О свойствах обучающих множеств*

Б.М. Гавриков[I], Н.В. Пестрякова[I], Р.В. Ставицкий[II]

[I]Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия
[II]ФГБУ Российский научный центр рентгенорадиологии МЗ РФ (ФГБУ РНЦРР МЗ РФ)

Аннотация. Рассматривается проблема описания множеств, на которых был обучен классификатор для оценивания состояния здоровья человека по результатам анализа периферической крови, основанный на статистическом методе, использующем полиномиально-регрессионный подход. Предлагается способ исследования структуры обучающего множества для каждой системы организма по четырем классам ее поражения.

Ключевые слова: состояние здоровья человека, система организма, периферическая кровь, классификация, полиномиальная регрессия, обучающее множество.

DOI 10.14357/207186321804010

Введение

В работах по методам классификации (распознавания) наличие базы обучения рассматривается обычно как некая данность, а ее качество не подлежит обсуждению. Однако вполне очевидно, что нельзя говорить о методе, его достоинстве и недостатках в отрыве от описания множества, на котором проходило обучение. Оно не только не является идеальным, но и может быть вообще непригодным для проведения этой процедуры. Результат обучения существенно зависит от его структуры. В то же время канонические принципы формирования и анализа таких множеств отсутствуют.

Рассматривается задача создания методов предварительной диагностики состояния здоровья человека (СЗЧ) по результатам анализа крови, базирующихся на данных, полученных в процессе медицинского обследования значительного количества людей [1].

Применяемый для решения данной задачи подход, основанный на полиномиальной регрессии [2-4], хорошо зарекомендовал себя при распознавании печатных и рукопечатных символов. Он является точным, быстрым, устойчивым к шумам, генерирует монотонные (надежные) оценки, имеющие вероятностную природу. Метод адаптирован для классификации объектов иного происхождения [5]. Требуется выяснить, какими свойствами обладают множества, на которых он обучался.

Проблема заключается в том, что до сих пор отсутствует некая стандартная методология и соответствующий инструментарий для изучения и описания конечных множеств объектов, используемых для обучения. В работе [4] предложен способ исследования обучающих множеств символьных объектов. Здесь он будет развит и применен к наборам данных, имеющим другую природу.

* Работа выполнена при финансовой поддержке РФФИ (грант №16-07-00742 а).

1. Классификатор для оценивания состояния систем организма человека

Разработан классификатор, при помощи которого можно оценивать состояние здоровья различных систем организма (СО) человека по результатам лабораторного анализа периферической крови (из пальца). В основе лежит статистический метод, использующий полиномиально-регрессионный подход и имеющий вероятностные оценки.

По данным анализа крови производится деление на четыре класса, которые соответствуют различным стадиям поражения СО человека:

1 класс – здоровые;

2 класс – начальные отклонения состояния здоровья;

3 класс – выраженное отклонение состояния здоровья;

4 класс – тяжелое заболевание.

Для каждой СО строится свой классификатор с использованием отдельного обучающего множества [1, 5]. Наборы показателей крови практически здоровых людей одинаковы у всех СО. Базы данных для мужчин и женщин различаются.

Приведем общепринятые обозначения и размерность показателей крови:

RBC [L^{-1}] – эритроциты,
HGB [g L^{-1}] – гемоглобин,
PLT [L^{-1}] – тромбоциты,
WBC [L^{-1}] – лейкоциты,
LIMPH [L^{-1}], [%] – лимфоциты,
GRAN [L^{-1}], [%] – гранулоциты
(GRAN = NEUT + EOS + BASO).

Для фиксированной СО и пола определим, какой из четырех градаций СЗЧ соответствует результат анализа периферической крови. Перечень градаций СЗЧ есть множество с $K = 4$ элементами. Введем вектор $\mathbf{v}\in\mathbf{R}^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови, причем $N = 8$.

Отождествляем k-й элемент множества градаций СЗЧ с базисным вектором $\mathbf{e}_k=(0\ldots1\ldots0)$ (здесь 1 находится на k-м месте, причем $1\le k\le K$) из $\mathbf{R}^K$. Обозначаем $Y=\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$.

Пусть можно найти $p_k(\mathbf{v})$ – вероятность того, что набор отнормированных показателей крови соответствует k–му элементу СЗЧ, где $1\le k\le K$. Искомый элемент СЗЧ будет иметь порядковый номер, получивший максимальное значение вероятности:

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

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:

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

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:

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

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

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

При получении правой части (4) используется рекуррентная процедура [2-5].

Рассматривались различные модификации вектора $\mathbf{x}(\mathbf{v})$. Указаны СО и пол, для которых при $\mathbf{x}(\mathbf{v})$ данного вида были получены наилучшие результаты и более сложные модификации не использовались. Точность классификации приведена в случаях, когда она меньше 100%.

1). Длина полинома 33.

$$\mathbf{x}=(1,\{v_i\},\{v_i^2\},\{v_i^3\},\{v_i^4\}),\ 1\le i\le 8. \qquad (5)$$

Имеются мономы степенного вида первого, второго, третьего и четвертого порядка. Перекрестные произведения отсутствуют.

2). Длина полинома 61. *ЦНС, органы чувствительности (мужчины). Печень и желчевыводящие пути (мужчины).*

$$\mathbf{x}=(1,\{v_i\},\{v_i^3\},\{v_i^4\},\{v_i v_j\}),\ 1\le i\le 8,\ i\le j\le 8. \quad (6)$$

Имеются мономы степенного вида первого, второго, третьего и четвертого порядка. Перекрестные произведения используются в качестве мономов второго порядка, а для более высоких порядков отсутствуют.

3). Длина полинома 69. *Органы дыхания (женщины). ЦНС, органы чувствительности (женщины). Гинекологическая система (женщины).*

$$\mathbf{x}=(1,\{v_i\},\{v_i^3\},\{v_i^4\},\{v_i^5\},\{v_i v_j\}),\ 1\le i\le 8,\ i\le j\le 8. \quad (7)$$

Имеются мономы степенного вида первого, второго, третьего, четвертого и пятого порядка. Перекрестные произведения используются в качестве мономов второго порядка, а для более высоких порядков отсутствуют.

4). Длина полинома 77. *Опорно-двигательный аппарат (мужчины).*

$$\mathbf{x}=(1,\{v_i\},\{v_i^3\},\{v_i^4\},\{v_i^5\},\{v_i^6\},\{v_i v_j\}),\ 1\le i\le 8,\ i\le j\le 8. \quad (8)$$

Имеются мономы степенного вида первого, второго, третьего, четвертого, пятого и шестого порядка. Перекрестные произведения используются в качестве мономов второго порядка, а для более высоких порядков отсутствуют.

5). Длина полинома 85. *Печень и желчевыводящие пути (женщины).*

$$\mathbf{x}=(1,\{v_i\},\{v_i^3\},\{v_i^4\},\{v_i^5\},\{v_i^6\},\{v_i^7\},\{v_i v_j\}),\ 1\le i\le 8,\ i\le j\le 8. \quad (9)$$

Имеются мономы степенного вида первого, второго, третьего, четвертого, пятого, шестого и седьмого порядка. Перекрестные произведения используются в качестве мономов второго порядка, а для более высоких порядков отсутствуют.

6). Длина полинома 165. *Пищеварительная система (женщины). Органы дыхания (мужчины). Опорно-двигательный аппарат (женщины). Эндокринная система (женщины). Грудные железы (женщины).*

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\}),\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8. \quad (10)$$

Имеются мономы первого, второго и третьего порядка. Перекрестные произведения используются для мономов второго и третьего порядка.

7). Длина полинома 495.

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\}),\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8. \quad (11)$$

Имеются мономы первого, второго, третьего и четвертого порядка. Перекрестные произведения используются для мономов второго, третьего и четвертого порядка.

8). Длина полинома 1287. *Пищеварительная система (мужчины – 98,2%). Урологическая система (женщины – 99,2%, мужчины – 97,3%). Эндокринная система (мужчины – 95%).*

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\},\{v_i v_j v_k v_l v_m\}),\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8,\ l\le m\le 8. \quad (12)$$

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

В (5 - 12) выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычисляемым по всем показателям крови из имеющегося набора.

Точность классификации на обучающих множествах для различных систем организма находится в диапазоне 95 – 100 %. Элементы класса «1» распознаются для всех СО при использовании полиномов наиболее простой структуры и минимальной длины.

2. Диапазоны расстояний между «своими» и «чужими» элементами

Пусть имеется обучающее множество некоторой СО и заданного пола.

Для каждого из четырех классов здоровья 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:

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

Максимальное расстояние:

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

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:

$$U\_k_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=j+1}^{J_k} \| \mathbf{w}^{k,j} - \mathbf{w}^{k,j1} \| / (J_k\ (J_k - 1)/2), \quad \mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k, \tag{15}$$

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 для пар свой – чужой по каждому из классов. Чужой вектор – не принадлежащий рассматриваемому классу.

Минимальное расстояние:

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

Максимальное расстояние:

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

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: $V^{\neg k} = V \setminus V^k$.

Среднее расстояние:

$$U\_kz_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=1}^{J_{\neg k}} \| \mathbf{w}^{k,j} - \mathbf{w}^{\neg k,j1} \| / (J_k\ J_{\neg k}), \quad \mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k, \quad \mathbf{w}^{\neg k,j1} \in V^{\neg k}, j1 = 1, \ldots, J_{\neg k}, \tag{18}$$

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: $\{\mathbf{w}^{\neg k,j1}, j1 = 1, \ldots, J_{\neg k}\} = V^{\neg k}$, $V^{\neg k} = V \setminus V^k$.

Везде ниже будет рассматриваться пищеварительная система для женщин. Обучающее множество в этом случае содержит 77 элементов. Классы «1» и «4» содержат по 24 набора крови, а классы «2» и «3» включают соответственно 9 и 20 элементов. Продемонстрируем, какие значения принимают перечисленные величины на этом примере. Расстояние между векторами определяем в метрике L_2.

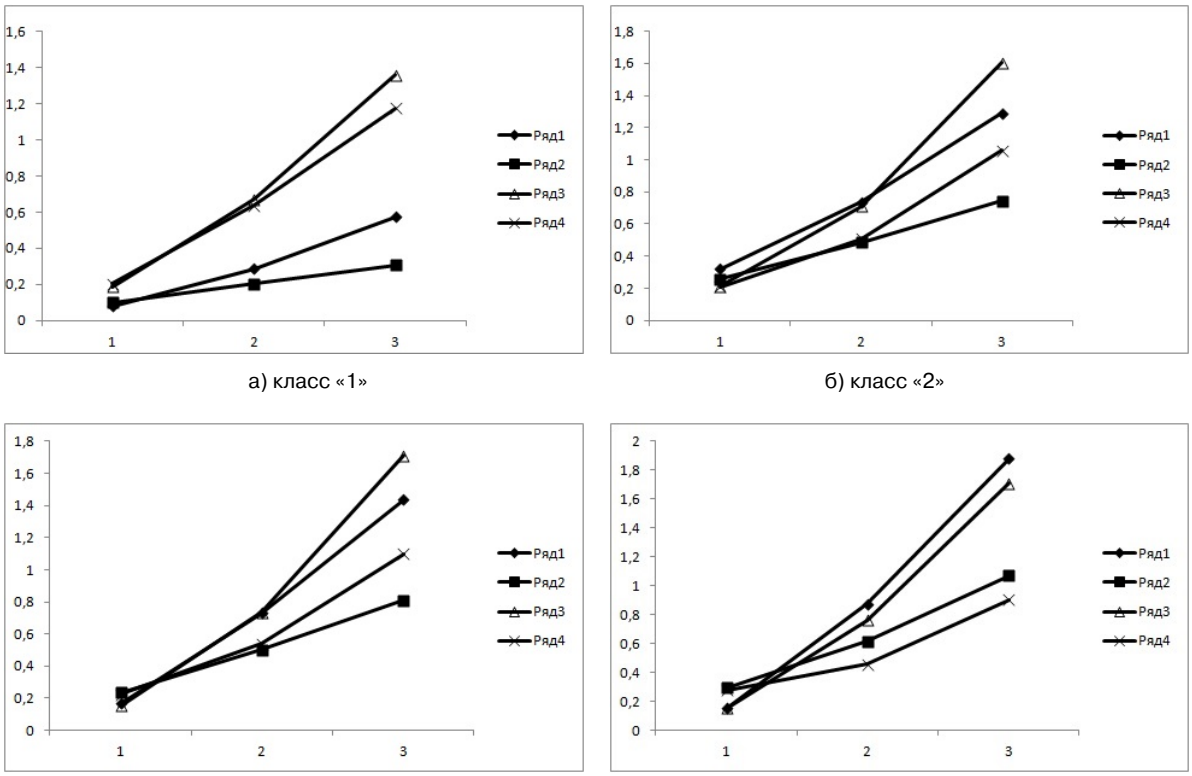
На Рис. 1 (а, б, в, г) соответственно для классов «1», «2», «3», «4» представлено минимальное, среднее и максимальное расстояние (значения ординат для точек 1, 2, 3 по оси абсцисс) между своими векторами (Ряд 1), аналогичные величины для пар свой - чужой (Ряд 3).

Как видно на Рис. 1, а для класса «1», минимальное, максимальное и среднее расстояние между своими векторами (Ряд 1), меньше значений соответствующих расстояний между парами свой - чужой (Ряд 3). Из Рис. 1 (б, в, г) следует, что этот порядок или существенно нарушается для классов «2», «3», «4», или соответствующие величины сближаются, разница их значений уменьшается. Например, по Рис. 1, б для класса «2» видно, что минимальное расстояние между своими векторами больше аналогичной величины для расстояния между своими и чужими векторами. Указанные средние величины примерно одинаковы как для класса «2», так и для класса «3» (Рис. 1, б, в). В то же время для класса «4» как среднее, так и максимальное расстояние между своими векторами больше, чем для пар свой - чужой (Рис. 1, г).

Элементы класса «1» распознаются при использовании полиномов более простой структуры с меньшими значениями их длины, чем для классов «2», «3», «4», что вполне соответствует отмеченным здесь закономерностям взаимного расположения своих и чужих элементов.

3. Диапазоны удаления от центра масс своих и чужих элементов

Для каждого из четырех классов здоровья в отдельности получим среднестатистический вектор длины 8, принадлежащий исходному



а) класс «1» б) класс «2»

в) класс «3» г) класс «4»

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

векторному пространству $\mathbf{R}^8$. Иногда такой вектор называют центром масс.

Для центра масс *k*-го класса СЗЧ 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:

$$v_i^{k,cp} = (\sum_{j=1}^{J_k} v_i^{k,j}) / J_k, \quad (19)$$

где $\mathbf{v}^{k,j}$ – перенумерованные элементы *k*-го класса: $\{\mathbf{v}^{k,j}=(v^{k,j}{}_1,\ldots,v^{k,j}{}_N), j = 1, \ldots, J_k\} = V^k$. Рассматриваемое обучающее множество, содержащее элементы всех четырех классов, суть $V = \{V^1 \cup V^2 \cup V^3 \cup V^4\}$.

Для каждого из четырех классов здоровья в отдельности 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:

$$D\_k_{\min} = \min_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\}, \mathbf{u}^k \in V^k. \quad (20)$$

Максимальное расстояние:

$$D\_k_{\max} = \max_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\}, \mathbf{u}^k \in V^k, \quad (21)$$

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:

$$D\_k_{\text{ср}} = \sum_{j=1}^{J_k} \|\mathbf{w}^{k,j} - \mathbf{v}^{k,cp}\| / J_k, \quad \mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k. \quad (23)$$

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. Минимальное расстояние:

$$D\_kz_{\min} = \min_{V^{\neg k}} \{\| \mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ , \mathbf{u}^{\neg k} \in V^{\neg k}. \qquad (23)$$

Максимальное расстояние:

$$D\_kz_{\max} = \max_{V^{\neg k}} \{\| \mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ , \mathbf{u}^{\neg k} \in V^{\neg k}, \qquad (24)$$

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:

$$D\_kz_{\text{ср}} = \sum_{j=1}^{J_{\neg k}} \| \mathbf{v}^{k,cp} - \mathbf{w}^{\neg k,j} \| / J_{\neg k}\,, \quad \mathbf{w}^{\neg k,j} \in V^{\neg k}, j = 1, \ldots, J_{\neg k}, \qquad (25)$$

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**.**

На Рис. 1 (а, б, в, г) соответственно для классов «1», «2», «3», «4» представлено минимальное, максимальное и среднее 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).

Как следует из Рис. 1, а для класса «1», минимальное, среднее и максимальное расстояние между центром масс и своими векторами (Ряд 2) меньше значений соответствующих расстояний между центром масс и чужими векторами (Ряд 4). По Рис. 1 (б, в, г) видно, что это соотношение либо в значительной степени нарушается для классов «2», «3», «4», либо различие между указанными величинами становится меньше. Так из Рис. 1, б для класса «2» следует, что минимальное расстояние между центром масс и своими векторами немного больше соответствующей величины для расстояния между центром масс и чужими векторами, а указанные средние величины примерно одинаковы. В то же время для класса «4» как среднее, так и максимальное расстояние между центром масс и своими векторами больше, чем между центром масс и чужими векторами (Рис. 1, г).

Для распознавания элементов класса «1» достаточно более простых и коротких полиномов, в отличие от классов «2», «3», «4», что согласуется с приведенными закономерностями расположения своих и чужих элементов по отношению к центру масс.

4. Ближайшие элементы. Распределения числа своих и чужих элементов

Диапазон расстояний между своими элементами k-го класса СЗЧ по рассматриваемой базе, согласно формулам (13), (14), находится на отрезке $[U\_k_{\min}, U\_k_{\max}]$. Перенумеруем все векторы этого класса: $\{\mathbf{w}^{k,j}, j = 1, \ldots, J_k\} = V^k$. Для каждого такого элемента $\mathbf{w}^{k,j}$ найдем расстояние до ближайшего вектора этого же класса (своего):

$$U\_k^j = \min_{V^k \setminus \mathbf{w}^{k,j}} \{\|\mathbf{w}^{k,j} - \mathbf{u}^k\|\}, \quad \mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k\,, \mathbf{u}^k \in \{V^k \setminus \mathbf{w}^{k,j}\}. \qquad (26)$$

Заметим, что один и тот же вектор $\mathbf{u}^k$ может оказаться ближайшим более чем для одного элемента, например, для $\mathbf{w}^{k,j1}$ и $\mathbf{w}^{k,j2}$, где $j1 \neq j2$. В то же время, для элемента $\mathbf{w}^{k,j}$ ближайшими могут быть одновременно несколько различных равноудаленных от него векторов, например $\mathbf{v}^k \in \{V^k \setminus \mathbf{w}^{k,j}\}$, $\mathbf{u}^k \in \{V^k \setminus \mathbf{w}^{k,j}\}$, $U\_k^j = \|\mathbf{w}^{k,j} - \mathbf{v}^k\| = \|\mathbf{w}^{k,j} - \mathbf{u}^k\|$, $\mathbf{v}^k \neq \mathbf{u}^k$.

Делим отрезок $[U\_k_{\min}, U\_k_{\max}]$ на десять равных по длине частей – один отрезок и девять полуинтервалов: $[U\_k_{\min}, U\_k_{\min} + udk]$, $(U\_k_{\min} + udk, U\_k_{\min} + 2udk]$, … , $(U\_k_{\min} + 9udk, U\_k_{\min} + 10udk]$, где $udk = (U\_k_{\max} - U\_k_{\min})/10$. Определим для каждого j, к какой из десяти этих частей относится величина U_k^j. Затем сосчитаем, какое количество перенумерованных векторов $\mathbf{w}^{k,j}$ попало в каждый такой участок, то есть имеет ближайший вектор из этого же класса, расстояние до которого лежит на этой части отрезка. Итак, мы получим распределение числа своих векторов $\mathbf{w}^{k,j}$ по расстоянию до

ближайшего своего вектора на отрезке [$U\_k_{\min}$, $U\_k_{\max}$] возможных расстояний между своими векторами.

Диапазон расстояний между элементами k-го класса и векторами всех других классов (чужими), согласно формулам (16), (17) находится на отрезке [$U\_kz_{\min}$, $U\_kz_{\max}$]. Перенумеруем все элементы рассматриваемого k-го класса: $\{\mathbf{w}^{k,j}, j = 1, \ldots, J_k\}=V^k$. Для каждого такого элемента $\mathbf{w}^{k,j}$ найдем расстояние до ближайшего вектора, не принадлежащего этому классу (чужого):

$$U\_kz^j = \min_{V^{\neg k}} \{\| \mathbf{w}^{k,j} - \mathbf{u}^{\neg k}\|\}, \quad (27)$$

$$\mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k, \mathbf{u}^{\neg k} \in V^{\neg k}.$$

Заметим, что один и тот же вектор $\mathbf{u}^{\neg k}$ может оказаться ближайшим более чем для одного элемента, например, для $\mathbf{w}^{k,j1}$ и $\mathbf{w}^{k,j2}$, где $j1 \neq j2$. В то же время для элемента $\mathbf{w}^{k,j}$ ближайшими могут быть одновременно несколько различных равноудаленных от него векторов, например, $\mathbf{v}^{\neg k} \in V^{\neg k}$, $\mathbf{u}^{\neg k} \in V^{\neg k}$, $U\_kz^j = \|\mathbf{w}^{k,j} - \mathbf{v}^{\neg k}\| = \|\mathbf{w}^{k,j} - \mathbf{u}^{\neg k}\|$, $\mathbf{v}^{\neg k} \neq \mathbf{u}^{\neg k}$.

Делим отрезок [$U\_kz_{\min}$, $U\_kz_{\max}$] на десять равных по длине частей – один отрезок и девять полуинтервалов: [$U\_kz_{\min}$, $U\_kz_{\min} + udkz$], ($U\_kz_{\min} + udk$, $U\_kz_{\min} + 2udkz$], … , ($U\_kz_{\min} + 9udkz$, $U\_kz_{\min} + 10udkz$], где $udkz = (U\_kz_{\max} - U\_kz_{\min})/10$. Определим для каждого j, к какой из десяти этих частей относится величина U_kz^j. Затем посчитаем, какое количество перенумерованных векторов $\mathbf{w}^{k,j}$ попало в каждый такой участок, то есть имеет ближайший вектор, не принадлежащий этому классу, расстояние до которого лежит на этой части отрезка. Итак, мы получим распределение числа своих векторов $\mathbf{w}^{k,j}$ по расстоянию до ближайшего чужого вектора на отрезке [$U\_kz_{\min}$, $U\_kz_{\max}$] возможных расстояний между своими и чужими векторами.

Диапазон расстояний между центром масс k-го класса и векторами этого же класса (своими), согласно формулам (20), (21), находится на отрезке [$D\_k_{\min}$, $D\_k_{\max}$]. Как ранее было описано, разделим этот отрезок на десять равных по длине частей (один отрезок и девять полуинтервалов), и определим, какое количество своих векторов попало в каждый такой участок. Затем рассмотрим распределение числа своих векторов на отрезке [$D\_k_{\min}$, $D\_k_{\max}$].

Диапазон расстояний между центром масс k-го класса СЗЧ и чужими векторами, согласно формулам (23), (24), занимает отрезок [$D\_kz_{\min}$, $D\_kz_{\max}$]. Аналогичным образом разделим данный отрезок на десять равных по длине частей (один отрезок и девять полуинтервалов), и определим, какое количество чужих векторов попало в каждый такой участок. Затем рассмотрим распределение числа чужих векторов на отрезке [$D\_kz_{\min}$, $D\_kz_{\max}$].

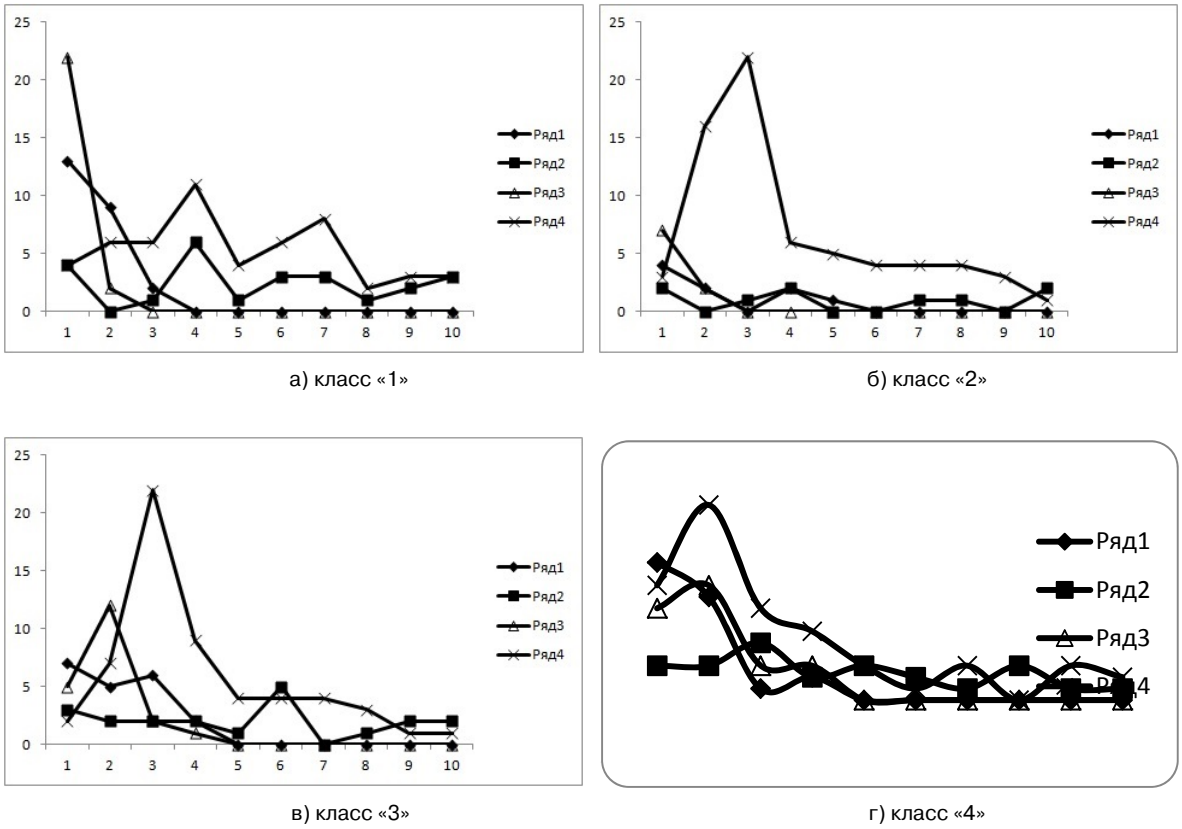
Совместим на оси абсцисс Рис. 2 все эти отрезки: [$U\_k_{\min}$, $U\_k_{\max}$], [$U\_kz_{\min}$, $U\_kz_{\max}$], [$D\_k_{\min}$, $D\_k_{\max}$], [$D\_kz_{\min}$, $D\_kz_{\max}$] так, чтобы их начальные точки совпали между собой, аналогичное условие выполним и для концов.

На Рис. 2 (а, б, в, г) соответственно по классам «1», «2», «3», «4» представлено распределение числа элементов k-го класса по расстоянию до ближайшего своего вектора на отрезке [$U\_k_{\min}$, $U\_k_{\max}$] (Ряд 1), распределение количества элементов этого класса по расстоянию до ближайшего чужого вектора на отрезке [$U\_kz_{\min}$, $U\_kz_{\max}$] (Ряд 3), распределение числа своих векторов по расстоянию до центра масс на отрезке [$D\_k_{\min}$, $D\_k_{\max}$] (Ряд 2), распределение числа чужих элементов по расстоянию до центра масс на отрезке [$D\_kz_{\min}$, $D\_kz_{\max}$] (Ряд 4).

Если рассмотреть Ряд 1 и Ряд 3 на Рис. 1 а и Рис. 2 а, то можно заметить, что максимальное для класса 1 расстояние от подавляющего большинства векторов этого класса до ближайшего своего элемента меньше, чем минимальное расстояние от каждого вектора этого класса до ближайшего чужого элемента; в частности, для подавляющего большинства векторов класса 1 расстояние до ближайшего своего элемента меньше, чем до ближайшего чужого.

Для классов 2, 3, 4 из сравнения значений ординат Ряда 1 и Ряда 3 в точке 1 на Рис. 1 (а, б, в), а именно, $U\_k_{\min} > U\_kz_{\min}$, следует, что для части векторов этих классов расстояние до ближайшего своего элемента больше, чем до ближайшего чужого.

Эти закономерности согласуются с фактом, что для распознавания элементов класса «1» достаточно более простых полиномов, в отличие от классов «2», «3», «4».



а) класс «1» б) класс «2»

в) класс «3» г) класс «4»

Рис. 2. Распределение числа своих элементов по расстоянию до ближайшего своего и чужого вектора на отрезках $[U\_k_{min}, U\_k_{max}]$ и $[U\_kz_{min}, U\_kz_{max}]$, числа своих и чужих элементов по расстоянию до центра масс на отрезках $[D\_k_{min}, D\_k_{max}]$ и $[D\_kz_{min}, D\_kz_{max}]$

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

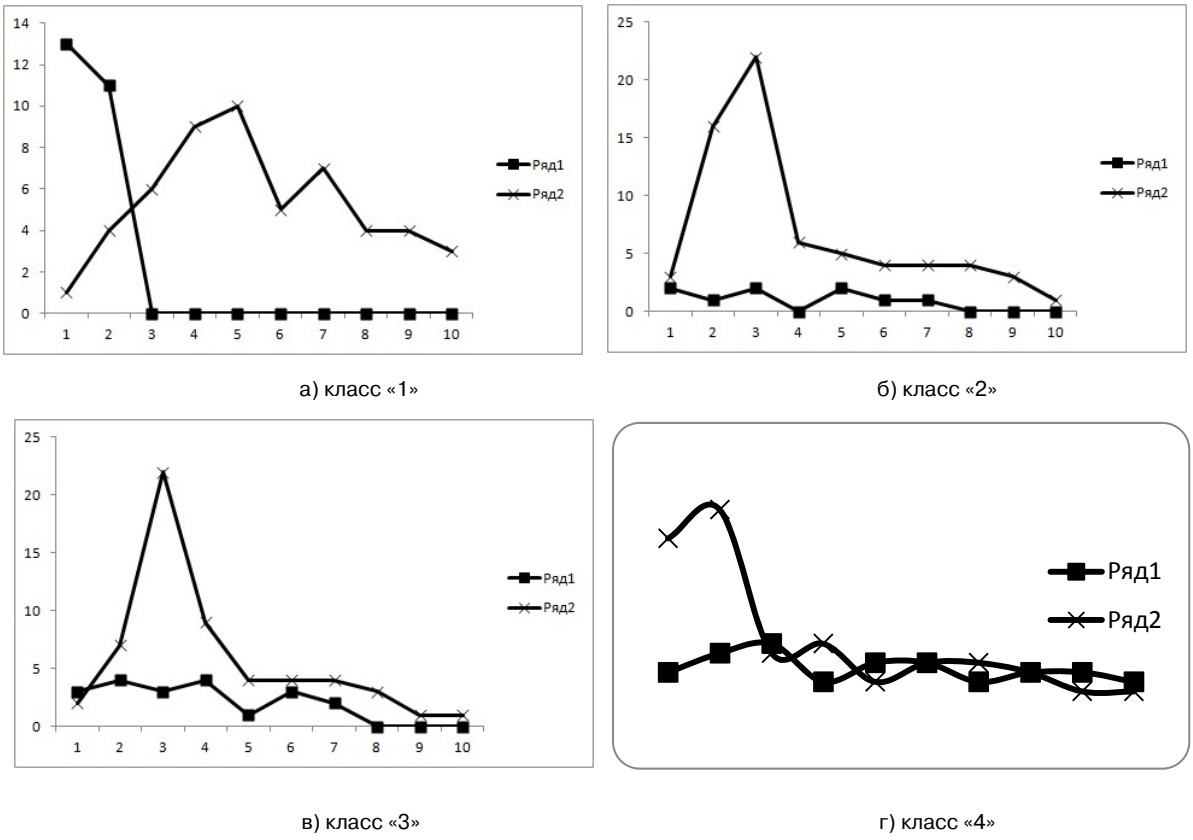
Диапазон расстояний между центром масс k-го класса СЗЧ и векторами этого же класса («своими», $\mathbf{v}^k \in V^k$) по рассматриваемой базе, согласно формулам (20), (21), находится на отрезке $[D\_k_{min}, D\_k_{max}]$. Диапазон расстояний между центром масс k-го класса СЗЧ и векторами всех других классов («чужими», $\mathbf{z}^k \in \{V \setminus V^k\}$), согласно формулам (23), (24), – на отрезке $[D\_kz_{min}, D\_kz_{max}]$. Пусть

$$Dk_{min} = \min(D\_k_{min}, D\_kz_{min})$$
$$Dk_{max} = \max(D\_k_{max}, D\_kz_{max}). \quad (28)$$

Делим отрезок $[Dk_{min}, Dk_{max}]$ (оси абсцисс на Рис. 3 (а, б, в, г) 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 $[Dk_{min}, Dk_{max}]$.

На Рис. 3 (а, б, в, г) соответственно для классов «1», «2», «3», «4» представлено распределение числа своих (Ряд 1) и чужих (Ряд 2) элементов на отрезке $[Dk_{min}, Dk_{max}]$.

Как нетрудно заметить, картина этих двух распределений на Рис. 3 а принципиально отличается от изображенных на остальных рисунках. А именно, вблизи центра масс элементов класса «1» имеется относительно небольшая окрестность, в которой находятся все элементы этого класса, причем их число убывает при удалении от центра масс. В то же время в этой окрестности есть небольшое коли-



а) класс «1» б) класс «2»

в) класс «3» г) класс «4»

Рис. 3. Распределение числа своих и чужих элементов при удалении от центра масс

чество чужих элементов, а подавляющее большинство их находится вне нее. Соответствующая функция распределения сначала нарастает при удалении от центра масс, а затем имеет тенденцию к убыванию, и ее максимум находится на удалении от этой окрестности, где отсутствуют элементы класса «1».

На Рис. 3 (б, в, г) вид распределений своих и чужих элементов для классов «2», «3», «4» совершенно иной. Свои элементы имеются до конца (Рис. 3 г) или почти до конца (Рис. 3 б, в) отрезка $[Dk_{\min}, Dk_{\max}]$, а в окрестности его начальной точки отсутствует заметный максимум соответствующей функции распределения. На этом отрезке число чужих элементов превышает или сопоставимо с числом своих. Поведение функции распределения чужих элементов схоже с Рис. 3 а), однако максимум этой функции значительно ближе к центру масс.

Описанная картина распределений количества своих и чужих элементов по классам «1», «2», «3», «4» может служить объяснением того факта, что для распознавания элементов класса «1» (здоровые) достаточно использовать полиномы, имеющие минимальную длину. Классы «2», «3», «4» являются более проблемными в этом отношении. Для их распознавания требуется усложнение структуры и соответственно увеличение длины полинома.

Заключение

Разработан классификатор СЗЧ по показателям периферической крови из пальца для различных СО с использованием статистического метода распознавания, основанного на полиномиальной регрессии.

Проведено исследование структуры обучающего множества для одной из СО, пищеварительной системы для женщин, по четырем классам ее поражения.

Показано, что для класса «1» минимальное, максимальное и среднее расстояние между своими векторами меньше значений соответствующих расстояний между парами свой - чужой. Этот порядок или существенно нарушается для классов «2», «3», «4», или

соответствующие величины сближаются, разница их значений уменьшается.

Для класса «1», минимальное, максимальное и среднее расстояние между центром масс и своими векторами меньше значений соответствующих расстояний между центром масс и чужими векторами. Это соотношение либо в значительной степени нарушается для классов «2», «3», «4», либо различие между указанными величинами становится меньше.

Элементы класса «1» распознаются при использовании полиномов более простой структуры с меньшими значениями их длины, чем для классов «2», «3», «4», что соответствует отмеченным здесь закономерностям, относящимся к взаимному расположению своих и чужих элементов, расположению своих и чужих элементов по отношению к центру масс, а также картине распределения количества своих и чужих элементов по классам «1», «2», «3», «4» при удалении от центра масс.

Литература

1. Количественная оценка гомеостатической активности здоровых и больных людей / Ставицкий Р.В. [и др.]. // М.: ГАРТ. 2013. 131 с.
2. Гавриков М.Б., Пестрякова Н.В. Метод полиномиальной регрессии в задачах распознавания печатных и рукопечатных символов. // Препринты ИПМ им.М.В.Келдыша. 2004. № 22. 12 с.
3. Об одном методе распознавания символов, основанном на полиномиальной регрессии / Гавриков М.Б. [и др.]. // Автоматика и Телемеханика. 2006. № 2. С. 119-134.
4. Пестрякова Н.В. Метод распознавания символов, основанный на полиномиальной регрессии. М.: УРСС. 2011. 141 с.
5. Гавриков Б.М., Пестрякова Н.В. О построении признакового пространства в задаче обучения. // Информационные технологии и вычислительные системы. 2018. №1. С.22-29.

Пестрякова Надежда Владимировна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва, Россия. Ведущий научный сотрудник, доктор технических наук. Количество печатных работ: более 90 (в т.ч.1 монография). Область научных интересов: вычислительная математика и физика, распознавание образов. E-mail: pestryakova@isa.ru

Гавриков Борис Михайлович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва, Россия. Аспирант. Количество печатных работ: 14 (в т.ч. 3 монографии). Область научных интересов: вычислительная математика, распознавание образов, медицинская физика. E-mail: bmgavrikov@gmail.com

Ставицкий Роман Владимирович. ФГБУ Российский научный центр рентгенорадиологии МЗ РФ (ФГБУ РНЦРР МЗ РФ), Москва, Россия. Главный научный сотрудник, доктор биологических наук, профессор. Количество печатных работ: 133 (в т. ч. 38 монографий). Область научных интересов: медицинская физика, радиология, биология, вычислительная математика и физика, распознавание образов.

On the properties of learning sets

B.M. Gavrikov[I], N. V. Pestryakova[I], R.V. Stavitskiy[II]

[I] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
[II] Russian Scientific Center of Roentgenoradiology, Moscow, Russia

The problem of describing the sets on which the classifier was trained to evaluate the state of human health from the results of the analysis of peripheral blood based on a statistical method using the polynomial regression approach is considered,. A method for studying the structure of a training set for each system of the body according to four classes of its defeat is proposed.

Keywords: state of human health, peripheral blood, classification, polynomial regression, training set

DOI 10.14357/207186321804010

References

1. Stavitskii, R.V. [et al.]. 2013. Kolichestvennaya otsenka gomeostaticheskoy aktivnosti zdorovih i bol’nih lyudey. [Quantitative estimation of homeostatic activity in healthy and sick people]. Moscow, GART. 131p.

2. Gavrikov, M.B., and N.V. Pestryakova 2004. Keldysh Institute Preprints. № 22. Metod polinomial'noy regressii v zadachah raspoznavaniya pechatnih i rukopechatnih simvolov. [Polynomial regression method in pattern recognition of printed and handprinted characters]. 12p.
3. Gavrikov, M. B. [et al.]. Automation and Remote Control. –2006. On a Pattern Recognition Method Based on Polynomial Regression // 67(2). p. 278-292. DOI: 10.1134/S000511790602007X.
4. Pestryakova, N. V. 2011. Metod raspoznavaniya simvolov, osnovannyy na polinomial'noy regressi [The method of character recognition based on polynomial regression]. M: URSS. 141p.
5. Gavrikov, B.M., Pestryakova N. V. 2018. ITVS. O postroyenii priznakovogo prostranstva v zadache obucheniya. [About building the feature space in the problem of learning]. 1:22-29.

N.V. Pestryakova laboratory of artificial intelligence methods ISA FRC CSC RAS. E-mail: pestryakova@isa.ru

B. M. Gavrikov post-graduate courses ISA FRC CSC RAS. E-mail: bmgavrikov@gmail.com

R.V. Stavitskiy Russian Scientific Center of Roentgenoradiology, Ministry of Health of Russian Federation, Moscow. Chief Researcher of the Radiation Therapy Laboratory, Doctor of Biological Sciences, Professor.