ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 22 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Б.М. Гавриков, М.Б. Гавриков,
И.М. Лебеденко, Н.В. Пестрякова

Статистический классификатор для определения области локализации онкопатологии по анализу крови

Рекомендуемая форма библиографической ссылки: Статистический классификатор для определения области локализации онкопатологии по анализу крови / Б.М. Гавриков [и др.] // Препринты ИПМ им. М.В.Келдыша. 2024. № 22. 15 с. https://doi.org/10.20948/prepr-2024-22
https://library.keldysh.ru/preprint.asp?id=2024-22

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Б.М. Гавриков, М.Б. Гавриков,
И.М. Лебеденко, Н.В. Пестрякова

Статистический классификатор для определения области локализации онкопатологии по анализу крови

Москва — 2024

Б.М. Гавриков, М.Б. Гавриков, И.М. Лебеденко, Н.В. Пестрякова

Статистический классификатор для определения области локализации онкопатологии по анализу крови

Описывается математический подход, позволяющий по данным анализа периферической крови онкопациента установить местонахождение опухоли. Для предварительной диагностики используется разработанный авторами статистический классификатор, основанный на полиномиальной регрессии и имеющий вероятностные оценки. Обучение классификатора осуществлялось на базах периферической крови онкологических больных мужского пола по семи различным системам организма. Предложен и реализован способ исследования структуры обучающего множества.

Ключевые слова: онкологическое заболевание, система организма, периферическая кровь, классификация, полиномиальная регрессия, обучающее множество

Boris Mikhailovich Gavrikov, Mikhail Borisovich Gavrikov, Irina Matveevna Lebedenko, Nadejda Vladimirovna Pestryakova

Statistical classifier for determining the localization area of oncopathology using a blood test

A mathematical approach is described that makes it possible to determine the location of the tumor based on the analysis of peripheral blood of a cancer patient. A statistical classifier developed by the authors is used for the preliminary diagnosis, based on polynomial regression and having probabilistic estimates. The classifier was trained on peripheral blood databases of male cancer patients in seven different body systems. A method for studying the structure of the training set is proposed and implemented.

Key words: cancer, body system, peripheral blood, classification, polynomial regression, training set

Оглавление

Введение

В данной публикации подведены первые итоги исследования о применении численного подхода для нахождения местоположения онкопатологии в организме человека по результатам анализа периферической крови [1-3].

На первый взгляд, используемые в медицинской диагностике высокотехнологичные методы перекрывают ставшую уже рутинной процедуру анализа крови. Однако даже компьютерная томография не дает точного ответа на вопрос относительно области локализации онкологии. Кроме того, доступ к таким дорогостоящим процедурам ограничен. В ожидании очереди на их проведение может быть потеряно время. В любом случае всегда лучше иметь в арсенале дополнительный подход, тем более весьма дешевый.

Исчерпаны ли возможности, которые предоставляет для первичной диагностики анализ крови? Какую именно полезную информацию практикующий врач способен извлечь из соответствующего набора данных? По сути, все сводится к выяснению, попадает ли каждый параметр крови в определенный для условно здоровых людей диапазон, причем последний время от времени меняется.

Кажется вполне логичным, что при принятии решения медики должны каким-то образом учитывать компенсаторные возможности человека. Но о них можно говорить, только если рассматривать весь набор параметров в качестве единого объекта. Сделать это проблематично даже для самого опытного врача.

В рамках изложенного мы решили выяснить, существует ли такая структурная взаимосвязь между различными характеристиками крови. Заметим, что ее наличие было показано нами в ранее проведенных исследованиях при оценке состояния здоровья человека (СЗЧ) для каждой из систем организма (СО), рассматриваемых по отдельности [4-10].

Подчеркнем, что настоящая работа описывает численный эксперимент: в рамках оригинального математического подхода на модельных задачах получены новые перспективные результаты. Они служат основанием для реализации предложенного метода на расширенных базах крови онкологических больных. Следующим возможным шагом является использование разработанной технологии в практической диагностике.

Данное исследование базируется на концепции крупнейших гематологов, согласно которой многие заболевания человека вносят изменения в состав его крови. При оценке СЗЧ гематологи предлагают использовать не менее пяти показателей периферической крови (из пальца) [11].

СЗЧ включает четыре градации: практически здоровые, начальные отклонения состояния здоровья, выраженные изменения, тяжелое заболевание (онкология).

Авторы ранее показали, что статистический метод классификации может успешно применяться для оценивания СЗЧ по данным лабораторного анализа периферической крови [4-10]. Человеческий организм представляется в виде совокупности СО: пищеварения, дыхания, урологической, эндокринной, гинекологической (для женщин), опорно-двигательного аппарата, печени и желчевыводящих путей, грудных желез (для женщин), центральной нервной системы и органов чувствительности.

Для каждой СО оценивание СЗЧ осуществлялось при помощи профильного классификатора, обученного на соответствующей выборке. Эти базы представляют собой совокупность наборов из восьми параметров периферической крови. Они были созданы на основе верифицированных диагнозов, поставленных большому количеству пациентов.

Для мужчин и женщин были разработаны различные классификаторы, поскольку диапазоны вариации показателей крови существенно зависят от пола.

К четвертой градации СЗЧ относится онкология. При ее диагностике не исключены ошибки относительно местонахождения опухоли. Кроме того, может возникнуть ситуация, когда для некоторого анализа крови ответом вышеописанного классификатора будет четвертый класс здоровья для нескольких СО.

Итак, можно ли по анализу крови онкологического больного определить область локализации патологии?

Для получения ответа на этот вопрос авторы разработали математическую модель и на ее основе построили статистический классификатор. Его обучение провели на наборах параметров периферической крови четвертого класса СЗЧ, относящихся ко всем имеющимся у мужчин семи различным СО. Тем самым были дополнены и обобщены результаты, полученные для двух, трех, четырех и пяти СО [1-3].

Проведено также изучение двух сопутствующих проблем. Как можно визуализировать векторные объекты восьмимерного пространства (по числу параметров крови)? Как соотносятся между собой фрагменты обучающей выборки, соответствующие каждому из указанных семи классов (по числу СО)? Подчеркнем, что в настоящей работе классами являются СО онкобольных мужского пола. В рамках поставленной задачи предложен и апробирован оригинальный способ исследования структуры обучающей выборки.

Метод классификации

Общепринятые обозначения и размерность используемых восьми показателей крови следующие: RBC [L^{-1}] – эритроциты, HGB [gL^{-1}] – гемоглобин, PLT [L^{-1}] – тромбоциты, WBC [L^{-1}] – лейкоциты, LIMPH [L^{-1}], [%]

– лимфоциты, GRAN [L^{-1}], [%] – гранулоциты (GRAN=NEUT+EOS+BASO, где NEUT[L^{-1}],[%] – нейтрофилы, EOS[L^{-1}],[%] – эозинофилы, BASO[L^{-1}],[%] – базофилы).

Рассматриваем K определенных перенумерованных СО человека, $1 \le k \le K$, K=7. Вводим вектор $v \in R^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови онкобольных, где N=8.

Отождествляем k-й элемент множества СО с базисным вектором из R^K: $e_k=(0\ldots1\ldots0)$, причем 1 находится на k-м месте, $1 \le k \le K$. Обозначим $Y=\{e_1,\ldots,e_K\}$.

Пусть существует $p_k(v)$ – вероятность того, что набор отнормированных показателей крови онкобольных соответствует k–му элементу СО, где $1 \le k \le K$. Искомый элемент СО будет иметь порядковый номер r, получивший максимальное значение вероятности:

$$p_r(\mathbf{v}) = \max_k \{p_k(\mathbf{v})\}, \ 1 \le k \le K. \tag{1}$$

Приближенные значения $p_1(\mathbf{v}),\ldots,p_K(\mathbf{v})$ представляются в виде конечных многочленов от координат $\mathbf{v}=(v_1,\ldots,v_N)$ и определяются выбором базисных мономов:

$$p_k(\mathbf{v}) \cong c_0^{(k)} + \sum_{i=1}^{N} c_i^{(k)} v_i \ + \sum_{i,j=1}^{N} c_{i,j}^{(k)} v_i v_j + \ldots, \ 1 \le k \le K. \tag{2}$$

Представим упорядоченные базисные мономы из (2) в виде вектора размерности L:

$$\mathbf{x}(\mathbf{v})=(1, v_1, \ldots, v_N, \ldots)^{\mathrm{T}}.$$

Тогда (2) можно записать в векторном виде:

$$\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}),\ldots,p_K(\mathbf{v}))^{\mathrm{T}} \cong A^{\mathrm{T}}\mathbf{x}(\mathbf{v}), \tag{3}$$

где A – матрица размера $L \times K$, столбцами которой являются векторы $\mathbf{a}^{(1)},\ldots, \mathbf{a}^{(K)}$. Каждый такой вектор составлен из коэффициентов при мономах соответствующей строки (2) (с совпадающим верхним индексом), упорядоченных так же, как в векторе $\mathbf{x}(\mathbf{v})$.

Значение A вычисляется приближенно в процессе обучения с использованием последовательности данных: $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}], \ldots, [\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$. Здесь $\mathbf{v}^{(j)}$ – набор параметров крови, соответствующий элементу СО с номером k ($1 \le k \le K$), $\mathbf{y}^{(j)}= (0\ldots1\ldots0)$ – его базисный вектор, где 1 стоит на k-м месте, $1 \le j \le J$):

$$A \cong \left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathrm{T}}\right)^{-1}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{y}^{(j)})^{\mathrm{T}}\right). \tag{4}$$

Поскольку проблема обращения заполненной матрицы большой размерности до сих пор не решена [12], правую часть (4) получаем посредством рекуррентной процедуры [13].

В данной работе рассмотрено семь СО для онкобольных пациентов мужского пола: пищеварительная система (C^1), органы дыхания (C^2), опорно-двигательный аппарат (C^3), урологическая система (C^4), эндокринная система (C^5), печень и желчевыводящие пути (C^6), центральная нервная система и органы чувствительности (C^7).

Нормирование проводим следующим образом. Рассмотрим совокупность семи обучающих выборок исследуемых СО. Для каждого i-го показателя крови находим минимальное и максимальное значение $v_i^{\min}$, $v_i^{\max}$, где $i = 1,..., N$.

$$v_i^{\min} = \min_j \{v_i^j\},\ j= 1, \ldots, J,$$

$$v_i^{\max} = \max_j \{v_i^j\},\ j= 1, \ldots, J\ ,$$

где J – общий объем выборки по всем рассматриваемым СО.

Затем выполняем преобразование:

$$v_i \rightarrow (v_i - v_i^{min})/(v_i^{max} - v_i^{\min}).$$

Использовалась следующая модификация $\mathbf{x}(\mathbf{v})$:

$$\begin{gathered}\mathbf{x} = (1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\},\{v_i v_j v_k v_l v_m\},\\ \{v_i v_j v_k v_l v_m v_n\},\ \{v_i v_j v_k v_l v_m v_n v_p\}),\\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8,\ l\le m\le 8,\ m\le n\le 8,\ n\le p\le 8.\end{gathered} \qquad (5)$$

В (5) выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычисляемым по всем показателям крови из имеющегося набора. Длина полинома 6435. Имеются мономы первого, второго, третьего, четвертого, пятого, шестого и седьмого порядка. Перекрестные произведения используются для мономов второго, третьего, четвертого, пятого, шестого и седьмого порядка.

Исследование проведено на классификаторе из семи классов. Обучающее множество $C^1UC^2UC^3UC^4UC^5UC^6UC^7$ имеет следующий объем (количество наборов крови), дифференцированный по классам: $|C^1| = |C^3| = |C^5| = 33$, $|C^2| = 21$, $|C^4| = |C^7| = 31$, $|C^6| = 25$.

Классификатор обеспечил точность более 95% на обучающем множестве из 207 элементов (остается 10 ошибок).

Проведем сравнительный анализ структуры обучающего множества $C^1UC^2UC^3UC^4UC^5UC^6UC^7$ [1-3].

Расстояние между «своими» и «чужими» элементами

Для каждого из семи рассматриваемых классов C^1, C^2, C^3, C^4, C^5, C^6, C^7 в отдельности найдем минимальное, максимальное и среднее расстояния между

своими векторами (принадлежащими данному классу). Для множества векторов k-го класса определяем их следующим образом.

Минимальное расстояние:

$$U\_k_{\min} = \min_{V^k} \{\|\mathbf{v}^k - \mathbf{u}^k\|\},\ \mathbf{v}^k \in V^k,\ \mathbf{u}^k \in V^k,\ \mathbf{v}^k \neq \mathbf{u}^k. \tag{6}$$

Максимальное расстояние:

$$U\_k_{\max} = \max_{V^k} \{\|\mathbf{v}^k - \mathbf{u}^k\|\},\ \mathbf{v}^k \in V^k,\ \mathbf{u}^k \in V^k, \tag{7}$$

где $\mathbf{v}^k$ и $\mathbf{u}^k$ – пары различных векторов, принадлежащих множеству элементов k-го класса V^k.

Среднее расстояние определим с приведением алгоритма нахождения этой величины:

$$U\_k_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=j+1}^{J_k} \| \mathbf{w}^{k,j} - \mathbf{w}^{k,j1} \| / (J_k\ (J_k\ - 1)/2),\ \mathbf{w}^{k,j} \in V^k,\ j = 1, \ldots, J_k, \tag{8}$$

где $\{\mathbf{w}^{k,j},\ j = 1, \ldots, J_k\} = V^k$ – представление совокупности элементов k-го класса в виде множества перенумерованных векторов.

Аналогично получим соответствующие значения для пар свой–чужой по каждому из классов. Чужой вектор – не принадлежащий рассматриваемому классу. Для обучающего множества, содержащего элементы всех семи классов, $V = \{C^1UC^2UC^3UC^4UC^5UC^6UC^7\}$.

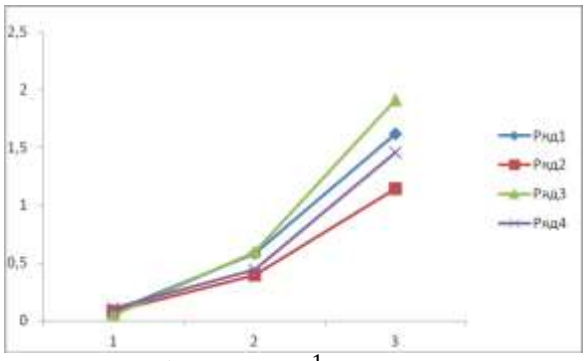
Минимальное расстояние:

$$U\_kz_{\min} = \min_{V} \{\|\mathbf{v}^k - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{v}^k \in V^k,\ \mathbf{u}^{\neg k} \in V^{\neg k}. \tag{9}$$

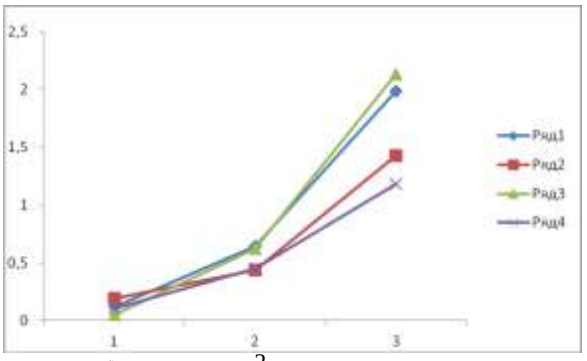
Максимальное расстояние:

$$U\_kz_{\max} = \max_{V} \{\|\mathbf{v}^k - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{v}^k \in V^k,\ \mathbf{u}^{\neg k} \in V^{\neg k}, \tag{10}$$

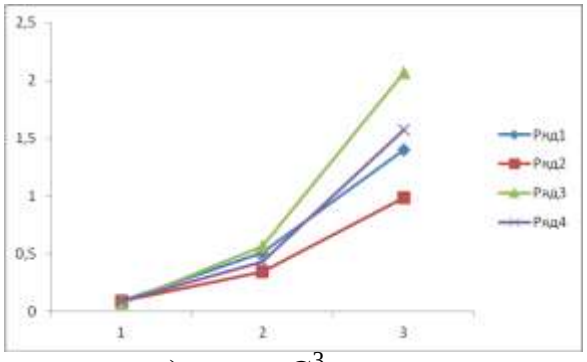
где $\mathbf{v}^k$ и $\mathbf{u}^{\neg k}$ – пары векторов, из которых $\mathbf{v}^k$ принадлежит множеству элементов k-го класса V^k, а $\mathbf{u}^{\neg k}$ принадлежит множеству чужих элементов $V^{\neg k}$ классов, отличных от k-го: $V^{\neg k} = V \setminus V^k$.



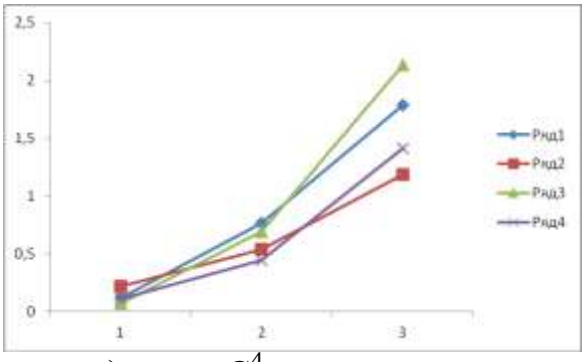
а) класс C^1



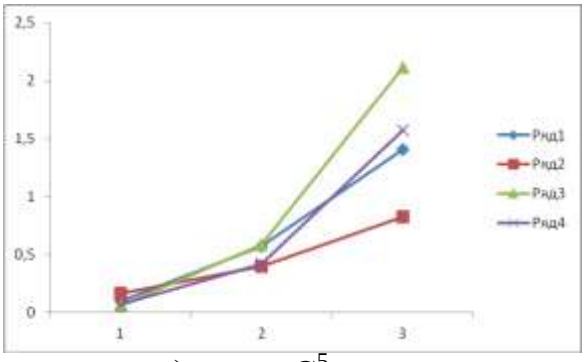
б) класс C^2



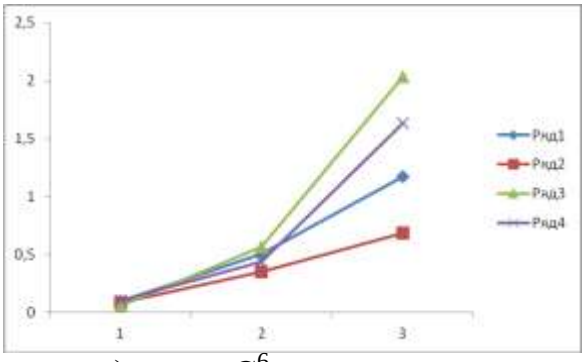
в) класс C^3



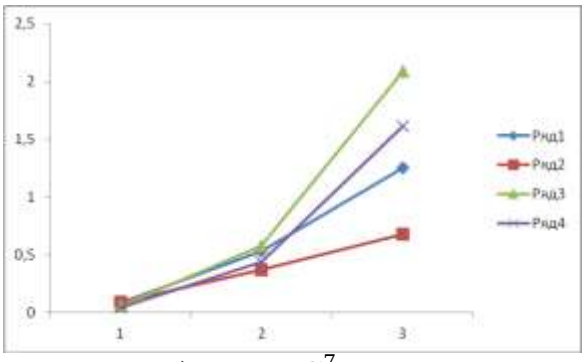
г) класс C^4



д) класс C^5



е) класс C^6



ж) класс C^7

Рис. 1. Минимальное, максимальное и среднее расстояния между парами векторов: свой–свой, свой–чужой, центр масс–свой, центр масс–чужой.

Среднее расстояние:

$$U\_kz_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=1}^{J_{\neg k}} \| \mathbf{w}^{k,j} - \mathbf{w}^{\neg k,j1} \| / (J_k \, J_{\neg k}), \; \mathbf{w}^{k,j} \in V^k, \; j = 1, \ldots, J_k, \quad \mathbf{w}^{\neg k,j1} \in V^{\neg k}, \; j1 = 1, \ldots, J_{\neg k}, \tag{11}$$

где $\{\mathbf{w}^{k,j}, \; j = 1, \ldots, J_k\} = V^k$ – представление совокупности своих элементов k-го класса в виде множества перенумерованных векторов, аналогично для множества чужих элементов классов, отличных от k-го: $\{\mathbf{w}^{\neg k,j1}, \; j1 = 1, \ldots, J_{\neg k}\} = V^{\neg k}$, $V^{\neg k} = V \setminus V^k$.

Продемонстрируем, какие значения принимают перечисленные величины. Расстояние между векторами определяем в метрике L_2.

Для классов C^1, C^2, C^3, C^4, C^5, C^6 и C^7 соответственно представлены (рис.1а, б, в, г, д, е, ж) минимальное, среднее и максимальное расстояния (значения ординат для точек 1, 2, 3 по оси абсцисс) между своими векторами (Ряд 1), аналогичные величины для пар свой-чужой (Ряд 3).

Для всех семи классов (рис. 1) Ряд 1 приблизительно в 1,5 – 2,5 раза превышает Ряд 3 по минимальным значениям (причем оба малы). Первая величина меньше второй для максимальных значений (как существенно, так и незначительно). Для средних имеются оба варианта различия (небольшого).

Аналогичное исследование приведено для наборов из двух и трех классов [1]; в [2, 3] рассматривались совокупности из четырех и пяти классов, соответственно. Сопоставление полученных результатов показывает их сходство.

Отклонение от центра масс своих и чужих элементов

Для каждого из семи рассматриваемых классов C^1, C^2, C^3, C^4, C^5, C^6, C^7 в отдельности получим среднестатистический вектор длины 8, принадлежащий исходному векторному пространству $\mathbf{R}^8$. Иногда такой вектор называют центром масс.

Для центра масс k-го класса значение i-го параметра крови равно среднему арифметическому значений i-х параметров крови по всем J_k, имеющимся в базе наборам показателей крови, относящихся к данному классу:

$$v_i^{k,cp} = (\sum_{j=1}^{J_k} v_i^{k,j}) / J_k, \tag{12}$$

где $\mathbf{v}^{k,j}$ – перенумерованные элементы k-го класса: $\{\mathbf{v}^{k,j} = (v^{k,j}{}_1, \ldots, v^{k,j}{}_N), \; j = 1, \ldots, J_k\} = V^k$.

Для каждого из классов C^1, C^2, C^3, C^4, C^5, C^6 и C^7 найдем минимальное, максимальное и среднее расстояния между центром масс и своими векторами.

Указанные величины для множества векторов k-го класса определяем следующим образом. Минимальное расстояние:

$$D\_k_{\min} = \min_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\},\ \mathbf{u}^k \in V^k. \tag{13}$$

Максимальное расстояние:

$$D\_k_{\max} = \max_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\},\ \mathbf{u}^k \in V^k, \tag{14}$$

где $\mathbf{u}^k$ – вектор, принадлежащий множеству элементов k-го класса V^k, $\mathbf{v}^{k,cp}$ – среднестатистический вектор этого класса.

Среднее расстояние определим более детально с приведением алгоритма нахождения этой величины:

$$D\_k_{\text{ср}} = \sum_{j=1}^{J_k} \|\mathbf{w}^{k,j} - \mathbf{v}^{k,cp}\| / J_k\ ,\ \mathbf{w}^{k,j} \in V^k,\ j = 1, \ldots, J_k, \tag{15}$$

где $\{\mathbf{w}^{k,j},\ j = 1, \ldots, J_k\} = V^k$ – представление совокупности элементов k-го класса в виде множества перенумерованных векторов.

Аналогично получим соответствующие значения по каждому из классов между центром масс и чужими векторами. Эти результаты зависят от количества классов, входящих в обучающее множество.

Минимальное расстояние:

$$D\_kz_{\min} = \min_{V^{\neg k}} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{u}^{\neg k} \in V^{\neg k}. \tag{16}$$

Максимальное расстояние:

$$D\_kz_{\max} = \max_{V^{\neg k}} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{u}^{\neg k} \in V^{\neg k}, \tag{17}$$

где $\mathbf{u}^{\neg k}$ – вектор, принадлежащий множеству чужих элементов $V^{\neg k}$ классов, отличных от k-го: $V^{\neg k} = V \setminus V^k$, $\mathbf{v}^{k,cp}$ – среднестатистический вектор k-го класса.

Среднее расстояние:

$$D\_kz_{\text{ср}} = \sum_{j=1}^{J_{\neg k}} \|\mathbf{v}^{k,cp} - \mathbf{w}^{\neg k,j}\| / J_{\neg k}\ ,\ \mathbf{w}^{\neg k,j} \in V^{\neg k},\ j = 1, \ldots, J_{\neg k}, \tag{18}$$

где $\{\mathbf{w}^{\neg k,j}, j = 1, \ldots, J_{\neg k}\} = V^{\neg k}$, $V^{\neg k} = V \setminus V^k$ – представление совокупности чужих элементов классов, отличных от k-го в виде множества перенумерованных векторов.

Для каждого класса соответственно (рис. 1) из объединения $C^1 \cup C^2 \cup C^3 \cup C^4 \cup C^5 \cup C^6 \cup C^7$ представлено минимальное, среднее и максимальное расстояния (значения ординат для точек 1, 2, 3 по оси абсцисс) между центром масс и своими векторами (Ряд 2), аналогично между парами центр масс–чужой вектор (Ряд 4).

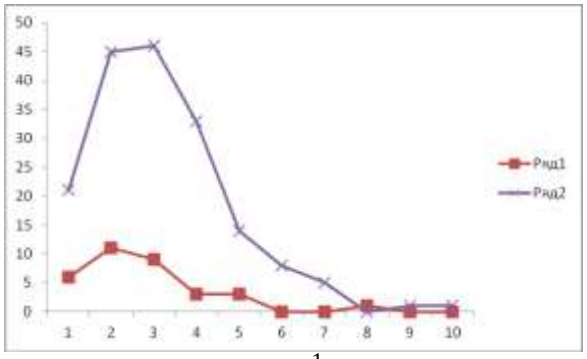
Для классов C^1, C^3 и C^6 (рис. 1а, в, е) Ряд 4 всюду превышает Ряд 2. Аналогичная картина наблюдается для классов C^5 и C^7, исключая минимальные значения (рис. 1д, ж). При рассмотрении этих пяти классов для минимальных и средних значений разница небольшая, а для максимальных – существенная. Для C^2 имеется такое же, как у C^1, C^3 и C^6, соотношение по средним значениям и противоположное – по минимальным и максимальным (рис. 1б). В классе C^4 (рис. 1г) Ряд 2 превышает Ряд 4 для минимальных и средних значений, а для максимальных – ситуация обратная.

Для рассмотренных ранее наборов из двух и трех классов [1], а также из четырех и пяти классов соответственно [2, 3] получены аналогичные результаты.

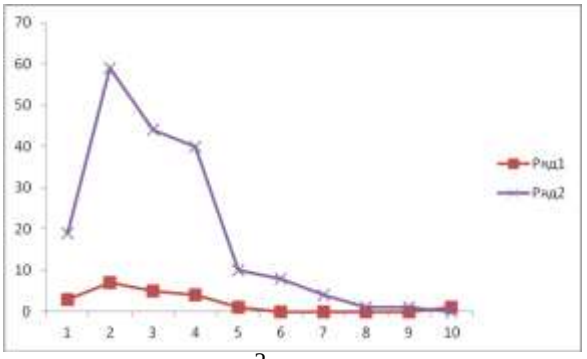
Распределение числа своих и чужих элементов при удалении от центра масс

Диапазон расстояний между центром масс k-го класса СО и векторами этого же класса («своими», $\mathbf{v}^k \in V^k$) по рассматриваемой базе, согласно формулам (13), (14), находится на отрезке $[D\_k_{\min}, D\_k_{\max}]$. Диапазон расстояний между центром масс k-го класса СО и векторами всех других классов («чужими», $\mathbf{z}^k \in \{V \setminus V^k\}$), согласно формулам (16), (17), – на отрезке $[D\_kz_{\min}, D\_kz_{\max}]$. Пусть

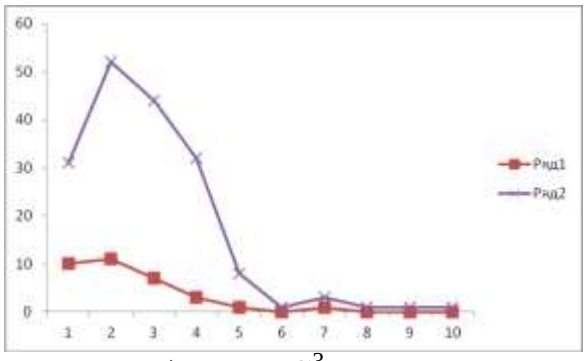
$$\begin{aligned} Dk_{\min} &= \min(D\_k_{\min}, D\_kz_{\min}), \\ Dk_{\max} &= \max(D\_k_{\max}, D\_kz_{\max}). \end{aligned} \qquad (19)$$



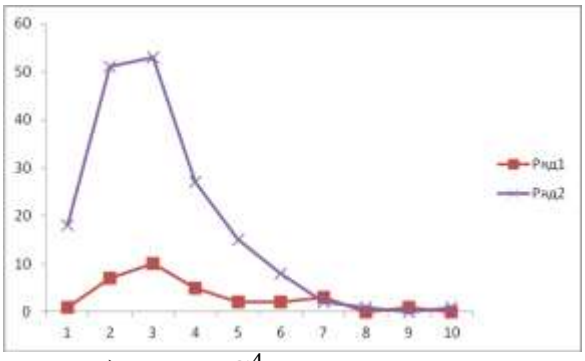
а) класс C^1



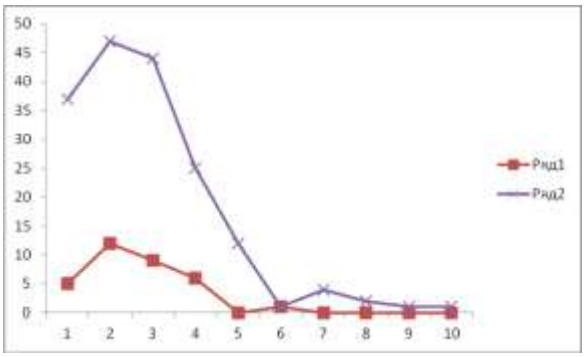
б) класс C^2



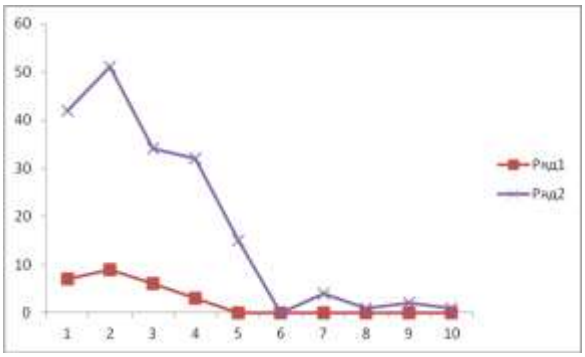
в) класс C^3



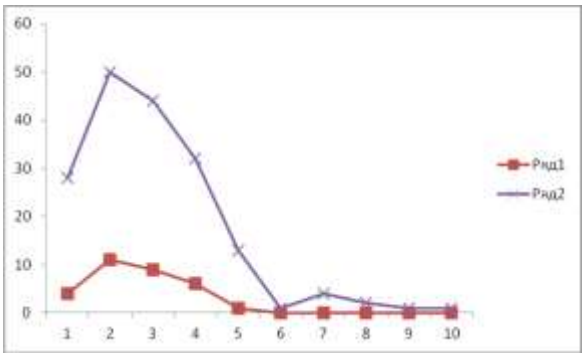
г) класс C^4



д) класс C^5



е) класс C^6



ж) класс C^7

Рис. 2. Распределение числа своих и чужих элементов при удалении от центра масс ($C^1UC^2UC^3UC^4UC^5UC^6UC^7$)

Делим отрезок [Dk_{min}, Dk_{max}] (оси абсцисс на рис. 2а, б, в, г, д), е), ж)) на десять равных по длине частей – один отрезок и девять полуинтервалов: [Dk_{min}, $Dk_{min} + d$], ($Dk_{min} + d$, $Dk_{min} + 2d$], … , ($Dk_{min} + 9d$, $Dk_{min} + 10d$], где $d = (Dk_{max} - Dk_{min})/10$. Определим, какое количество своих векторов попало в каждый такой участок (аналогично для чужих векторов). Затем рассмотрим распределение числа своих (чужих) векторов на отрезке [Dk_{min}, Dk_{max}].

Для каждого класса (рис. 2) соответственно из объединения $C^1UC^2UC^3$ UC^4UC^5 UC^6UC^7 представлено распределение количества своих (Ряд 1) и чужих (Ряд 2) элементов на отрезке [Dk_{min}, Dk_{max}].

В характере полученных распределений числа своих (чужих) элементов имеется структурное сходство между всеми семью классами. Эта же аналогия наблюдается в отношении распределений своих–чужих элементов. Отметим, что для классов C^1, C^3, C^4, C^5, C^6 и C^7 получены результаты (рис. 2а, в, г, д, е, ж), отличающиеся от C^2 (рис. 2б). Для C^2 (рис. 2б) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$, что не выполняется для C^1, C^3, C^4, C^5, C^6 и C^7 (рис. 2а, в, г, д, е, ж).

Схожие результаты были получены для множеств из двух и трех классов [1], а также из четырех и пяти классов соответственно [2, 3]. Исключением является класс C^4. Дело в том, что в данной работе для преодоления проблемы с обучением классификатора был удален один элемент из множества C^4, определяющий указанную зависимость.

Можно также убедиться в сходстве структуры обучающих множеств [1, 2, 6, 10] классификаторов двух типов: вычисляющих СО – описанных в данной работе и в [1-3], а также находящих СЗЧ [4-10].

Заключение

Предложена математическая модель, и на ее основе разработан классификатор, с помощью которого по показателям периферической крови онкологического пациента мужского пола можно конкретизировать область локализации патологии. Данный подход использует статистический метод распознавания, базирующийся на полиномиальной регрессии.

Полиномиальный вектор имеет 6435 элементов, представляющих собой мономы до седьмого порядка. Он построен и обучен для семи классов, соответствующих следующим СО для мужчин: (C^1) пищеварительная, (C^2) урологическая и (C^3) эндокринная системы, (C^4) органы дыхания, (C^5) опорно-двигательный аппарат, (C^6) печени и желчевыводящих путей, (C^7) центральной нервной системы и органов чувствительности.

На обучающем множестве из 207 наборов крови точность классификатора составляет более 95% (остается 10 ошибок).

Проведено исследование структуры обучающего множества.

Для каждого из семи классов C^1, C^2, C^3, C^4, C^5, C^6 и C^7 в отдельности найдено минимальное, максимальное и среднее расстояние между своими (принадлежащими данному классу) векторами – Ряд 1.

Также получены соответствующие значения для пар свой–чужой (элемент, не относящийся к рассматриваемому классу) – Ряд 3.

Для всех семи классов Ряд 1 немного превышает Ряд 3 по минимальным значениям (обе малы). Ряд 1 меньше, чем Ряд 3, для максимальных значений (как существенно, так и незначительно). Для средних величин имеются оба варианта различия (небольшого).

Ранее для наборов из двух, трех, четырех и пяти классов ранее были получены схожие результаты.

Для классов C^1, C^2, C^3, C^4, C^5, C^6 и C^7 вычислили среднестатистический вектор, принадлежащий исходному векторному пространству $\mathbf{R}^8$ (центр масс). Найдено минимальное, максимальное и среднее расстояние между центром масс и своими (чужими) векторами – Ряд 2 (Ряд 4).

Для классов C^1, C^3 и C^6 Ряд 4 всюду превышает Ряд 2. Аналогичная картина наблюдается для класса C^5 и C^7, исключая минимальные значения. При рассмотрении этих пяти классов для минимальных и средних значений разница небольшая, а для максимальных – существенная. Для C^2 имеется такое же, как у C^1, C^3 и C^6, соотношение по средним значениям и противоположное – по минимальным и максимальным. В классе C^4 Ряд 2 превышает Ряд 4 для минимальных и средних значений, а для максимальных – ситуация обратная.

Для рассмотренных ранее наборов из двух, трех, четырех и пяти классов получены аналогичные результаты.

Построено распределение количества своих и чужих элементов на отрезке их нахождения при удалении от центра масс.

В характере распределения числа своих (чужих) элементов имеется структурное сходство между всеми семью классами. Это же наблюдается в отношении распределений своих–чужих элементов.

Аналогичные результаты были получены ранее для множеств, состоящих из двух, трех, четырех и пяти классов. Имеющееся исключение (класс C^4) объясняется уменьшением соответствующего ему фрагмента обучающего множества на один элемент, который отвечает за указанную зависимость.

Кроме того, наблюдается сходство структуры соответствующих обучающих множеств классификаторов двух типов: с одной стороны, вычисляющих СО, а с другой стороны – определяющих СЗЧ.

Библиографический список

1. Гавриков Б.М., Пестрякова Н.В. Статистический классификатор для диагностики онкологических заболеваний // Информационные технологии и вычислительные системы. 2023. № 1. С. 39-49.

2. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Статистический подход для диагностики онкологических заболеваний по параметрам крови // Препринты ИПМ им. М.В.Келдыша. 2022. № 72. 12 с. DOI:10.20948/prepr-2022-72

3. Гавриков Б.М., Гавриков М.Б., Лебеденко И.М., Пестрякова Н.В. Нахождение области локализации онкопатологии по параметрам крови больного // Препринты ИПМ им. М.В. Келдыша. 2023. №24. С.1-15. DOI: 10.20948/prepr-2023-24.

4. Гавриков Б.М., Лебеденко И.М., Пестрякова Н.В., Ставицкий Р.В. Об одном статистическом методе оценивания состояния здоровья человека // Труды ИСА РАН, 2016. Т. 66. № 2. С. 54-59.

5. Гавриков Б.М., Пестрякова Н.В. О построении признакового пространства в задаче обучения // Информационные технологии и вычислительные системы. 2018. № 1. С. 22-29. DOI: 10.14357/20718632180104

6. Гавриков Б.М., Пестрякова Н.В., Ставицкий Р.В. О свойствах обучающих множеств // Информационные технологии и вычислительные системы. 2018. № 4. С.97-107. DOI: 10.14357/207186321804010

7. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Статистический метод распознавания на основе нелинейной регрессии // Математическое моделирование. 2020. Т.32. №4. С.116-130. DOI: 0.20948/mm-2020-04-09

8. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. О способности статистического классификатора к обобщениям // Информационные технологии и вычислительные системы. 2021. № 4. С. 38-50. DOI: 10.14357/20718632210404.

9. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. О структуре базы обучения классификатора для оценивания состояния здоровья человека // Препринты ИПМ им. М.В.Келдыша. 2018. № 126. 18 с. DOI:10.20948/prepr-2018-126

10. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В., Ставицкий Р.В. Структура базы обучения статистического классификатора состояний систем организма человека // Препринты ИПМ им. М.В.Келдыша. 2018. № 255. 40 с. DOI:10.20948/prepr-2018-255

11. Ставицкий Р.В., Лебедев Л.А., Лебедев А.Л., Смыслов А.Ю. Количественная оценка гомеостатической активности здоровых и больных людей. — М.: ГАРТ. 2013. 131 с.

12. Гавриков М.Б., Локуциевский О.В. Начала численного анализа. — М.: Янус, 1995.

13. Schürmann J. Pattern Classification. — New York: John Wiley&Sons, Inc., 1996.