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

Препринты ИПМ • Препринт № 126 за 2018 г.

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

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

О структуре базы обучения классификатора для оценивания состояния здоровья человека

Рекомендуемая форма библиографической ссылки: 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
URL: http://library.keldysh.ru/preprint.asp?id=2018-126

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, Н.В. Пестрякова

О структуре базы обучения классификатора для оценивания состояния здоровья человека

Москва — 2018

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

О структуре базы обучения классификатора для оценивания состояния здоровья человека

Разработан классификатор, позволяющий на основании лабораторного анализа периферической крови оценивать состояние здоровья человека по четырем градациям. Используется статистический метод классификации, базирующийся на полиномиально-регрессионном подходе и имеющий вероятностные оценки. Для каждой из систем организма раздельно по мужчинам и женщинам выполнен анализ структуры обучающей базы.

Ключевые слова: состояние здоровья человека, периферическая кровь, классификация, полиномиальная регрессия

Boris Mikhailovich Gavrikov, Mikhail Borisovich Gavrikov, Nadejda Vladimirovna Pestryakova

On the structure of the classifier's training base for assessing the state of human health

A classifier has been developed that allows assessing the state of human health in four grades based on laboratory analysis of peripheral blood. A statistical classification method based on a polynomial-regression approach and having probability estimates is used. For each of the body systems, the analysis of the structure of the training base was carried out separately for men and women.

Key words: state of human health, peripheral blood, classification, polynomial regression

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 16-07-00742-а.

Оглавление

Введение

Решается проблема оценивания состояния здоровья человека (СЗЧ) на основе результатов лабораторного анализа периферической крови (из пальца). Каждая система организма (СО) рассматривается для мужчин и женщин раздельно, поскольку характерные диапазоны вариации показателей крови людей существенно зависят от пола.

Имеется четыре градации СЗЧ: от практически здорового состояния до высшей стадии поражения организма (онкологические заболевания):

1 класс – здоровые;

2 класс – начальные отклонения состояния здоровья;

3 класс – выраженные отклонения состояния здоровья;

4 класс – тяжелые заболевания.

Разработаны версии классификатора по обоим полам для систем дыхания, пищеварительной, урологической, эндокринной, опорно-двигательного аппарата, центральной нервной системы (ЦНС) и органов чувствительности, печени и желчевыводящих путей, а для гинекологической системы и грудных желез по женщинам. Состояние здоровья грудных желез у мужчин не рассматривается ввиду достаточно низкой заболеваемости.

Для конкретной СО из рассматриваемого перечня к каждой из четырех возможных градаций СЗЧ относится список заболеваний, соответствующих этой СО [1].

Статистический метод классификации, базирующийся на полиномиально-регрессионном подходе и имеющий вероятностные оценки, реализованный авторами для распознавания изображений печатных и рукопечатных символов [2-5], адаптирован на рассматриваемые несимвольные объекты. В качестве независимых компонент полиномиального вектора используется восемь показателей крови, отнормированных на отрезок [0,1] по обучающей базе. Наборы комбинаций базисных мономов включают степени и перекрестные произведения степеней независимых компонент.

При обучении используются выборки, полученные в результате детального обследования пациентов большой группой специалистов из различных областей медицины [1]. База показателей крови практически здоровых людей одинакова для всех СО, но значительно различается для мужчин и женщин.

При построении дифференцированных по полу обучающих выборок для каждой из СО используются измеренные на автоматизированном анализаторе крови стабильные показатели. Они перечислены ниже, причем приведены их общепринятые обозначения и размерность:

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).

Ввиду объективных трудностей составления обучающих выборок принцип соразмерности не выполняется в необходимой степени. А именно, для ряда СО некоторые из градаций СЗЧ недостаточно заполнены или вообще отсутствуют. По каждому из полов градация, соответствующая здоровым людям (класс «1»), наиболее обширная и ввиду универсальности используется для всех СО. Следующим по заполненности является класс «4». Наборы выборок показателей крови для классов «2», «3», «4» уникальны для каждой из СО. Предъявляемое к выборкам условие случайности, несомненно, выполняется для параметров крови. Это обусловлено большим разнообразием человеческих организмов.

Математическая постановка задачи

Для конкретной СО по имеющемуся анализу периферической крови пациента известного пола определим, какой из четырех градаций СЗЧ он соответствует. Перечень градаций СЗЧ представляет собой множество с $K = 4$ элементами.

Пространство первичных признаков определяется параметрами крови, которые измеряются несопоставимыми величинами, а поэтому принципиально различаются и по наименованию, и по порядкам величин, и по диапазону вариации. Число независимых компонент полиномиального вектора соответствует количеству используемых при построении классификатора параметров крови и равно восьми в данном исследовании. Их набор определяется типом автоматического анализатора крови и включает существенно большее число наименований. Однако нет необходимости использовать какие-либо параметры крови в дополнение к этим восьми, поскольку они представляют собой их комбинации и не являются независимыми.

Для представления наборов показателей крови вводится вектор $\mathbf{v} \in \mathbf{R}^N$, i-я компонента которого суть отнормированная величина i-го показателя крови, лежащая на отрезке [0,1], причем $N = 8$.

Нормирование на отрезок [0,1] проводим следующим образом. Пусть по рассматриваемой СО имеется обучающая выборка, включающая все градации СЗЧ. 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 – объем выборки по данной СО.

Затем выполняем следующее преобразование:

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

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 r, где

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

Приближенные значения компонент $(p_1(\mathbf{v}),\ldots,p_K(\mathbf{v}))$ представляются в виде многочленов от координат $\mathbf{v}=(v_1,\ldots,v_N)$:

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

Суммы в правых частях равенств (2) конечные и определяются выбором базисных мономов. А именно, если

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

конечный вектор размерности L из приведенных в (2) базисных мономов, упорядоченных некоторым образом и определяющих соответствующее признаковое пространство, тогда (2) можно записать следующим образом:

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

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 $\mathbf{x}(\mathbf{v})$. Следовательно, приближенный поиск вектора вероятностей $\mathbf{p}(\mathbf{v})$ сводится к нахождению матрицы A.

Значение A вычисляется приближенно в процессе обучения, используя содержащиеся в некоторой базе данных наборы пар векторов $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$. Здесь первый элемент $\mathbf{v}^{(j)}$ – набор параметров крови, соответствующий элементу СЗЧ с каким-либо номером k ($1\le k\le K$), а второй элемент – его базисный вектор $\mathbf{y}^{(j)}= (0\ldots1\ldots0)$, 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 следующая рекуррентная процедура, где A_0 и G_0 заданы:

$$A_j = A_{j-1} - \alpha_j G_j \mathbf{x}^{(j)} [A_{j-1}^{\mathrm{T}} \mathbf{x}^{(j)} - \mathbf{y}^{(j)}]^{\mathrm{T}}, \quad \alpha_j = 1/(m\,J/2), \quad m \geq 1$$

$$G_j = \frac{1}{1-\alpha_j} [G_{j-1} - \alpha_j \frac{G_{j-1}\mathbf{x}^{(j)} (\mathbf{x}^{(j)})^{\mathrm{T}} G_{j-1}}{1+\alpha_j ((\mathbf{x}^{(j)})^{\mathrm{T}} G_{j-1} \mathbf{x}^{(j)} - 1)}], \qquad 1 \leq j \leq J \qquad (5)$$

$$G_j \equiv D^{-1}, \quad D = \mathrm{diag}\,(\,\mathrm{E}\{x_1^2\}, \mathrm{E}\{x_2^2\}, \ldots, \mathrm{E}\{x_L^2\}).$$

Здесь $x_1, x_2, \ldots, x_L$ – компоненты вектора $\mathbf{x}(\mathbf{v})$, m – экспериментально подбираемый множитель, приведенный в таблице 1. Получаемые оценки могут выходить за рамки отрезка [0,1] из-за того, что используемый метод является приближенным. Отрицательные значения обнуляли, а превышающие 1 делали равными 1.

Рассматривались различные модификации вектора $\mathbf{x}(\mathbf{v})$. Указанной в таблице 1 длине полинома соответствует следующая его структура.

1). Длина полинома 33.

$$\mathbf{x} = (1, \{v_i\}, \{v_i^2\}, \{v_i^3\}, \{v_i^4\}),\ 1 \leq i \leq 8. \qquad (6)$$

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

2). Длина полинома 61.

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

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.

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

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.

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

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.

$$\mathbf{x} = (1, \{v_i\}, \{v_i^3\}, \{v_i^4\}, \{v_i^5\}, \{v_i^6\}, \{v_i^7\}, \{v_i v_j\}),\ 1 \leq i \leq 8,\ i \leq j \leq 8. \qquad (10)$$

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.

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

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

7). Длина полинома 495.

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

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

8). Длина полинома 1287.

$$\mathbf{x} = (1,\{v_i\},\{v_i\, v_j\},\{v_i\, v_j\, v_k\},\{v_i\, v_j\, v_k\, v_l\},\{v_i\, v_j\, v_k\, v_l\, v_m\}),\quad 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8,\ l\le m\le 8. \qquad (13)$$

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

В (6 - 13) выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычисляемым по всем восьми параметрам крови из имеющегося набора.

Таблица 1

Зависимость точности классификации от структуры полинома для основных систем организма

Пищеварительная система												
м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-109							13	11,9%	2	12	11,0%	2
ж-77	12	15,6%	2	4	5,2%	2	4	5,2%	2	4	5,2%	2
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*

м-109	13	11,9%	2	4	3,7%	2	2	1,8%	4	2	1,8%	8
ж-77	1	0,8%	2	0	0%	2						

Органы дыхания

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-76							4	5,3%	2	1	1,3%	2
ж-62							0	0%	2			
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-76	1	1,3%	2	0	0%	4						

Опорно-двигательный аппарат

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-76				1	1,3%	2				0	0%	2
ж-72				1	1,4%	2	1	1,4%	2	1	1,4%	2
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*									
м-76												
ж-72	1	1,4%	2	0	0%	4						

Урологическая система

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-110										18	16,4%	2
ж-119							6	5,0%	2	4	3,4%	2
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-110	9	8,2%	2	3	2,7%	2	1	0,9%	6	3	2,7%	18
ж-119	расх		2	3	2,5%	4	1	0,8%	16	1	0,8%	60

Эндокринная система

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-121	29	24,0%	2	22	18,2%	4	21	17,4%	4	24	19,8%	3
ж-80	6	7,5%	2	2	2,5%	2	2	2,5%	2	2	2,5%	2
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-121	расх		2	9	7,4%	6	6	5,0%	20	6	5,0%	60
ж-80	2	2,5%	2	0	0%	4						

ЦНС, органы чувствительности

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-61				0	0%	2						
ж-79							0	0%	2			

Печень и желчевыводящие пути

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
м-74				0	0%	2						
ж-87							1	1,1 %	2	1	1,1 %	2
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*									
ж-87	0	0%	2									

Гинекологическая система

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*
ж-56							0	0%	2			

Грудные железы

м/ж-объем базы	*длина полинома*											
	33			61			69			77		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*	*mis*	*mis-%*	*m*

ж-129							4	3,1%	2			
м/ж-объем базы	*длина полинома*											
	85			165			495			1287		
	mis	*mis-%*	*m*	*mis*	*mis-%*	*m*						
ж-129	4	3,1%	2	0	0%	2						

Численная реализация метода оценивания СЗЧ и полученные результаты

Используемые базы показателей крови обладали значительной вариабельностью и небольшим объемом. Кроме того, исходные обучающие множества имели неравноценные по объему подмножества для различных градаций заболевания организма. Поэтому при получении обучающей последовательности использовалась лишь часть объема данных по классу «1» (здоровых людей), сопоставимая с количеством элементов класса «4» (наивысшая степень заболевания), далее добавлялись элементы классов «2» и «3», а затем проводилось перемешивание по всему объему обучающего множества.

Этот прием, конечно, не позволил осуществить приблизительное выравнивание числа элементов для всех четырех классов по каждой СО. Вычислительная практика показала, что используемого количества элементов класса «1» достаточно для безошибочного распознавания всех элементов этого класса, в том числе не включенных в процесс обучения. Более того, они распознаются уже на самых низких значениях длины полинома.

В таблице 2 приведены данные для обоих полов о том, как именно количественно распределены по классам элементы обучающих последовательностей различных СО.

В СО «Эндокринная система» (женщины), «Печень и желчевыводящие пути» (мужчины) нет элементов класса «3». Для женщин в базах СО «Опорно-двигательный аппарат», «Урологическая система» и «Грудные железы» отсутствуют элементы класса «2». Однако классификаторы обучаются и обеспечивает высокую точность при распознавании обучающего множества. В наборах альтернатив распознавания эти классы отсутствуют.

В таблице 1 в левом столбце указан пол («*м»/* «*ж»*) и объем обучающей базы, а в ячейках, озаглавленных «*mis»* и «mis-%» соответственно приведено число ошибок и их доля в процентном выражении по отношению к числу элементов в обучающем множестве.

Наиболее проблемными в отношении точности классификации являются для мужчин пищеварительная, урологическая и эндокринная системы. Они характеризуются высокой заполненностью классов «2» и «3».

С целью повышения точности распознавания проводилось многократное обучение на одной и той же базе с контролем точности распознавания, поскольку при неограниченном увеличении числа таких итераций точность

сначала стабилизируется на некотором минимальном значении, а затем может начинать нарастать. В ряде случаев процесс имеет вид периодических колебаний с одним и тем же значением минимума количества ошибочных распознаваний, но с частично изменяющимся набором неправильно распознанных элементов, а для одинаковых элементов – с различным перечнем альтернатив.

Как выяснилось в ходе проведения численных экспериментов, наиболее успешным оказался следующий способ обучения на используемых базах показателей крови. Полученные из них обучающие последовательности повторялись десять раз. Один цикл обучения проводился на такой десятикратной последовательности. Это позволило ускорить данную процедуру без потери информативности в отношении особенностей процесса обучения.

Из приведенных в таблице 1 результатов видно, что использование перекрестных произведений для мономов второго, третьего, четвертого и пятого порядка позволило значительно увеличить точность классификации. Однако эту процедуру удалось продолжить только до пятого порядка перекрестных произведений (длина полинома 165, 495, 1287). В случае, когда использовались мономы с перекрестными произведениями только второго порядка, мономы степенного вида имели порядок до семи включительно (длина полинома 69, 77, 85). При этом в обоих вариантах для успешного проведения расчетов требовалось увеличивать значение *m*. В таблице 1 в качестве примера приведено несколько случаев, когда недостаточно большое значение *m* приводило к расхождению в счете. Это же происходило и при увеличении порядка полиномов выше указанных максимальных значений.

Таблица 2

Распределение объема базы по классам основных систем организма

Пищеварительная система					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	33	17	26	33	109
ж	24	9	20	24	77
Органы дыхания					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»

м	32	11	12	21	76
ж	24	4	10	24	62
Опорно-двигательный аппарат					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	33	3	7	33	76
ж	33	0	6	33	72
Урологическая система					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	33	18	26	33	110
ж	42	0	42	35	119
Эндокринная система					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	34	26	27	34	121
ж	35	10	0	35	80
ЦНС, органы чувствительности					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	26	2	7	26	61
ж	31	13	4	31	79
Печень и желчевыводящие пути					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
м	31	12	0	31	74
ж	34	15	4	34	87
Гинекологическая система					

Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
ж	30	6	5	15	56
Грудные железы					
Распределение объема базы по классам					
м/ж	«1»	«2»	«3»	«4»	«1-2-3-4»
ж	50	0	29	50	129

Анализ обучающей базы

Зачастую в работах по методам классификации (распознавания) наличие базы обучения рассматривается как некая данность, не подлежащая обсуждению, хотя вполне очевидно, что нельзя говорить о методе, его достоинстве и недостатках, в отрыве от анализа множества, на котором проводилось обучение.

Используемый при решении описанной задачи подход, основанный на полиномиальной регрессии [2-5], хорошо зарекомендовавший себя при распознавании столь сложных объектов, как печатные и рукопечатные символы, являющийся точным, быстрым, устойчивым к шумам, генерирующий монотонные (надежные) оценки, имеющие вероятностную природу, был адаптирован для классификации объектов иного происхождения. На этих объектах он может проявить совершенно иные свойства.

В работах [4-5] описан подход к исследованию свойств обучающих множеств символьных объектов.

Проведем анализ множеств, содержащих наборы параметров крови, используемых при обучении метода.

Заметим, что исходное векторное пространство описывается векторами длины 8, имеющими следующий вид:

$$\mathbf{v}=(v_1, \ldots, v_N),$$

где $N = 8$.

Рассмотрим обучающее множество некоторой СО и заданного пола. Для каждого из четырех классов здоровья в отдельности получим среднестатистический вектор длины 8, относящийся к исходному векторному пространству. 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:

$$v_i^{k,cp} = (\sum_{j=1}^{J_k} v_i^{k,j}) / J_k \qquad (14)$$

В таблице 2 по каждой СО для обоих полов приведены значения J_k, где k=1, 2, 3, 4.

Расстояние между векторами $\mathbf{v}=(v_1,\ldots,v_N)$ и $\mathbf{u}=(u_1,\ldots,u_N)$ определяем в метрике L_2:

$$\|\mathbf{v}\text{-}\mathbf{u}\| = \sqrt{(\sum_{i=1}^{N}(v_i - u_i)(v_i - u_i))} \qquad (15)$$

Диапазон расстояний между центром масс k-го класса СЗЧ и векторами этого же класса («своими») по рассматриваемой базе находится на отрезке $[D\_k_{\min},\ D\_k_{\max}]$. Диапазон расстояний между центром масс k-го класса СЗЧ и векторами всех других классов («чужими») на отрезке $[D\_z_{\min},\ D\_z_{\max}]$. Пусть

$$\begin{aligned} Dk_{\min} &= \min(D\_k_{\min}, D\_z_{\min}) \\ Dk_{\max} &= \max(D\_k_{\max}, D\_z_{\max}) \end{aligned} \qquad (16)$$

Делим отрезок $[Dk_{\min},\ Dk_{\max}]$ (оси абсцисс на рис. 1, 2, 3, 4) 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 $[Dk_{\min}, Dk_{\max}]$.

Продемонстрируем описанную процедуру на примере пищеварительной системы для женщин. Обучающее множество в этом случае содержит 77 элементов (табл. 1, 2). Классы «1» и «4» содержат по 24 набора крови, а классы «2» и «3» включают соответственно 9 и 20 элементов.

На рисунках 1, 2, 3, 4 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 рисунке 1 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 количество 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 рисунках 2, 3, 4 вид распределений своих и чужих элементов для классов «2», «3», «4» совершенно иной. Свои элементы имеются до конца (рис. 4) или почти до конца (рис. 2, 3) 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. 1, 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 отмеченного ранее интересного факта, что элементы класса «1» (здоровые) распознаются при использовании полиномов, имеющих минимальную длину.

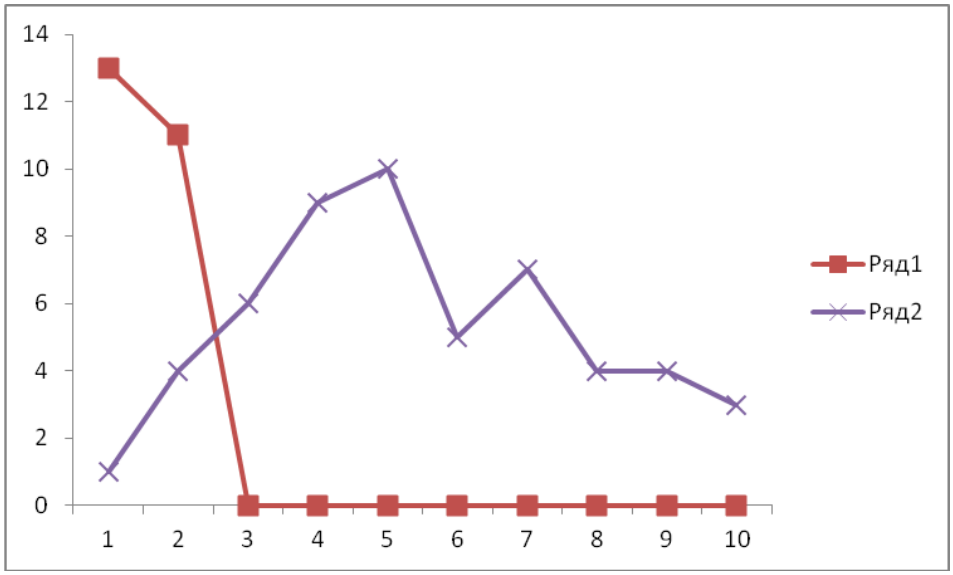

Рис. 1. Класс «1»

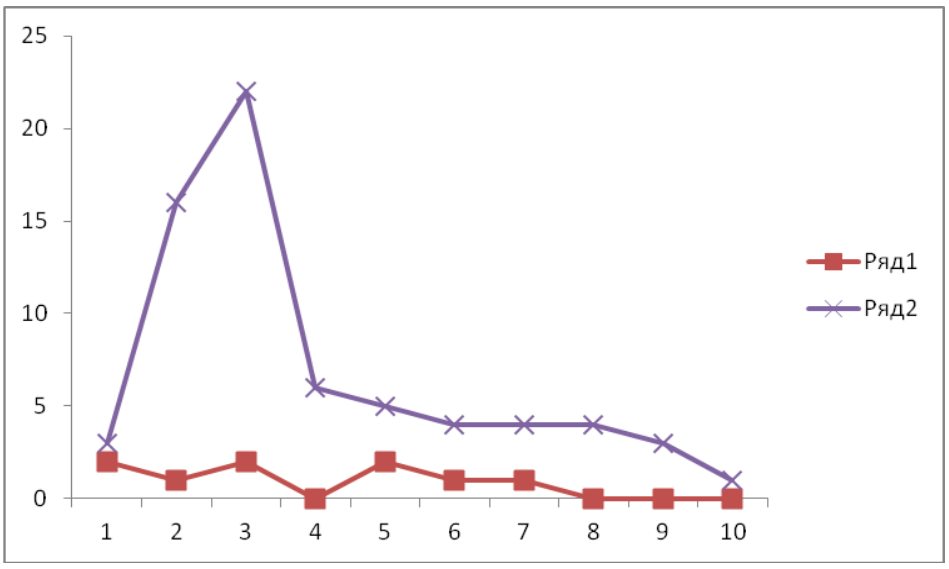

Рис. 2. Класс «2»

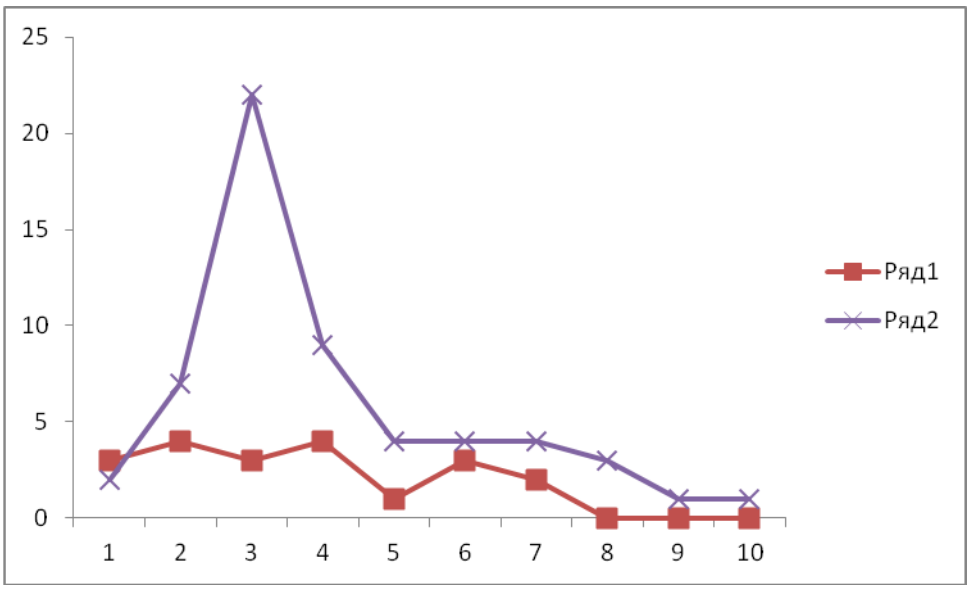

Рис. 3. Класс «3»

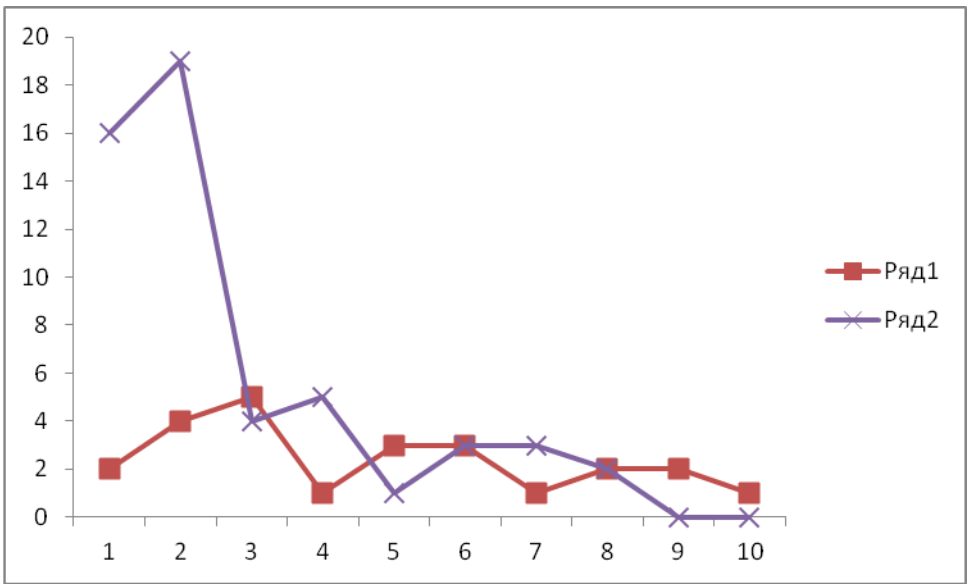

Рис. 4. Класс «4»

Заключение

Реализовано эффективное приложение статистического метода распознавания на основе полиномиальной регрессии, имеющего вероятностные оценки, в качестве классификатора состояния здоровья человека по показателям периферической крови (из пальца) для различных систем организма.

Рассмотрены четыре градации состояния здоровья – от практически здорового до максимальной степени поражения организма. Исследования по мужчинам и женщинам проведены раздельно.

Рассматриваемая проблематика является актуальной ввиду необходимости разработки статистических методов предварительной диагностики в медицине.

Данный метод по самой своей природе подходит для решения задач медицинской диагностики, поскольку он именно посредством статистической обработки данных, относящихся к большому количеству людей, позволяет найти некоторый принцип, по которому принимается то или иное решение.

Сравнительно небольшой объем и возможная неполнота обучающих множеств, несомненно, являются существенным недостатком и могут привести к уменьшению уровня достоверности полученных результатов и, как следствие, к понижению точности классификации в реальной медицинской практике. Однако это обстоятельство нисколько не умаляет значимость исследований, поскольку речь идет о создании метода, в принципе пригодного для классификации таких объектов, как наборы показателей крови. Кроме того, в целях «чистоты эксперимента» рассматривается распознавание элементов самих обучающих множеств.

Используемые в работе базы показателей крови обладает значительной вариабельностью и небольшим объемом. Были разработаны и реализованы новые подходы в организации обучающей последовательности.

Проведено исследование структуры обучающего множества. Для каждого из четырех классов СЗД получено распределение количества своих элементов (принадлежащих этому классу) и чужих элементов (относящихся к другим классам). Найдены закономерности распознавания, обусловленные структурой обучающих множеств.

Полиномы имеют различную структуру. Во-первых, содержащие мономы степенного вида до седьмого порядка, при этом перекрестные произведения используются в качестве мономов второго порядка. Также рассмотрены полиномы с мономами, использующими перекрестные произведения до пятого порядка.

Точность классификации на обучающих множествах для различных систем организма находится в диапазоне 95 – 100 %.

Библиографический список

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., Гавриков М.Б., Пестрякова Н.В. Статистический анализ характеристик метода распознавания при распознавании заданной модификации обучающего множества. // Труды ИСА РАН. 2015. Т.65. Вып.1. С. 82-88.