

Статистический классификатор для диагностики онкологических заболеваний

Б. М. Гавриков, Н. В. Пестрякова

Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", Москва, Россия

Аннотация. Описывается подход к диагностике онкологических заболеваний человека различной локализации. С этой целью используется разработанный авторами статистический классификатор, основанный на полиномиально-регрессионном подходе и имеющий вероятностные оценки. Обучение классификатора проводилось на базах данных анализа периферической крови онкобольных для различных систем организма.

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

DOI 10.14357/20718632230104

Введение

Настоящая работа основана на концепции крупнейших гематологов, исходящей из того, что многие заболевания человека вносят изменения в состав его крови. Организм рассматривается как совокупность систем органов. При оценке состояния здоровья человека (СЗЧ) гематологи предлагают использовать не менее пяти показателей периферической крови [1].

Проведенные авторами исследования доказали приемлемость статистического метода классификации для оценивания СЗЧ по показателям периферической крови [2-8]. Получены результаты, относящиеся к структуре наборов крови, на основе которых строились обучающие выборки [3, 4, 7, 8].

В настоящей работе мы делаем новый шаг: приступаем непосредственно к разработке компьютерных методов медицинской диагностики, а именно: рассматривается способ, позволяющий определить область локализации опухоли

при постановке диагноза в онкологии. Способ дополняет арсенал уже известных методов постановки диагноза.

Для этого обучение статистического классификатора предлагается проводить на наборах параметров периферической крови онкологических больных, относящихся к нескольким различным системам организма (СО).

Нет необходимости рассматривать все СО, вполне допустимо ограничиться только проблемными. Был построен и обучен классификатор для трех СО. Аналогичная уточняющая процедура проведена также для трех пар из имеющегося набора.

В данной постановке не возникает вопрос о «чистоте эксперимента», хотя эта проблема практически всегда имеет место при численном моделировании на основе измеренных данных. Дело в том, что наборы параметров крови онкобольных получены одним и тем же способом и имеют достаточный (сопоставимый) объем. Отсутствуют также вырожденные случаи по

отдельным признакам. Все это выгодно отличается проводимое здесь моделирование диагностики от оценивания состояния здоровья [1-3].

1. Метод классификации

Используется восемь показателей крови. Их общепринятые обозначения и размерность следующие: RBC [L⁻¹] – эритроциты, HGB [gL⁻¹] – гемоглобин, PLT [L⁻¹] – тромбоциты, WBC [L⁻¹] – лейкоциты, LIMP [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, $1 \leq k \leq K$. Вводим вектор $\mathbf{v} \in \mathbf{R}^N$, i -я компонента которого – отнормированная на отрезок, не выходящий за пределы [0,1], величина i -го показателя крови онкобольных, где $N=8$.

Отождествляем k -й элемент множества СО с базисным вектором из $\mathbf{R}^K$: $\mathbf{e}_k = (0 \dots 1 \dots 0)$, причем 1 находится на k -м месте, $1 \leq k \leq K$. Обозначим $\mathbf{Y} = \{\mathbf{e}_1, \dots, \mathbf{e}_K\}$.

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

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

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

$$p_k(\mathbf{v}) \cong c_0^{(k)} + \sum_{i=1}^N c_i^{(k)} v_i + \sum_{i,j=1}^N c_{i,j}^{(k)} v_i v_j + \dots, \quad (2)$$

$$1 \leq k \leq K.$$

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

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

Тогда (2) можно записать в векторном виде:

$$\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}), \dots, p_K(\mathbf{v}))^T \cong A^T \mathbf{x}(\mathbf{v}), \quad (3)$$

где A – матрица размера $L \times K$, столбцами которой являются векторы $\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(K)}$. Каждый

такой вектор составлен из коэффициентов при мономах соответствующей строки (2) (с совпадающим верхним индексом), упорядоченных так же, как в векторе $\mathbf{x}(\mathbf{v})$.

Приближенное вычисление A производится при обучении на конечной последовательности: $[\mathbf{v}^{(1)}, \mathbf{y}^{(1)}], \dots, [\mathbf{v}^{(J)}, \mathbf{y}^{(J)}]$. Здесь $\mathbf{v}^{(j)}$ – набор параметров крови, соответствующий элементу СО с номером k ($1 \leq k \leq K$), $\mathbf{y}^{(j)} = (0 \dots 1 \dots 0)$ – его базисный вектор, где 1 стоит на k -м месте, $1 \leq j \leq J$:

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

Поскольку проблема обращения заполненной матрицы большой размерности до сих пор не решена [9], правую часть (4) получаем посредством рекуррентной процедуры [10].

В данной работе рассмотрены три СО для мужчин: пищеварительная система (C^1), органы дыхания (C^2), опорно-двигательный аппарат (C^3).

Нормирование проводим следующим образом. Рассмотрим все три обучающие выборки изучаемых СО. Для каждого i -го показателя крови находим минимальное и максимальное значение $v_i^{\min}, v_i^{\max}$, где $i = 1, \dots, N$.

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

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

где J – объем выборки по данной совокупности СО.

Затем выполняем следующее преобразование:

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

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

$$\mathbf{x} = (1, \{v_i\}, \{v_i v_j\}, \{v_i v_j v_k\}, \{v_i v_j v_k v_l\}, \{v_i v_j v_k v_l v_m\}),$$

$$1 \leq i \leq 8, i \leq j \leq 8, j \leq k \leq 8, k \leq l \leq 8, l \leq m \leq 8. \quad (5)$$

Длина полинома 1287. Имеются мономы первого, второго, третьего, четвертого и пятого порядка. Перекрестные произведения используются для мономов второго, третьего, четвертого и пятого порядка.

В (5) выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычис-

ляемым по всем восьми параметрам крови из имеющегося набора.

Опишем четыре обученных классификатора, для каждого из которых далее проведем анализ обучающего множества.

Пример 1. Обучающее множество $C^1UC^2UC^3$ имеет следующий объем (количество наборов крови), дифференцированный по классам: $|C^1| = |C^3| = 33$, $|C^2| = 21$. Классификатор обеспечил точность 94,3% на обучающем множестве, состоящем из трех классов при совокупности из 87 элементов (сохраняется пять ошибок).

Также рассматриваются классификаторы, основанные на попарных объединениях классов из трех имеющихся. Такие комбинации позволяют детализировать структуру обучающего множества. Заметим, что в случаях, когда производилось обучение на двух классах (**Пример 2:** C^1UC^2 , **Пример 3:** C^1UC^3 , **Пример 4:** C^2UC^3) получена точность 100%. Для **Примера 2** (C^1UC^2) достаточно использовать модификацию вектора $\mathbf{x}(\mathbf{v})$ четвертого порядка с длиной полинома 495, без комбинации элементов в последней паре фигурных скобок (5).

2. Расстояния между «своими» и «чужими» элементами

Для каждого из трех рассматриваемых классов C^1 , C^2 , C^3 в отдельности найдем минимальное, максимальное и среднее расстояние между своими векторами (принадлежащими данному классу). Для множества векторов k -го класса определяем их следующим образом.

Минимальное расстояние:

$$U_{-k_{\min}} = \min_{V^k} \{ \|\mathbf{v}^k - \mathbf{u}^k\| \}, \mathbf{v}^k \in V^k, \mathbf{u}^k \in V^k, \mathbf{v}^k \neq \mathbf{u}^k \quad (6)$$

Максимальное расстояние:

$$U_{-k_{\max}} = \max_{V^k} \{ \|\mathbf{v}^k - \mathbf{u}^k\| \}, \mathbf{v}^k \in V^k, \mathbf{u}^k \in V^k \quad (7)$$

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:

$$U_{-k_{\text{cp}}} = \sum_{j=1}^{J_k} \sum_{j1=j+1}^{J_k} \|\mathbf{w}^{k,j} - \mathbf{w}^{k,j1}\| / (J_k (J_k - 1) / 2),$$

$$\mathbf{w}^{k,j} \in V^k, j = 1, \dots, J_k \quad (8)$$

где $\{\mathbf{w}^{k,j}, j = 1, \dots, J_k\} = V^k$ – 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. Обучающее множество, содержащее элементы всех трех классов $C^1UC^2UC^3$, суть $V = \{V^1U^2U^3\}$; для C^1UC^2 : $V = \{V^1U^2\}$; для C^1UC^3 : $V = \{V^1U^3\}$; для C^2UC^3 : $V = \{V^2U^3\}$. Поэтому вычисления ниже зависят от количества классов (два или три), из которых состоит обучающее множество.

Минимальное расстояние:

$$U_{-k_{\min}} = \min_V \{ \|\mathbf{v}^k - \mathbf{u}^{-k}\| \}, \mathbf{v}^k \in V^k, \mathbf{u}^{-k} \in V^{-k} \quad (9)$$

Максимальное расстояние:

$$U_{-k_{\max}} = \max_V \{ \|\mathbf{v}^k - \mathbf{u}^{-k}\| \}, \mathbf{v}^k \in V^k, \mathbf{u}^{-k} \in V^{-k} \quad (10)$$

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

Среднее расстояние:

$$U_{-k_{\text{cp}}} = \sum_{j=1}^{J_k} \sum_{j1=1}^{J_{-k}} \|\mathbf{w}^{k,j} - \mathbf{w}^{-k,j1}\| / (J_k J_{-k}),$$

$$\mathbf{w}^{k,j} \in V^k, j = 1, \dots, J_k, \mathbf{w}^{-k,j1} \in V^{-k}, j1 = 1, \dots, J_{-k} \quad (11)$$

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

Продemonстрируем на примерах, какие значения принимают перечисленные величины. Расстояние между векторами определяем в метрике L_2 .

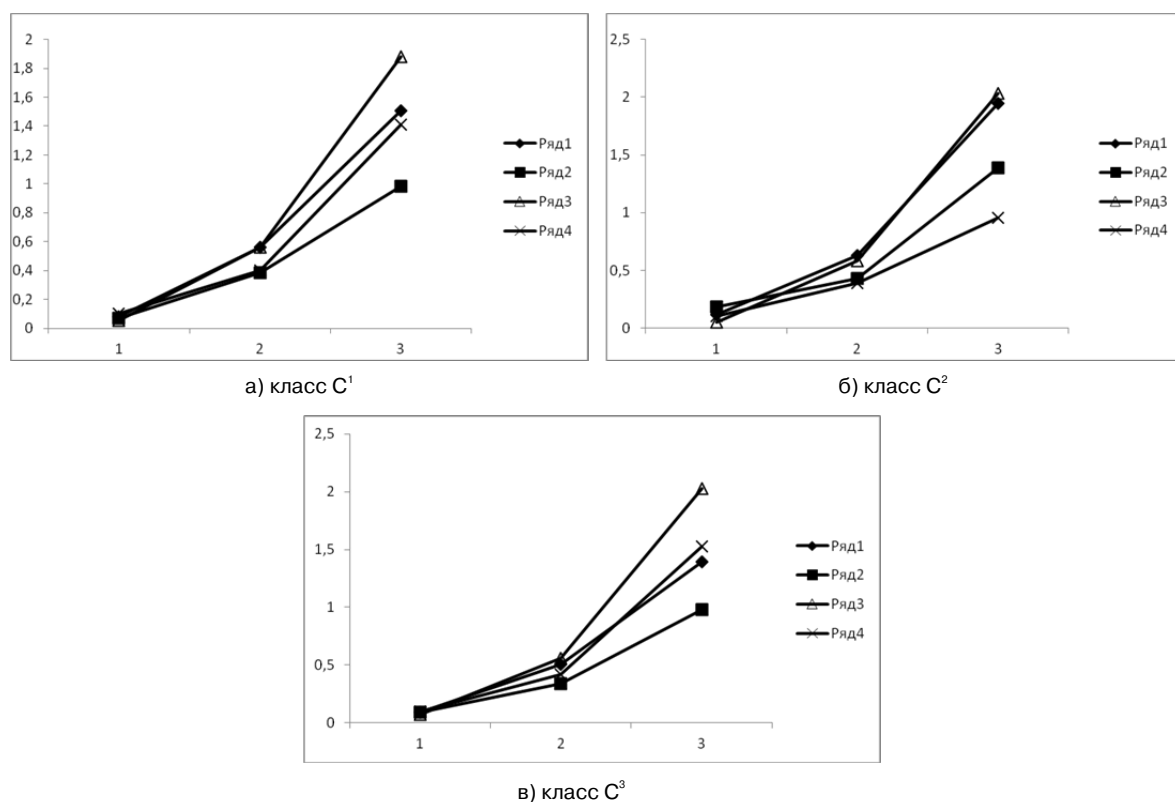


Рис. 1. Минимальное, максимальное и среднее расстояние между парами векторов: свой – свой, свой – чужой, центр масс – свой, центр масс – чужой ($C^1UC^2UC^3$)

Пример 1. Рассмотрим совокупность всех трех классов $C^1UC^2UC^3$ (Рис. 1). Для классов C^1 , C^2 и C^3 соответственно представлены (Рис. 1 а, б, в) минимальное, среднее и максимальное расстояние (значения ординат для точек 1, 2, 3 по оси абсцисс) между своими векторами (Ряд 1), аналогичные величины для пар свой – чужой (Ряд 3).

Для класса C^1 (Рис. 1 а) Ряд 1 несколько превышает Ряд 3 по минимальным значениям (причем обе малы); максимальное значение первой величины существенно меньше, чем второй, а средние практически равны. Для классов C^2 и C^3 (Рис. 1 б, в) этот порядок сохраняется в отношении минимальных и максимальных значений, однако в случае C^2 максимальные параметры близки. По средним значениям Ряд 1 и Ряд 3 отличаются немного, причем для C^2 первый из них превышает второй, а для C^3 наоборот.

Пример 2. Рассмотрим совокупность двух классов C^1UC^2 (Рис. 2). Для C^1 и C^2 соответственно представлены (Рис. 2 а, б) минимальное, среднее и максимальное расстояние (значения ординат для точек 1, 2, 3 по оси абсцисс) между своими векторами (Ряд 1), аналогичные величины для пар свой – чужой (Ряд 3).

Для класса C^1 (Рис. 2 а) Ряд 1 несколько превышает Ряд 3 по минимальным значениям (причем, обе малы); первая величина меньше, чем вторая как по максимальным значениям (различаются существенно), так и по средним (разница небольшая). Для класса C^2 (Рис. 2 б) Ряд 1 немного превышает Ряд 3 в отношении минимальной, средней и максимальной величин.

Пример 3. Рассмотрим совокупность двух классов C^1UC^3 (Рис. 3). Для класса C^1 (Рис. 3 а) минимальные значения Ряда 1 и Ряда 3 почти равны; среднее значение первой из величин незначительно больше второй; первое максимальное значение немного меньше второго. В классе C^3 (Рис. 3 б) по минимальным значениям (весьма малым) Ряд 1 несколько превышает Ряд 3; первое максимальное значение существенно меньше, чем второе, аналогичным образом различаются и средние значения, но немного.

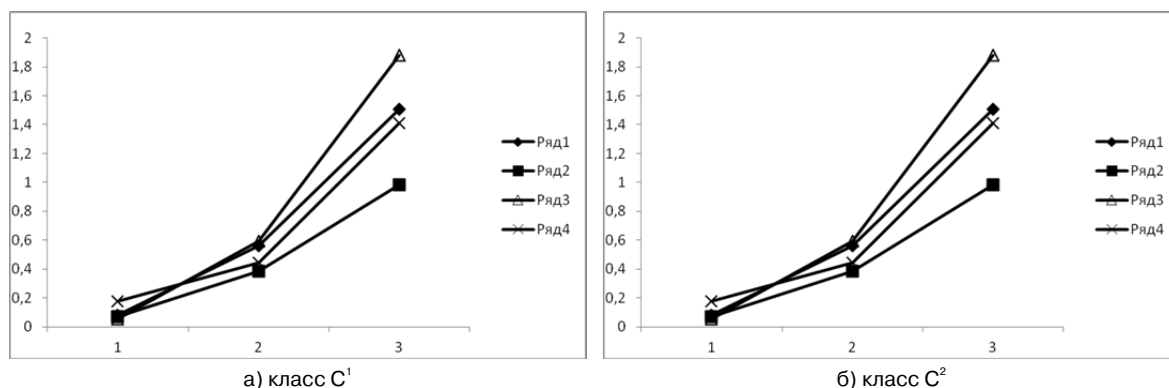


Рис. 2. Минимальное, максимальное и среднее расстояние между парами векторов: свой – свой, свой – чужой, центр масс – свой, центр масс – чужой (C^1UC^2)

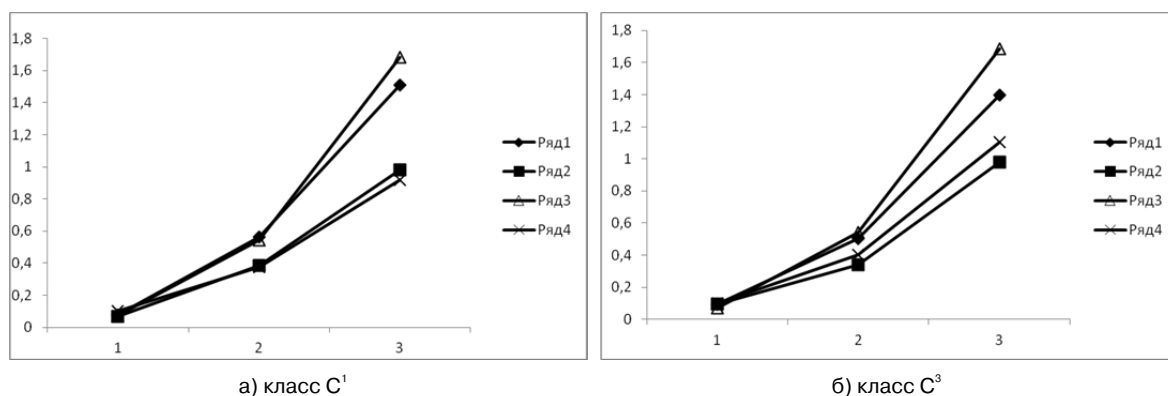


Рис. 3. Минимальное, максимальное и среднее расстояние между парами векторов: свой – свой, свой – чужой, центр масс – свой, центр масс – чужой (C^1UC^3)

Пример 4. Рассмотрим совокупность двух классов C^2UC^3 (Рис. 4).

Для класса C^2 (Рис. 4 а) по минимальным и средним значениям Ряд 1 незначительно превышает Ряд 3; соответствующее максимальное значение первой величины немного меньше, чем второй. Для класса C^3 (Рис. 4 б) Ряд 3 превышает Ряд 1 в отношении минимальной величины (незначительно), средней (немного) и максимальной (существенно).

Приведенные в **Примере 2** и **Примере 4** результаты (Рис. 2, 4), показывают, что множества C^1UC^2 и C^2UC^3 аналогичны по структуре, C^1 и C^3 схожи.

3. Отклонение от центра масс своих и чужих элементов

Для каждого из трех рассматриваемых классов C^1 , C^2 , C^3 в отдельности получим среднестатистический вектор длины 8, принадлежа-

щий исходному векторному пространству 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:

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

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

Для каждого из трех классов C^1 , C^2 , C^3 найдем минимальное, максимальное и среднее расстояние между центром масс и своими векторами.

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

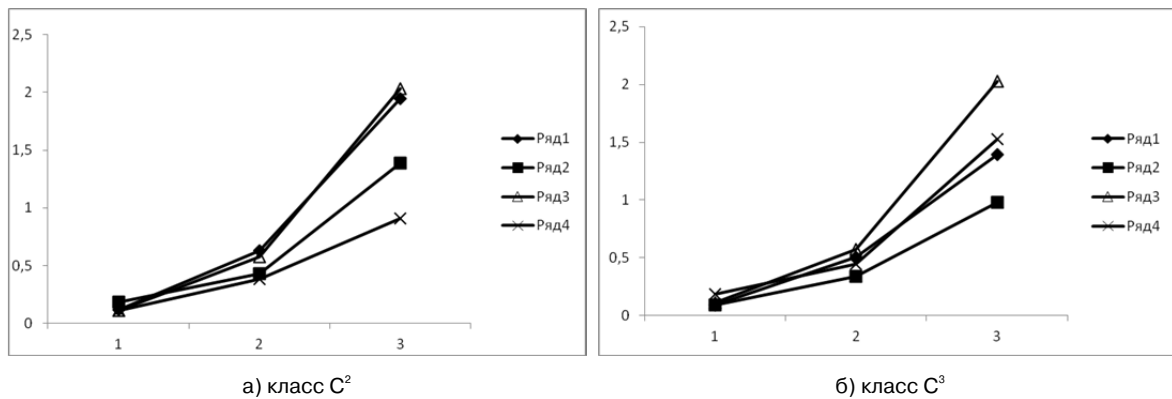


Рис. 4. Минимальное, максимальное и среднее расстояние между парами векторов: свой – свой, свой – чужой, центр масс – свой, центр масс – чужой (C^2UC^3)

$$D_{\min} = \min_{V^k} \{ \|v^{k,cp} - u^k\| \}, u^k \in V^k \quad (13)$$

Максимальное расстояние:

$$D_{\max} = \max_{V^k} \{ \|v^{k,cp} - u^k\| \}, u^k \in V^k \quad (14)$$

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

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

$$D_{\text{ср}} = \sum_{j=1}^{J_k} \|w^{k,j} - v^{k,cp}\| / J_k, \quad w^{k,j} \in V^k, j = 1, \dots, J_k \quad (15)$$

где $\{w^{k,j}, j = 1, \dots, J_k\} = V^k$ – 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:

$$D_{\min} = \min_{V^{-k}} \{ \|v^{k,cp} - u^{-k}\| \}, u^{-k} \in V^{-k} \quad (16)$$

Максимальное расстояние:

$$D_{\max} = \max_{V^{-k}} \{ \|v^{k,cp} - u^{-k}\| \}, u^{-k} \in V^{-k} \quad (17)$$

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

Среднее расстояние:

$$D_{\text{ср}} = \sum_{j=1}^{J_{-k}} \|v^{k,cp} - w^{-k,j}\| / J_{-k}, \quad w^{-k,j} \in V^{-k}, j = 1, \dots, J_{-k} \quad (18)$$

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

Пример 1. Для каждого класса (Рис. 1 а, б, в) соответственно из объединения $C^1UC^2UC^3$ представлено минимальное, среднее и максимальное расстояние (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 класса C^1 (Рис. 1 а) Ряд 4 всюду превышает Ряд 2, причем для минимальных и средних значений разница небольшая, а для максимальных – существенная. Аналогичная картина наблюдается для класса C^3 (Рис. 1 в). Для C^2 соотношения противоположные (Рис. 1 б).

Следует отметить, что нераспознанные элементы относятся к классу C^1 , причем в перечне альтернатив на первом месте стоит C^3 . Имеющиеся проблемы с классификацией коррелируются с полученной структурой обучающего множества.

Пример 2. Для каждого класса (Рис. 2 а, б) соответственно из объединения C^1UC^2 представлено минимальное, среднее и максимальное расстояние (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 класса C^1 (Рис. 2 а) Ряд 4 всюду превышает Ряд 2. Для C^2 имеют место противоположные соотношения, но для средних значений разница небольшая (Рис. 2 б).

Пример 3. Рассмотрим совокупность C^1UC^3 . По классу C^1 (Рис. 3 а) Ряд 4 превышает Ряд 2 для минимальных значений; для средних и максимальных – ситуация противоположная, причем, в двух последних случаях разница соответственно незначительная и небольшая. По классу C^3 (Рис. 3 б) Ряд 4 превышает Ряд 2, но для минимальных значений различие несущественное, а для средних и максимальных – небольшая.

Необходимость использования более сложных полиномов в **Примере 1** и **Примере 3** по сравнению с **Примером 2** определяется особенностями структуры обучающего множества, относящимися к классам C^1 и C^3 . Это позволяет понять сопоставление результатов, полученных по классу C^1 (Рис. 2 а, 3 а).

Пример 4. Во множестве C^2UC^3 для класса C^2 (Рис. 4а) Ряд 2 всюду превышает Ряд 4, причем для минимальных и средних значений разница небольшая, а для максимальных – существенная. Для C^3 имеют место противоположные соотношения (Рис. 4 б).

Результаты, полученные в **Примере 2** и **Примере 4** (Рис. 2, 4), показывает сходство структуры множеств C^1UC^2 и C^2UC^3 , а также соответственно классов C^1 и C^3 .

4. Распределение числа своих и чужих элементов при удалении от центра масс

Диапазон расстояний между центром масс k -го класса CO и векторами этого же класса («своими», $v^k \in V^k$) по рассматриваемой базе, согласно формулам (13), (14), находится на отрезке $[D_{k_{\min}}, D_{k_{\max}}]$. Диапазон расстояний между центром масс k -го класса CO и векторами всех других классов («чужими»,

$z^k \in \{V \setminus V^k\}$), согласно формулам (16), (17), – на отрезке $[D_{kz_{\min}}, D_{kz_{\max}}]$. Пусть

$$\begin{aligned} Dk_{\min} &= \min(D_{k_{\min}}, D_{kz_{\min}}) \\ Dk_{\max} &= \max(D_{k_{\max}}, D_{kz_{\max}}) \end{aligned} \quad (19)$$

Делим отрезок $[Dk_{\min}, Dk_{\max}]$ (оси абсцисс на Рис. 5 а, б, в) 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 $[Dk_{\min}, Dk_{\max}]$.

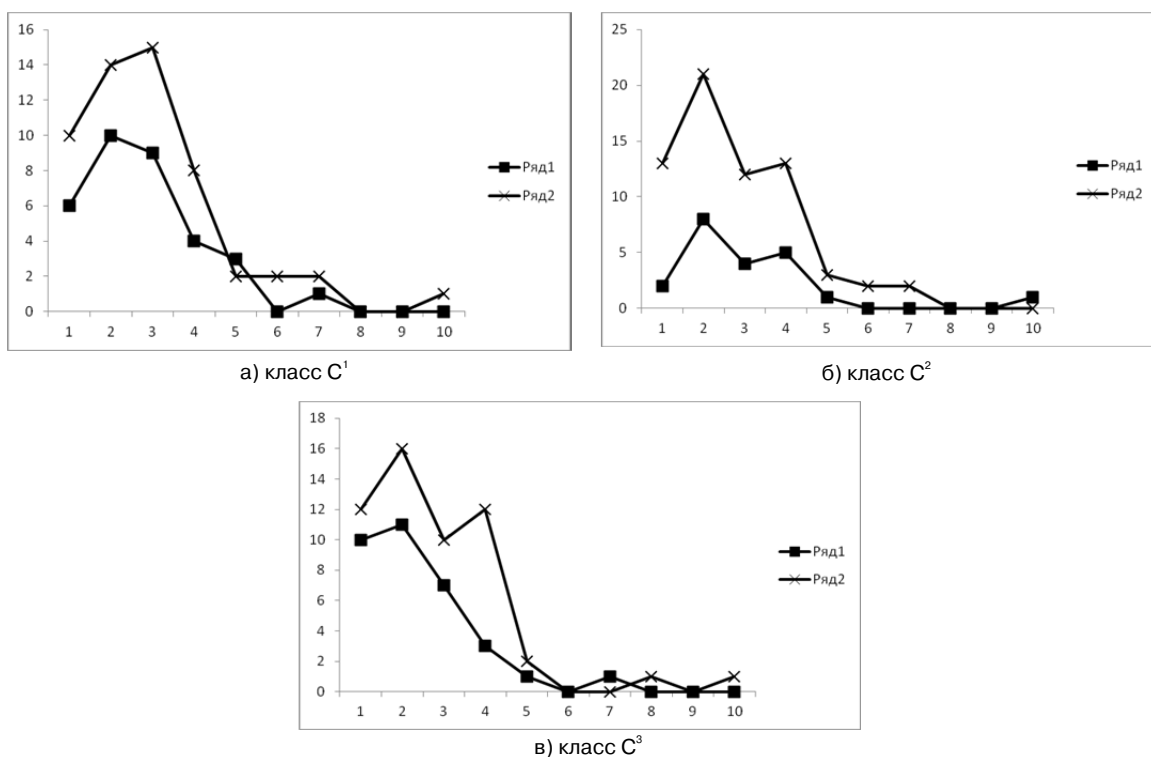
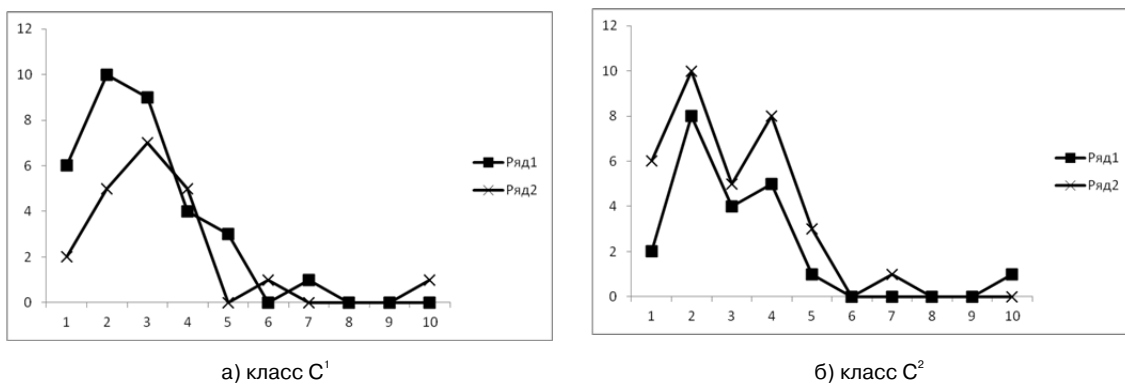
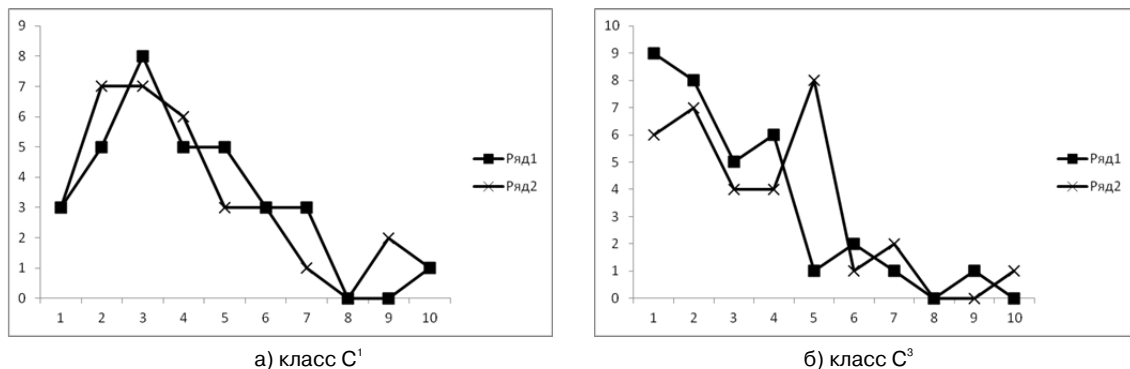
Пример 1. Для каждого класса (Рис. 5 а, б, в) соответственно из объединения $C^1UC^2UC^3$ представлено распределение количества своих (Ряд 1) и чужих (Ряд 2) элементов на отрезке $[Dk_{\min}, Dk_{\max}]$.

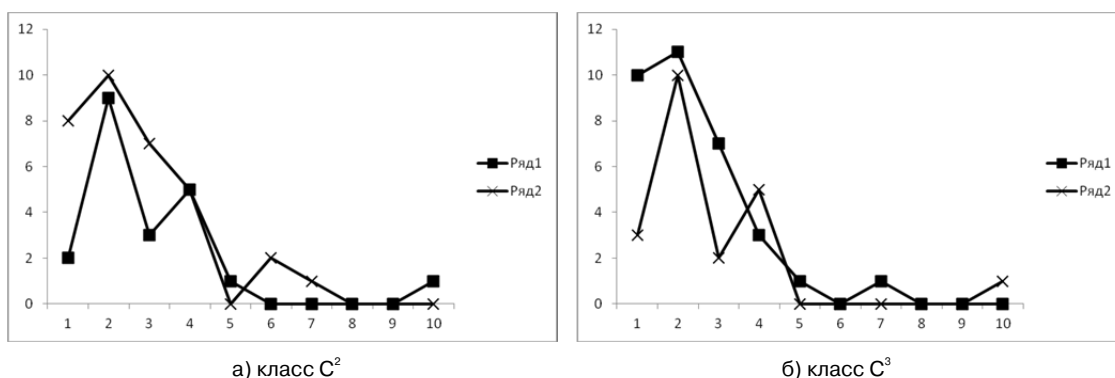
Характер распределения для классов C^1 и C^3 (Рис. 5а, в) отличается от C^2 (Рис. 5б). Для C^2 (Рис. 5б) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$, в отличие от C^1 и C^3 (Рис. 5а, в).

Пример 2. Для каждого класса (Рис. 6 а, б) соответственно из объединения C^1UC^2 представлено распределение числа своих (Ряд 1) и чужих (Ряд 2) элементов на отрезке $[Dk_{\min}, Dk_{\max}]$.

Для C^2 (Рис. 6 б) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$, в отличие от C^1 (Рис. 6 а).

Пример 3. По множеству C^1UC^3 имеем: для C^1 (Рис. 7а) максимум распределения количества своих элементов находится дальше от центра масс, чем аналогичная величина по чужим элементам. Для C^1 (Рис. 7 а) свои и чужие элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$. Для C^3 (Рис. 7б) максимум распределения числа своих элементов совпадает с левой границей $[Dk_{\min}, Dk_{\max}]$, а аналог по чужим элементам немонотонен: один локальный максимум находится вблизи левого края этого отрезка, а второй – на его середине. Заметим, что эта немонотонность не вызвана излишним дроблением рассматриваемого отрезка при небольшом количестве элементов базы. Она является проявлением непростой структуры множества C^1UC^3 . Вследствие этого для разделения классов C^1 и C^3 требуется усложнение полинома и соответственно увеличение его длины.

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

Рис. 8. Распределение числа своих и чужих элементов при удалении от центра масс (C^2UC^3)

Пример 4. В случае множества C^2UC^3 распределение числа своих и чужих элементов для классов C^2 (Рис. 8 а) отличается от C^3 (Рис. 8 б). Для C^2 (Рис. 8 а) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$, в отличие от C^3 (Рис. 8 б). Наблюдается сходство с **Примером 2** (множество C^1UC^2).

Заключение

Разработан классификатор, позволяющий уточнить область локализации опухоли при постановке диагноза в онкологии по показателям периферической крови с использованием статистического метода распознавания, основанного на полиномиальной регрессии. Способ дополняет арсенал уже известных методов постановки диагноза.

Проведено исследование структуры обучающего множества, состоящего из трех классов (C^1 , C^2 , C^3), соответствующих следующим СО для мужчин: пищеварительная система, органы дыхания, опорно-двигательный аппарат. Аналогичная процедура проведена для трех пар, составленных из имеющегося набора.

Для каждого из трех классов C^1 , C^2 , C^3 в отдельности найдено минимальное, максимальное и среднее расстояние между своими векторами (принадлежащими данному классу).

Также получены соответствующие значения для пар свой – чужой (не относится к рассматриваемому классу); эти вычисления зависят от количества классов (два или три), из которых состоит обучающее множество.

Показано, что в указанных терминах множества C^1UC^2 и C^2UC^3 аналогичны по структуре.

Для классов C^1 , C^2 , C^3 вычислили среднестатистический вектор, принадлежащий исходному векторному пространству R^8 (центр масс). Найдено минимальное, максимальное и среднее расстояние между центром масс и своими (чужими) векторами.

При распознавании $C^1UC^2UC^3$ получены ошибки по классу C^1 , причем в перечне альтернатив на первом месте стоит C^3 . Имеющиеся проблемы с классификацией коррелируются с описанной структурой обучающего множества.

Показано, что необходимость использования более сложных полиномов определяется особенностями структуры обучающего множества, относящимися к классам C^1 и C^3 .

Подтверждено сходство структуры множеств C^1UC^2 и C^2UC^3 .

Для каждого класса, входящего в двух- или трехклассовое обучающее множество, получено распределение количества своих и чужих элементов на отрезке их нахождения при удалении от центра масс.

Показано характерное проявление проблемной структуры множества C^1UC^3 . Для разделения классов C^1 и C^3 потребовалось усложнение полинома и соответственно увеличение его длины.

Литература

1. 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.
2. 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.

3. 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
4. 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
5. 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
6. 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.
7. 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. 18с. DOI:10.20948/prepr-2018-126
8. 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. 40с. DOI:10.20948/prepr-2018-255
9. Гавриков М.Б., Локуцкий О.В. Начала численного анализа. — М.: Янус, 1995.
10. Schürmann J. Pattern Classification. — New York: John Wiley&Sons, Inc., 1996.

Гавриков Борис Михайлович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Аспирант. Количество печатных работ: 25 (в т.ч.3 монографии). Область научных интересов: вычислительная математика, распознавание образов, медицинская физика. E-mail: bmgavrikov@gmail.com

Пестрякова Надежда Владимировна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Ведущий научный сотрудник, доктор технических наук. Количество печатных работ: более 100 (в т.ч.1 монография). Область научных интересов: вычислительная математика и физика, распознавание образов. E-mail: pestryakova@isa.ru, bmgavrikov@gmail.com

Statistical Classifier for Diagnosing Oncological Diseases

B. M. Gavrikov, N. V. Pestryakova

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. An approach to the diagnosis of human oncological diseases of various localization is described. For this purpose, a statistical classifier developed by the authors is used, which is based on a polynomial-regression approach and has probabilistic estimates. Training was conducted on databases of analysis of peripheral blood of cancer patients for various body systems.

Keywords: cancer, body system, peripheral blood, classification, polynomial regression, learning set.

DOI 10.14357/20718632230104

References

1. R.V.Stavitskii, L.A.Lebedev, A.L.Lebedev, A.IU.Smyslov. Kolichestvennaia otsenka gomeostaticheskoi aktivnosti zdorovykh i bolnykh liudei - M.: GART. 2013. 131 s.
2. B.M. Gavrikov, I.M. Lebedenko, N.V. Pestryakova, R.V. Stavitskiy. Ob odnom statisticheskom metode otsenivaniya sostoyaniya zdorov'ya cheloveka. // Trudy ISA RAN, 2016. T. 66. № 2. S. 54-59.
3. Gavrikov B.M., Pestryakova N.V. O postroyenii priznakovogo prostranstva v zadache obucheniya // Informatsionnyye tekhnologii i vychislitel'nyye sistemy. 2018. №1. S. 22-29. DOI: 10.14357/20718632180104
4. B.M. Gavrikov, N.V. Pestryakova, R.V.Stavitskiy. O svoystvakh obuchayushchikh mnozhestv // Informatsionnyye tekhnologii i vychislitel'nyye sistemy. 2018. №4. S.97-107. DOI: 10.14357/207186321804010
5. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. Statisticheskii metod raspoznavaniya na osnove nelineynoy regressii. // Matematicheskoye modelirovaniye. 2020. T.32. №4. S.116-130. DOI: 0.20948/mm-2020-04-09
6. Gavrikov B. M., Gavrikov M. B., Pestryakova N. V. O sposobnosti statisticheskogo klassifikatora k obobshcheniyam// Informatsionnyye tekhnologii i vychislitel'nyye sistemy. 2021. № 4. S. 38-50. DOI: 10.14357/20718632210404.
7. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. O strukture bazy obucheniya klassifikatora dlya otsenivaniya

- sostoyaniya zdorov'ya cheloveka // Preprinty IPM im. M.V.Keldysha. 2018. №126. 18s. DOI:10.20948/prepr-2018-126
8. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V., Stavitskiy R.V. Struktura bazy obucheniya statisticheskogo klassifikatora sostoyaniy sistem organizma cheloveka // Preprinty IPM im. M.V.Keldysha. 2018. №255. 40s. DOI:10.20948/prepr-2018-255
9. Gavrikov M.B., Lokutsiyevskiy O.V. Nachala chislennogo analiza. — M.: Yanus, 1995.
10. Schürmann J. Pattern Slassification. — New York: John Wiley&Sons, Inc. 1996.

Gavrikov B. M. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia, e-mail: bmgavrikov@gmail.com

Pestryakova N. V. Doctor of Technical Sciences, PhD in Physics and Mathematics. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia, e-mail: pestryakova@isa.ru