наивысшим качеством диагностики. При этом качество диагностики определяется критерием безошибочности – это относительная частота принятия безошибочных решений как по отношению к истинно больным, так и истинно здоровым пациентам.

Для решения задачи классификации предложена дискретная функция ошибок, которая равна отношению числа неверно распознанных объектов к общему числу объектов. Далее предлагается выявлять единственный минимум функции ошибок во множестве локальных.

Поведение линейной разделяющей дискриминантной функции вида $d(x, \alpha) = (x, \alpha)$ зависит от изменения её весовых коэффициентов α и переменной x . Функция ошибок в свою очередь зависит от точности классификации объектов. Работая в условиях недоступности аналитического вида функции ошибок, задачу поиска минимума всё чаще решают методом визуальной аналитики, что ускоряет и направляет поиск.

Покажем, при каких условиях качество разделения сохраняется на постоянном уровне и насколько оно чувствительно к изменению коэффициентов линейной разделяющей дискриминантной функции. Продemonстрируем возможную степень отклонения от заданных коэффициентов, решая таким образом задачу визуального управления процессом оптимизации.

Актуальность визуально-аналитического подхода продиктована требованием создания более совершенных методов обработки данных.

II. ПОСТАНОВКА ЗАДАЧИ РАЗДЕЛЕНИЯ МНОЖЕСТВ

Пусть в пространстве m диагностических показателей есть два множества объектов: X_A и X_B , из которых n_A соотносятся с объектами класса A , а n_B – с объектами класса B ($n_A + n_B = n$). Пусть построена начальная гиперплоскость

$$D(\alpha^0, x) = \alpha_0^0 + \alpha_1^0 x_1 + \dots + \alpha_m^0 x_m,$$

разделяющая X_A и X_B с ошибкой классификации r^0 , равной отношению числа неверно распознанных объектов ($n_A^0 + n_B^0$) к общему числу объектов n :

$$r^0 = \frac{n_A^0 + n_B^0}{n} \quad (1).$$

Далее приступим к поворачиванию гиперплоскости, таким образом минимизируя число объектов, распознанных неверно. Для этого предложим функцию ошибок классификации:

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

где n_A^k, n_B^k – число неверно распознанных объектов на k -м шаге приближений к минимуму функции ошибок $r(\alpha)$. В результате требуется построить последовательность векторов $\alpha^0, \alpha^1, \dots$ такую, что

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

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

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

и разработать алгоритм движения к минимуму, используя эту информацию для модификации известного метода оврагов [14].

В данной работе усовершенствован один из известных способов визуализации градиентного спуска. Для исключения возможного перескакивания с одного склона оврага на другой и обратно, что существенно замедляет сходимость, предлагается, построив семейство поверхностей равного уровня, визуально определять направление ко «дну оврага». Визуальный контроль обеспечивает более быстрый спуск.

III. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Материалом исследования служили базы медицинских данных из библиотек с открытым доступом передовых университетов. В данной работе приведены результаты на материале базы данных рака молочной железы РМЖ-569 [15]. Всего 569 пациентов с опухолью молочной железы и 30 непрерывных диагностических показателя. По результатам анализа характеристик ядер клеток проводится разделение по двум группам: доброкачественная опухоль (357 пациентов) или злокачественная (212 пациентов).

Начальная точка α^0 находится методом минимизации средне-квадратической ошибки, что аналогично первому шагу известной процедуры Хо-Кашьяпа [16-21]. В качестве метода поиска минимума функции $r(\alpha)$ выбран метод покоординатного спуска, позволяющий войти в область фактического нахождения минимума. Для движения в этой области к точке минимума избран метод градиента.

При поиске минимума можно рассмотреть семейство поверхностей равного уровня, или одну из них. Это ограничение необходимо при поиске минимума.

Учитывая геометрические особенности поведения дискретной функции ошибок в пространстве параметров линейной дискриминантной функции, авторы наметили цель статьи в виде выявления этих особенностей, разработки и программирования алгоритма движения к минимуму, визуально показывающего в 2D и 3D динамику изменения значений функции ошибок на примере базы данных рака молочной железы.

IV. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Описанные в предыдущем разделе методы мы применили к базе данных рака молочной железы (РМЖ-569). В качестве критериальной была использована дискретная функция ошибок, регистрирующая количество неверно диагностированных пациентов. Далее мы сократили общее число диагностических показателей до двух наиболее информативных: наибольший периметр - показатель X_{23} и наибольшая доля вогнутых участков контура - показатель X_{28} , см. Табл. 1.

Численные значения функции ошибок для диагностики рака молочной железы.

Параметр	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010	0.011	0.012	0.013	0.014	0.015	0.016	0.017	0.018	0.019	0.020
0.0	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.1	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.2	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.3	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.4	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.5	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.6	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.7	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.8	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
0.9	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
1.0	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26

Оптимальные значения находятся в центре

Специально для журнала International Journal of Open Information Technologies

Табл. 1. Табличное представление изолаты в 2D: окрестности оврага вблизи минимума.

Результат построения поверхностей равного уровня в координатном пространстве $\alpha_0, \alpha_1, \alpha_2$, где свободный член α_0 , показан ниже на рис.1. Визуально – это три "кристалла", вложенные друг в друга. Эти три поверхности с постоянным уровнем ошибок $r(\alpha^k) = \bar{r}^k = \text{const}$ ($k=1,2,3$) представлены, соответственно, значениями $\bar{r}^1=7,56\%$, $\bar{r}^2=6,15\%$ и $\bar{r}^3=5,27\%$.

Примечание. Сравнение изолаты с кристаллом – издержки недостатка визуализированных данных в целях ускорения проведения исследования. Иными словами, про большем количестве точек *угловатый кристалл* превращается в весьма *сглаженную оболочку*. Именно ввиду подобия куриному яйцу с наличием вложенности: сердцевина – желтка, средней оболочки – белка, а также внешней – скорлупы, – одними из первых названий данных вложенных поверхностей равного уровня были «яйцо» или «матрешка». Причем линии равного уровня, названные изолатами (см. Приложение), по определению не пересекаются, т.к. их смысловая нагрузка – наглядная визуализация дискретного изменения значений функции ошибок ввиду скачкообразного перехода с одного значения на соседнее с заданным и уникальным для каждой базы шагом $\delta=1/n$ (где n – количество пациентов). Заметим, что нельзя сделать итерацию более дробной, т.к. нельзя «раздробить» одного пациента на части.

Поверхности равного уровня дают представление о том, как выглядит трехмерное размытое образование в пространстве $\alpha_0, \alpha_1, \alpha_2$ линейной дискриминантной функции $D(x, \alpha) = (x, \alpha)$. После поворачивания разделяющей гиперплоскости она принимает следующий вид: $D(x, \alpha) = -2,5 + 0,0135x_{23} + 6,8168x_{28}$, где $\alpha_0 = -2,5$, $\alpha_1 = 0,0135$ и $\alpha_2 = 6,8168$ – параметры.

Построение поверхностей равного уровня была реализована с использованием метода линейных оболочек в среде динамического моделирования MATLAB/ Simulink.

Самая крайняя наружная поверхность, которой соответствует функция ошибок $\bar{r}^1=7,56\%$ включает в себе область, содержащую минимум. На рисунке 1 она выглядит как "кристалл" с максимальной площадью поверхности - семейство вложенных одна в другую трех поверхностей равного уровня r_1, r_2, r_3 в пространстве параметров дискретной функции ошибок $r(\alpha_0, \alpha_1, \alpha_2)$.

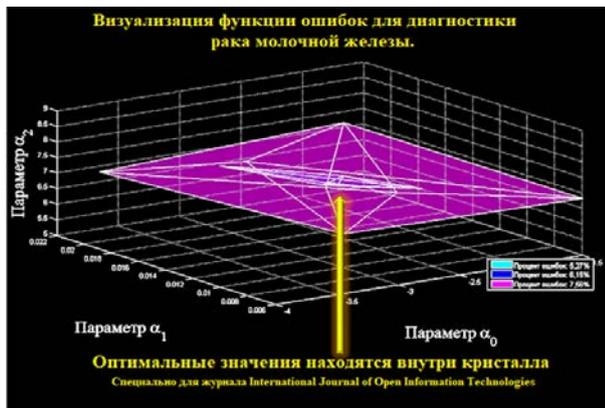


Рис. 1. Визуализация изолаты в 3D

В соответствии с определением размытого многомерного образования, предложенным ранее в статье [22], нами получена некая сущность, характеризующаяся последовательностью поверхностей равного уровня $r(\alpha^k) = \bar{r}^k = \text{const}$ ($k=1,2,3$) таких, что $r^1(\alpha) > r^2(\alpha) > r^3(\alpha)$, при $\bar{r}^1 \leq \bar{r}^0$, где r^0 – начальная ошибка. В соответствии с введенным определением *экстремальной области размытого многомерного образования*, нами также получен некий набор точек α локального минимума дискретной функции $r(\alpha) = 5,27\%$. Примечательно, что мы получили эту минимальную ошибку классификации двумя способами: с помощью дискретной функции ошибок и используя суррогатную непрерывную функцию. Злокачественную опухоль с точностью 94,73% удалось отделить от доброкачественной, используя лишь три диагностических показателя. Таким образом, мы вышли на достойный уровень качества диагностики, работая в сокращенном признаковом пространстве и, учитывая, что точность классификации по всем диагностическим показателям составила всего лишь на 1,77% больше, а именно – 96,5%.

V. ИЗОЛАТА – ЛИНИЯ ОШИБОК РАВНОГО УРОВНЯ

В соответствии с общепринятым определением, *изолиния*, или линия уровня (функции) (от др.-греч. ἴσος — «равный») — условное обозначение на карте, чертеже, схеме или графике, представляющее собой линию, в каждой точке которой измеряемая величина сохраняет одинаковое значение. *Изолинии* — способ представления скалярной функции от двух переменных на плоскости. Приведем общеизвестные термины, полученные аналогичным образом:

- изотерма — изолиния одинаковых температур;
- изохора — изолиния одинаковых объемов;
- изобара — изолиния одинакового давления;
- изобара в термодинамике — график изобарного процесса;
- изобара в метеорологии — линия на карте, обозначающая область с той или иной границей давления.

Аналогичным образом вводим термин *изолата*.

Изолата в теории распознавания образа – линия ошибок равного уровня. За термином «изолата» — линия (функция) ошибок равного уровня (от др.-греч.

ἴσος — «равный» и λάθος — ошибка). Изолата показывает равные значения функции ошибок на плоскости двух ее параметров. Термин изолата заменяет объединение нескольких слов: за этим термином стоит иной уровень внимания к некоему явлению, другой уровень его понимания и развития знания о нем. Данный факт подтверждается тем, что при исследовании изолат были выявлены общие для всех баз особенности - математическое явление, исследование которого таким образом еще не проводилось.

Проводя исследование в пространстве параметров линейной дискриминантной функции, мы выявляли специфику структуры многомерных образований, присущих различным базам данных. Охарактеризуем выявленные общие черты при исследовании поверхностей равного уровня. Напомним, в пространстве 2D при $\alpha = (\alpha_0, \alpha_1)$ размытое многомерное образование соотносится с оврагом, а экстремальная область – с его дном, см. рис. 2.

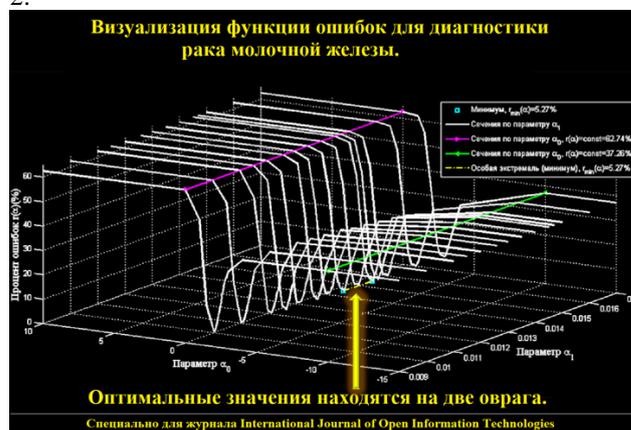


Рис. 2. Визуализация изолаты в 3D: вид на овраг вблизи минимума с "направляющими".

В пространстве 3D при $\alpha = (\alpha_0, \alpha_1, \alpha_2)$ размытое многомерное образование соотносится с линейными оболочками – вложенными кристаллами, а экстремальная область – с его сердцевиной. Минимум в 2D достигается при движении по дну оврага, и в 3D при движении к сердцевине внутреннего кристалла.

Проводя визуальную аналитику мы добились построения оптимального классификатора, который обладает максимальным диагностическим качеством.

Как во внешнем виде оврага, так и в ориентации кристаллов есть одна особенность – и овраг, и кристалл вытянуты. Это говорит о том, что изменение значения одних коэффициентов больше влияет на изменение значения функции ошибок, чем других.

Данное явление имеет общие для всех баз особенности, соответственно, оно может быть исследовано, просчитано.

Введение обобщающего термина «изолата» дало возможность визуально исследовать и формализовать поведение функции ошибок. Исследование изолат может быть полезно, поскольку сокращает издержки на опыты, экономит трудозатраты, даёт возможность создавать более экономные вычислительные алгоритмы.

Введённый в текущем исследовании термин "изолата" – это очевидная формализованное обобщение уже известных дефиниций, введенных ранее в статье [23].

Ввиду неудобства многократного использования в тексте длинного устоявшегося словесного сочетания, обладающего совершенной четкой дефиницией, а также по причине объективной потребности мышления – постепенного повышения уровня абстрагирования – коллегиально было решено ввести краткий термин, повышающий качество и скорость восприятия текста.

Что такое «изолата»? Напомним, что здесь речь идет о поведении дискретной функции ошибок (или цифровой и визуализированной модели оценки качества диагностики пациента), которая представлена в виде дискретной совокупности точек. Причем набор точек с равными величинами соответствует отдельной изолате. В табличном виде характеристика изолаты – это уникальное значение функции ошибок g , соответствующее совокупности двух координат (x, y) на плоскости и уникальное значение функции ошибок g , соответствующее совокупности трёх координат (x, y, z) в пространстве. При визуализации изолата представляет собой совокупность линий в случае 2D, напоминающие карты высот, и «вложенные» кристаллы в случае 3D наподобие матрешки.

Наличие дефиниции у термина «изолата», отсутствие дублирования существующих терминов, несение специальной дополнительной информации – всеми этими аргументами мы доказали такое обозначение явления, которое является термином – независимым элементом понятийного поля. Так же выше была приведена дефиниция понятия «изолата» для акцентирования наличия новизны в рассматриваемых понятиях.

Одним из оснований введения термина «изолата» послужило то, что для данного явления – формализуемого поведения линии ошибок равного уровня – до сих пор отсутствовало понятие. Данное понятие созрело в результате исследования закономерного для различных баз данных поведения линии ошибок равного уровня, позволяющее ввести понятийный механизм, содержащий в себе объяснение отдельных явлений.

После сравнения этого механизма с множеством понятий, присущим разделу науки распознавания образов в части биомедицины, производится соотнесение начальной понятийной схемы именно на предмет наличия или отсутствия дефиниций. Если для объяснения описываемого явления дефиниций нет, тогда имеем право внедрять обособленный термин и его определение.

Вводу термина «изолата» предшествовал скрупулёзный анализ множества используемых в разделе науки распознавания образов в части биомедицины понятий, что привело к обнаружению пробела в виде отсутствия соответствия термина для обнаруженного явления.

Обобщая вышеприведенные аргументы касательно термина «изолата», мы видим, что введенный термин обладает собственной дефиницией. При отсутствии

дефиниции можно было бы говорить не о ином термине, а лишь о ином обозначении. Далее, сформировав иную дефиницию, было сформировано иное понятие для иной сущности.

Причем эта дефиниция подверглась экспертной оценке на предмет совместимости с уже имеющимися дефинициями, относящимися к разделу науки распознавания образов в части биомедицины. В результате оценки был сделан вывод об уместности введения иной дефиниции с выбором подходящего термина. Введение термина произошло не в свободной форме, а строго соблюдая терминологическую системность в рамках филологической грамотности, в виде лексической единицы информационно-поискового языка.

VI. ОСОБЕННОСТИ ПОВЕДЕНИЯ ИЗОЛАТ

1. Визуально легко проследить, при каких численных изменениях коэффициентов разделяющей гиперплоскости изолата принимает минимальное значение, т.е. максимально возрастает качество классификации.
2. При минимальных значениях изолаты коэффициенты имеют наименьшую степень свободы своей варибельности, иными словами, при малейшей попытке изменения оптимальных коэффициентов изолата уходит из области минимума. В то же время, на периферии изолата не столь чувствительная к вырыванию коэффициентов.
3. При приближении изолаты к оптимуму динамика изменения значений коэффициентов характеризуется меньшей вариативностью.
4. Для того, чтобы изолата сохраняла свое оптимальное значение, набор значений коэффициентов нужно менять крайне аккуратно и с наименьшим шагом.
5. Около «проблемных» областей изолаты нужно исследовать более тщательно, уменьшая шаг изменения коэффициентов.
6. Наихудшее качество разделения характеризуется максимальным постоянным значением изолаты. Выход изолаты с постоянными максимальными значениями на оба берега оврага в случае 2D по смыслу идентичен выходу на крайнюю внешнюю оболочку в случае 3D.
7. Соответственно, оптимальные изолаты находятся на дне оврага в случае 2D или в сердцевине кристаллов в случае 3D. Причем изолаты на дне оврага находятся чуть выше нулевой отметки. Также и изолаты в сердцевине безуспешно стремятся превратиться в точку, но не могут этого достичь по причине неразделимости двух классов.
8. Нахождение оптимальных значений коэффициентов разделяющей гиперплоскости – минимальных значений изолат – в общем случае характеризуются движением ко дну

оврага в случае 2D и движением к сердцевине вложенных кристаллов в случае 3D.

Представленные выше особенности визуального поведения изолат ранее были неочевидными при численном исследовании, т.к. функция ошибок не имеет аналитического вида и является весьма непростой для нахождения глобального минимума различными оптимизационными подходами. Однако удалось заметить, что функция ошибок $r(\alpha)$ в 2D имеет одинаковые черты с функциями Букина, являющиеся почти фрактальными с острым изрезанным дном (рис. 3 и 4). Например, в области задания функции $x_1 \in [-15, -5]$, $x_2 \in [-3, 3]$ $f_4(x_1, x_2)$ и $f_6(x_1, x_2)$ определяются следующим образом:

$$f_4(x_1, x_2) = 100x_2^2 + 0,01|x_1 + 10|; f_{\min}(-10,0) = 0,$$

$$f_6(x_1, x_2) = 100\sqrt{|x_2 - 0,01x_1^2 + 0,01|x_1 + 10|}; f_{\min}(-10,0) = 0.$$

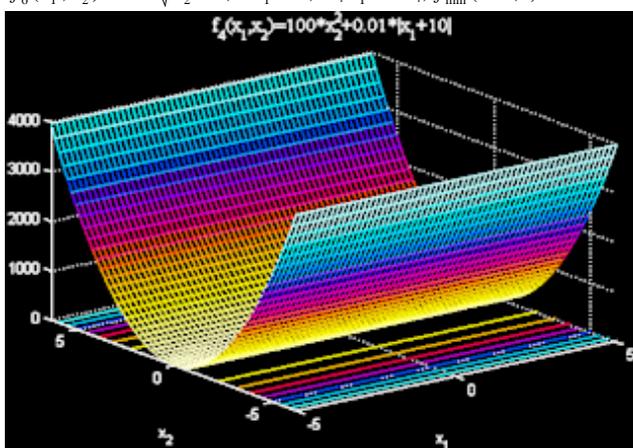


Рис. 3. Графическое представление функций Букина $f_4(x_1, x_2)$ и $f_6(x_1, x_2)$ в трехмерном пространстве (x_1, x_2, f) .

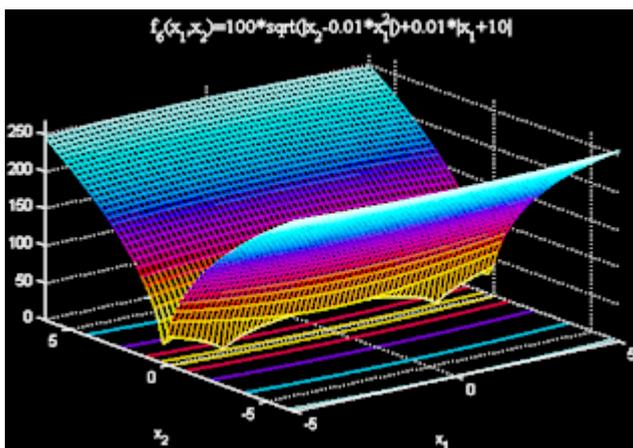


Рис. 4. Графическое представление функций Букина $f_4(x_1, x_2)$ и $f_6(x_1, x_2)$ в трехмерном пространстве (x_1, x_2, f) .

VII. ЗАКЛЮЧЕНИЕ

При исследовании медицинских баз данных методом визуализации авторам удалось выявить ряд

особенностей поведения критериальной функции ошибок.

С помощью предложенного визуального метода можно представлять критериальную функцию ошибок и анализировать ее характеристики как на плоскости, так и в трехмерном пространстве, решая таким образом задачу поиска оптимальной гиперплоскости для разделения двух множеств.

Формализуемое поведение линий ошибки равного уровня соотнесено с известной моделью распознавания образов; по аналогии с общепринятыми терминами для линий равного уровня введен термин изолата.

Визуальная аналитика качества диагностики ввиду относительной несложности используемого алгоритма целесообразна для использования на практике как одного из вспомогательных методов ранней экспресс-диагностики.

БЛАГОДАРНОСТИ

Авторы благодарны специалистам ВМедА им. С.М. Кирова за конструктивное обсуждение и практические рекомендации.

БИБЛИОГРАФИЯ

- [1] ВОЗ <https://www.who.int/topics/cancer/breastcancer/ru/index1.html> (дата обращения: 10.02.2019).
- [2] Журавлев, Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлев, В.В. Рязанов, О.В. Сенько. – М.: Фазис, 2006. – 176 с.
- [3] Бондаренко, Н.Н. Алгоритм выбора конъюнкций для логических методов распознавания / Н.Н. Бондаренко, Ю.И. Журавлев // Ж. вычисл. матем. и матем. физ. – 2012. – № 52. – С. 746-749.
- [4] Вапник, В.Н. Теория распознавания образов / В.Н. Вапник, А.Я. Червоненкис. – М.: Наука, 1974. – 416 с.
- [5] 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.
- [6] Vapnik, V. Estimation of Dependences Based on Empirical Data / V. Vapnik. – Springer, New York, 2006. – 505 pp.
- [7] Vapnik V. Statistical learning theory / V. Vapnik. – Wiley, New York, 1998. – 768 pp.
- [8] Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.
- [9] Bukin A.D. New Minimization Strategy for Non-Smooth Functions. Novosibirsk, 1997. 24 p.
- [10] Генкин, А.А. Новая информационная технология анализа медицинских данных: программный комплекс ОМИС / А.А. Генкин. – СПб.: Политехника, 1999. – 191 с.
- [11] Демьянов, В.Ф. Прогнозирование эффективности химиотерапии при лечении онкологических заболеваний / В.Ф. Демьянов, В.В. Демьянова, А.В. Кокорина, В.М. // Вестник Санкт-

- Петербургского университета, Прикладная математика. – 2006. – № 4. – с. 30-36.
- [12] Дюк В.А. Инструментальные средства интеллектуального анализа данных. – Изд-во РГПУ им. А.И. Герцена, 2012. – 161 с.
- [13] Дюк, В.А. Информационные технологии в медико-биологических исследованиях / В. Дюк, В. Эмануэль. – СПб.: Питер, 2003. – 528 с.
- [14] Калиткин, Н.Н. Численные методы / Н.Н. Калиткин. – М.: Наука, 1978. – 512 с.
- [15] Bache, K. UCI Machine Learning Repository / K. Bache, M. Lichman. – Irvine, CA: University of California, School of Information and Computer Science, 2013. Url: <http://archive.ics.uci.edu/ml>.
- [16] Ho, Y.-C. An algorithm for linear inequalities and its applications / Y.-C. Ho, R. Kashyap // IEEE Trans. Elec.Comp. – 1965. – Vol. 14. – pp. 683–688.
- [17] Hassoun, M.H. Ho-Kashyap Rules for Perceptron Training / M.H. Hassoun, J.Song // IEEE Transactions on Neural Networks. – 1992. – Vol. 3. – pp. 51–61.
- [18] Ho, Y.-C. A class of iterative procedures for linear inequalities / Y.-C. Ho, R. Kashyap // Journal of SIAM Control. – 1966. – Vol. 4. – pp. 112–115.
- [19] Lauer, F. Ho-Kashyap with early stopping vs soft-margin SVM for linear classifiers – an application / F. Lauer, M. Bentoumi, G. Bloch, G. Millrerioux // Advances in Neural Networks. – 2004. – Vol. 1. – pp. 524-530.
- [20] Leski, J. An ϵ -margin nonlinear classifier based on if-then rules / J. Leski // IEEE Transactions on Systems, Man and Cybernetics. – 2004. – Vol. 34. – pp. 68-76.
- [21] Leski, J. Ho-Kashyap classifier with generalization control / J. Leski // Pattern Recognition Letters. – 2003. – Vol. 24. – pp. 2281-2290.
- [22] Трояножко, О.А. Оптимизация диагностики рака молочной железы на основе дискретной функции ошибок / О.А. Трояножко, И.Д. Колесин // Журнал "Известия Юго-Западного государственного университета" Серия Управление, вычислительная техника, информатика. Медицинское приборостроение. – 2016. – №3(20) – С. 125-132 (из перечня ВАК, Url: https://www.swsu.ru/izvestiya/seriesivt/archiv/3_2016.pdf)
- [23] Трояножко, О.А. Визуализация процесса оптимизации параметров критериальной функции ошибок в задаче медицинской диагностики / И.Д. Колесин, О.А. Трояножко, П.П. Сивашенко // Научная визуализация. 2017. Т. 9. С. 26-40. (из БД SCOPUS, Url: <http://sv-journal.org/2017-4/03.php?lang=ru>).

Visual analytics in digital biomedicine on the example of the breast cancer diagnostics quality

O.A. Trojanozhko, I.D. Kolesin

Abstract—The optimization of the breast cancer diagnostics is considered in this paper. The issue of classifying a tumor as benign or malignant is being resolved. Histochemistry of breast tumor punctates is performed using a fine-needle aspiration puncture.

Next, a digital analysis of images of cell nuclei in micrographs obtained from sections of tissue pieces is performed. Working within the framework of the standard approaches of medical diagnostics, the authors carried out two stages: diagnostic features ranking was done at the first stage, and then patients classification was fulfilled at the second stage.

A discrete error function was used to register the number of incorrectly diagnosed patients. The authors studied the geometric characteristics of the behavior of a discrete error function in the parameter space of a linear discriminant function. We developed and programmed an algorithm for movement to a minimum, visually showing in 2D and 3D the dynamics of changes in the error function values.

As a result, it was shown that a sufficiently reliable diagnosis with an accuracy of 94.73% can be carried out based only by two diagnostic features.

Key words—automated system, biomedicine, diagnostics, scientific visualization, classification accuracy, optimization.

REFERENCES

- [1] VOZ <https://www.who.int/topics/cancer/breastcancer/ru/index1.html> (data obrashhenija: 10.02.2019).
- [2] Zhuravlev, Ju.I. Raspoznavanie. Matematicheskie metody. Programmaja sistema. Prakticheskie primenenija / Ju.I. Zhuravlev, V.V. Rjazanov, O.V. Sen'ko. – M.: Fazis, 2006. – 176 s.
- [3] Bondarenko, N.N. Algoritm vybora kon'junkcij dlja logicheskikh metodov raspoznavanija / N.N. Bondarenko, Ju.I. Zhuravlev // Zh. vychisl. matem. i matem. fiz. – 2012. – # 52. – S. 746-749.
- [4] Vapnik, V.N. Teorija raspoznavanija obrazov / V.N. Vapnik, A.Ja. Chervonenkis. – M.: Nauka, 1974. – 416 s.
- [5] 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.
- [6] Vapnik, V. Estimation of Dependences Based on Empirical Data / V. Vapnik. – Springer, New York, 2006. – 505 pp.
- [7] Vapnik V. Statistical learning theory / V. Vapnik. – Wiley, New York, 1998. – 768 pp.
- [8] Zagorujko, N.G. Prikladnye metody analiza dannyh i znaniy / N.G. Zagorujko.– Novosibirsk: IM SO RAN, 1999.– 270 s.
- [9] Bukin A.D. New Minimization Strategy for Non-Smooth Functions. Novosibirsk, 1997. 24 p.
- [10] Genkin, A.A. Novaja informacionnaja tehnologija analiza medicinskih dannyh: programnyj kompleks OMIS / A.A. Genkin. – SPb.: Politehnika, 1999. – 191 s.
- [11] Dem'janov, V.F. Prognozirovanie jeffektivnosti himioterapii pri lechenii onkologicheskikh zabolevanij / V.F. Dem'janov, V.V. Dem'janova, A.V. Kokorina, V.M. // Vestnik Sankt-Peterburgskogo universiteta, Prikladnaja matematika. – 2006. – # 4. – c. 30-36.
- [12] Djuk V.A. Instrumental'nye sredstva intellektual'nogo analiza dannyh. – Izd-vo RGPU im. A.I. Gercena, 2012. – 161 s.
- [13] Djuk, V.A. Informacionnye tehnologii v mediko-biologicheskikh issledovanijah / V. Djuk, V. Jemanujel'. – SPb.: Piter, 2003. – 528 s.
- [14] Kalitkin, N.N. Chislennye metody / N.N. Kalitkin. – M.: Nauka, 1978. – 512 s.
- [15] Bache, K. UCI Machine Learning Repository / K. Bache, M. Lichman. – Irvine, CA: University of California, School of Information and Computer Science, 2013. Url: <http://archive.ics.uci.edu/ml>.
- [16] Ho, Y.-C. An algorithm for linear inequalities and its applications / Y.-C. Ho, R. Kashyap // IEEE Trans. Elec.Comp. – 1965. – Vol. 14. – pp. 683–688.
- [17] Hassoun, M.H. Ho-Kashyap Rules for Perceptron Training / M.H. Hassoun, J.Song // IEEE Transactions on Neural Networks. – 1992. – Vol. 3. – pp. 51–61.
- [18] Ho, Y.-C. A class of iterative procedures for linear inequalities / Y.-C. Ho, R. Kashyap // Journal of SIAM Control. – 1966. – Vol. 4. – pp. 112–115.
- [19] Lauer, F. Ho–Kashyap with early stopping vs soft-margin SVM for linear classifiers – an application / F. Lauer, M. Bentoumi, G. Bloch, G. Millrerioux // Advances in Neural Networks. – 2004. – Vol. 1. – pp. 524-530.
- [20] Leski, J. An ϵ -margin nonlinear classifier based on if-then rules / J. Leski // IEEE Transactions on Systems, Man and Cybernetics. – 2004. –Vol. 34. – pp. 68-76.
- [21] Leski, J. Ho-Kashyap classifier with generalization control / J. Leski // Pattern Recognition Letters. – 2003. – Vol. 24. – pp. 2281-2290.
- [22] Trojanozhko, O.A. Optimizacija diagnostiki raka molochnoj zhelezy na osnove diskretnoj funkcii oshibok / O.A. Trojanozhko, I.D. Kolesin // Zhurnal "Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta" Serija Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie. – 2016. – #3(20) – S. 125-132 (iz perechnja VAK, Url: https://www.swsu.ru/izvestiya/seriesivt/archiv/3_2016.pdf)
- [23] Trojanozhko, O.A. Vizualizacija processa optimizacii parametrov kriterial'noj funkcii oshibok v zadache medicinskoj diagnostiki / I.D. Kolesin, O.A. Trojanozhko, P.P. Sivashhenko // Nauchnaja vizualizacija. 2017. T. 9. S. 26-40. (iz BD SCOPUS, Url: <http://sv-journal.org/2017-4/03.php?lang=ru>).