Об одном статистическом методе оценивания состояния здоровья человека*

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

Аннотация. Описывается способ получения оценки состояния здоровья обследуемого пациента по результатам лабораторного анализа периферической крови из пальца. Для этого используется классификатор, базирующийся на методе полиномиальной регрессии.

Ключевые слова: *классификация, полиномиальная регрессия, периферическая кровь.*

Введение

В настоящей работе описывается новое приложение метода классификации, основанного на полиномиальной регрессии. А именно, предлагается способ определения оценки состояния здоровья человека (СЗЧ) по параметрам периферической крови, полученным в результате лабораторного анализа.

Для мужчин и женщин строятся и используются различные классификаторы, поскольку диапазоны вариации показателей крови среди множества людей существенно зависят от пола. Кроме того, гинекологическая система имеется только у женщин.

По каждой системе организма (СО) – пищеварения, дыхания и пр., – проводится самостоятельное исследование СЗЧ посредством своего классификатора. При его построении используется обучающая выборка для рассматриваемой СО (табл.1).

СЗЧ включает четыре градации – от практически здорового состояния до максимальной степени поражения организма. Условное деление в проценном выражении следующее:

1 класс – здоровые – 0–20%;
2 класс – начальные отклонения состояния здоровья – 21–40%;
3 класс – выраженное отклонение состояния здоровья – 41–70 %;
4 класс – тяжелое заболевание – 71–100%.

При обучении используются выборки, полученные в результате детального обследования пациентов большой группой специалистов из различных областей медицины. Для определенной СО из рассматриваемого перечня к каждой из четырех возможных градаций относится список заболеваний, соответствующих этой СО (табл.1). База показателей крови практически здоровых людей одинакова для всех СО.

Идея использовать при решении описанной задачи подход, основанный на полиномиальной регрессии [2-4], основывалась на том обстоятельстве, что данный метод хорошо зарекомендовал себя при распознавании столь сложных объектов, как печатные и рукопечатные символы. Он является точным, быстрым, генерирует монотонные (надежные) оценки, имеющие вероятностную природу. Представилась уникальная возможность адаптировать этот подход для классификации объектов принципиально иной этимологии.

Несходство в отношении пространства первичных признаков основывалось на существенном отличии полиномиальных векторов как по структуре, так и по размерности. При распознавании печатных или рукопечатных символов изображения представляются в виде серого растра размера 16x16, состоящего из пикселов, состояние которых соответствуют их яркости, лежащей в диапазоне от 0 до 1. Именно они определяют ряд независимых компонент полиномиального вектора, а остальные вычисляются в виде некоторым образом заданных комбинаций этих элементов, имеющих одну и ту же природу и диапазон изменения. Напротив, параметры крови измеряются несопоставимыми величинами, а поэтому принципиально различаются и по наименованию, и по порядкам величин, и по диапазону вариации. Кроме того, количество параметров крови, равное 8 в данном исследовании, значительно меньше, чем 256 – число независимых компонент полиномиального вектора, построенного по растру изображения символа.

Все обучающие выборки дифференцируются по полу. В программе используется измеренные на автоматизированном анализаторе крови стабильные показатели. Они перечислены ниже, приведены их общепринятые обозначения и размерность:

RBC [L^{-1}] – эритроциты,
HGB [g L^{-1}] – гемоглобин,
PLT [L^{-1}] – тромбоциты,
WBC [L^{-1}] – лейкоциты,

* Работа выполнена при финансовой поддержке РФФИ (гранты №13-07-00262 а, № 13-07-12176 офи_м, №16-07-00742 а).

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

Таблица 1.

Наименование основных систем организмаи классов заболеваний

Пищеварительная система
Полость рта, глотка, пищевод, желудок, тонкая кишка, толстая кишка, поджелудочная железа.
1 класс – здоровые;
2 класс – гастрит, геморрой, аппендицит, грыжа пищевого отверстия, диафрагмы, эзофагит, хронический колит, гастроитоз, диспенсия, дуоденит, язвенный колит, эрозия желудка и двенадцатиперстной кишки;
3 класс – язва желудка, язва двенадцатиперстной кишки, панкреатит, энтерит, неспецифический язвенный колит, желудочно-кишечное кровотечение, туберкулез;
4 класс – онкологические заболевания.
Органы дыхания
Гортань, легкие, трохея, бронхи, диафрагма.
1 класс – здоровые;
2 класс – ларингит, трахеит, бронхит, гайморит, плеврит;
3 класс – бронхиальная астма, туберкулез, инфаркт легкого, пневмокониоз, пневмония, спонтанный пневмоторакс;
4 класс – онкологические заболевания.
Опорно-двигательный аппарат
Кости, связки, сухожилия, суставы, мышцы, фасции.
1 класс – здоровые;
2 класс – артрозы, бурситы, тендовагиниты, вывихи, растяжения, радикулит, остеохондроз, артрит, миозиты;
3 класс – подагра, ревматизм, миопатия, ревматоидный артрит, туберкулез;
4 класс – онкологические заболевания.
Урологическая система
Мочевыделительная система, половые органы.
1 класс – здоровые;
2 класс – цистит, простатит, водянка яичка, аденома предстательной железы, аднексит, бартолинит;
3 класс – пиелонефрит, макрогематурия, почечная колика, мочекаменная болезнь, амлоидоз, эндометрит, хроническая почечная недостаточность, киста почек, туберкулез;
4 класс – онкологические заболевания.
Гинекологическая система
1 класс – здоровые;
2 класс – эрозия шейки матки, гонорея;
3 класс – маточное кровотечение миомы, крауроз вульвы, туберкулез;
4 класс – онкологические заболевания.

Эндокринная система
Щитовидная, поджелудочная железы, надпочечники, гипофиз.
1 класс – здоровые;
2 класс – зоб, гипертиреоз, сахарный диабет 2-го типа, несахарный диабет, ожирение, аллергия;
3 класс – акромегалия, тиреотоксикоз, сахарный диабет 1-го типа, гипсерпаратериоз, микседема, болезнь Иценко-Кушинга, надпочечная недостаточность, туберкулез;
4 класс – онкологические заболевания.
ЦНС, органы чувствительности
Головной мозг, спинной мозг, зрение, обоняние, вкусовые железы, осязание, периферические нервы.
1 класс – здоровые;
2 класс – невралгии, невриты, дистония, воспаление среднего уха, тугоухость, хронический отит;
3 класс – энцефалиты, менингиты, инсульт, миопатии наследственные, туберкулез;
4 класс – онкологические заболевания.
Грудные железы (мужские и женские)
1 класс – здоровые;
2 класс – фиброаденомы, мастопатии, мастодении;
3 класс – острый мастит, туберкулез;
4 класс – онкологические заболевания.
Печень и желчевыводящие пути
1 класс – здоровые;
2 класс – дискенезия желчевыводящих путей, жировой гепатоз, синдром Жилебова, ротара, Дубина–Джонсона;
3 класс – гепатиты, холецистит, желчекаменная болезнь, абсцесс печени, холангиты, механическая желтуха, туберкулез;
4 класс – онкологические заболевания.
Общее состояние всего организма

В отношении обучающего множества следует заметить, что для печатных и рукопечатных символов имелось количественное расхождение. А именно, выборка рукопечатных цифр в сотни раз превосходила по объему множество печатных цифр. Это было вполне обоснованным, поскольку рукопечатное написание более вариативное, чем печатное. В то же время в каждой из этих выборок доли изображений различных символов были достаточно соразмерными и отличались не очень существенно. Однако ввиду объективных трудностей составления обучающих выборок анализов крови принцип соразмерности не выполнялся в необходимой степени. Для ряда СО некоторые из градаций СЗЧ были недостаточно заполнены.

Предъявляемое к выборкам условие случайности, несомненно, выполняется для анализов крови, как и для множества рукопечатных изображений. Это обусловлено большим разнообразием как человеческих организмов, так и почерков различных людей. В отношении печатных символов соблюдение этого условия должно проверяться более тщательно, поскольку количество типов печатных шрифтов не так уж велико.

Еще одна особенность заключается в том, что имеется очевидный изначальный порядок расположения градаций СЗЧ от здорового до максимальной степени поражения организма. Однако при решении поставленной задачи классификации в полученном перечне альтернатив он может нарушаться.

1. Оценивание СЗЧ по анализу крови.

Поскольку организм человека является сложной биологической системой, то разработка методологии оценивания его состояния – большая проблема, решением которой занимались многие поколения врачей. В медицине за долгие века ее существования и развития, несомненно, был накоплен и систематизирован огромный фактический

материал по диагностике заболеваний. Однако зачастую при общении с пациентом врач полагается в основном на самого себя, и гарантом правильной постановки диагноза являются его квалификация и опыт. Высокий уровень технического оснащения современной медицины имеет существенное значение, но именно врач должен принять безошибочное решение относительно направления дальнейших исследований состояния различных органов и систем, а затем уж стратегии и тактики лечения пациента.

Одним из «подручных» средств, используемых при первичной диагностике, является анализ периферической крови, состоящий из ряда показателей. Их набор определяется типом автоматического анализатора и включает 15 – 20, а иногда и большее число наименований. Нужно отметить, что только использование достаточного количества показателей (более пяти) позволяет судить о состоянии организма.

Известные гематологи, такие как Кассирский И.А., Воробьев А.И., Бергану Ш., Vinatier I., Hayachy Y. и другие, отмечали, что любое заболевание организма и его систем проявляется в виде изменения показателей крови.

Показатели крови здорового человеческого организма варьируются в некоторых известных диапазонах (табл.2). Значительные отклонения от нормы могут быть характерными проявлениями определенных заболеваний, выявить которые не составляет большого труда. Но во многих случаях ситуация неоднозначная, и врачу необходима помощь в принятии решения о состоянии той или иной системы организма.

Академик РАН Роберт Нигматулин в интервью изданию «Аргументы недели», опубликованном 23.04.2015 под названием «Власть «послушных», сказал: «В наше время эффективных информационных систем нельзя управлять «на глазок». Перефразируем его слова: «В наше время эффективных информационных систем нельзя диагностировать «на глазок».

Сама проблема оценивания СЗЧ непосредственно относится к понятию гомеостаза (в переводе с греческого homoios – подобный, тот же самый, stasis – состояние, подвижность), который характеризует относительное динамическое постоянство внутренней среды (крови, лимфы, тканевой жидкости) и устойчивости основных физиологических функций (кровообращения, дыхания,

Таблица 2.

Средние показатели периферической крови здоровых мужчин и женщин России

% колебаний		Пол		Основные показатели крови
М	Ж	М	Ж	
6,5	12	4,3- 4,6	3,9-4,2	Эритроциты RBC 10^{12}1/n
1,5	7	148-149	130-138	Гемоглобин HGB г/л
7,5	11	9±2	8,0±1,5	Ретикулоциты RET до 20 % от RBC
1,0	1,5	198±16	214±28	Тромбоциты PLT 10^9 1/л
39		4,46-7,28		Лейкоциты WBC 10^9 1/л
21		48,6-63,4		Нейтрофилы NEUT 10^9 1/л
--		2,39-4,39		------//------- NEUT%%
44		1,34-2,38		Лимфоциты LIMPH 10^9 1/л
--		2,47-3,89		------//------- LIMPH %%
--		0,32-1,26		Базофилы BASO %%
88		0,035-0,3		Эозинофилы EOS 10^9 1/л
--		0-5,9		------//------- EOS %%
107		0,02-0,14		Палочкоядерные P 10^9 1/л
--		0,24-2,36		------//------- P %%
46		2,31-4,31		Сегметоядерные S10^9 1/л
--		48,6-63,4		------//------- S %%
42		0,3-0,52		Моноциты MONO 10^9 1/л
--		5,72-8,62		------//------- MONO %%
5,0		1,0-2,0		Гранулоциты GRAN 10^9 1/л

терморегулирования, обмена веществ и пр.) организма. Как отмечается в монографии [1], при оценке гомеостаза и его динамики (гомеостатической активности) в клинической практике применяется комплексный подход. А именно, производится сопоставление ряда измеряемых параметров (температуры тела, пульсации сердца, отдельных показателей периферической крови и др.). Этот способ оценивания крайне неточен, поскольку ряд измеряемых величин нестабильны, они могут изменяться под влиянием внешних факторов (режима питания, физических нагрузок и т.д.). В качестве более объективного способа определения уровня гомеостаза и гомеостатической активности предлагается использовать не менее пяти показателей периферической крови.

2. Метод оценивания СЗЧ и полученные результаты.

Пусть проводится исследования СЗЧ по конкретной СО для пациента, пол которого известен. По предъявляемому анализу периферической крови требуется определить, какому элементу из некоторого конечного множества с $K = 4$ элементами, соответствующими градациям СЗЧ, он соответствует. Для рассматриваемых наборов показателей крови вводится вектор $\mathbf{v}\in\mathbf{R}^N$, i-ая компонента которого соответствует отнормированной величине i-го показателя крови, лежащей на отрезке [0,1], причем $N = 8$.

Нормировка проводится следующим образом. По обучающей выборке для данной СО, включающей все градации СЗЧ, для каждого i-го показателя крови находим минимальное и максимальное значение $v_i^{\min}$, $v_i^{\max}$, причем $i = 1,..., N$.

$$v_i^{\min} = \min (v_i^j),\ j = 1,\ Q_i$$

$$v_i^{\max} = \max (v_i^j),\ j = 1,\ Q_i,$$

где Q_i – объем выборки по i-му показателю крови.

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 $1\le k\le K$. На выходе имеем элемент СЗЧ с порядковым номером r, где

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

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

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

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 выбранных и приведенных в (2) базисных мономов, упорядоченных определенным образом, то в векторном виде соотношения (2) можно записать так:

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

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 $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$ ($\mathbf{v}^{(j)}$ набор параметров крови, соответствующий элементу СЗЧ с каким-либо номером k ($1\le k\le K$) и его базисный вектор $\mathbf{y}^{(j)}= (0\ldots1\ldots0)$, где 1 стоит на k-м месте, $1\le j\le J$):

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

При получении правой части (4) используется следующая рекуррентная процедура, где A_0 и G_0 заданы:

$$A_j = A_{j-1} - \alpha_j G_j \mathbf{x}^{(j)}[A_{j-1}^{\mathrm{T}}\mathbf{x}^{(j)} - \mathbf{y}^{(j)}]^{\mathrm{T}},\quad \alpha_j = 1/J \quad (5)$$

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

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

Здесь $x_1, x_2,\ldots,x_L$ - компоненты вектора $\mathbf{x}(\mathbf{v})$. Получаемые оценки могут выходить за рамки отрезка [0,1] из-за того, что используемый метод является приближенным. Отрицательные значения искусственно обнулялись, а те, которые были больше 1, делались равными 1.

Для показателей крови использовалась следующая модификация вектора $\mathbf{x}(\mathbf{v})$:

$$\mathbf{x}=(1, \{v_i\}, \{v_i^2\}, \{v_i^3\}, \{v_i^4\}\}, \{v_i v_j\}) \quad 1\le i\le 8, 1\le j\le 8, i\ne j \quad (6)$$

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

Поскольку обучающие множества имели неравноценные по объему подмножества для различных градаций заболевания организма, то использовался следующий прием, суть которого сводилась к приблизительному выравниванию числа элементов для четырех классов по каждой СО. А именно, для тех классов, в которых элементов было недостаточно, добавлялись повторно имеющиеся элементы в требуемом количестве, а затем проводилось перемешивание по всему объему обучающего множества.

Для количественной оценки качества классификации требуется ввести следующее понятие.

Точностью распознавания по базе *B* называется величина

$$1-\frac{\sum_{b\in B}(1-\rho(C(b)\ P(b)))}{|B|} \quad (7)$$

где *b* – элементы тестовой базы анализов периферической крови *B*, $|B|$ – число наборов показателей крови в базе *B*, *C(b)* – класс СЗЧ, известный для каждого набора из тестовой базы, *P(b)* – класс СЗЧ, полученный в результате распознавания, $\rho(s,t)$ – расстояние между известным и распознанным классами СЗЧ (функция сравнения, равная 1, если *s* и *t* неразличимы, и равная 0 в противоположном случае).

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 начинает нарастать.

Тестирование классификатора проводилось на той же базе, которая использовалась для обучения. Достигнутая точность распознавания для различных рассматриваемых СО находится в диапазоне 88 – 93 %. Эти результаты соответствуют данным, полученным при помощи алгебраического подхода Журавлева [1].

Заключение

Для разработанного авторами статистического метода распознавания на основе полиномиальной регрессии реализовано приложение в качестве классификатора СЗЧ по показателям периферической крови из пальца для различных СО. Подтверждена возможность применения данного метода для объектов, природа которых принципиально отлична от изображений печатных и рукопечатных символов, а структура обучающего множества значительно более сложная, чем в случае изображений символов.

Литература

1. *Ставицкий Р.В., Лебедев Л.А., Лебедев А.Л., Смыслов А.Ю.* Количественная оценка гомеостатической активности здоровых и больных людей. М., «ГАРТ»,2013г., 131 с.
2. *Гавриков М.Б., Пестрякова Н. В.* Метод полиномиальной регрессии в задачах распознавания печатных и рукопечатных символов. //Препринт ИПМатем. РАН, М., 2004, №22, 12 стр.
3. *Гавриков М.Б., Мисюрев А.В., Пестрякова Н.В., Славин О.А.* Об одном методе распознавания символов, основанном на полиномиальной регрессии. // Автоматика и Телемеханика. 2006, №2, с. 119-134.
4. *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.

Гавриков Борис Михайлович. Медицинский физик МГОБ №62. Окончил в 2015 г. НИИЯУ «МИФИ». Кол-во печатных работ: 7 (1 монография). Область научных интересов: математическое моделирование, распознавание образов, медицинская физика. E-mail: bmgavrikov@gmail.com

Лебеденко Ирина Матвеевна. В.н.с. ФГБУ «РОНЦ им. Н. Н. Блохина» МЗ РФ. Д.б.н., профессор. Окончила МЭИ в 1977 г. Количество печатных работ 170 (в т.ч. 10 монографий). Область научных интересов: медицинская физика, лучевая терапия. E-mail: imlebedenko@mail.ru

Пестрякова Надежда Владимировна. В.н.с. ИСА ФИЦ ИУ РАН. Д.т.н. Окончила в 1983 г. МФТИ. Кол-во печатных работ: более 60 (1 монография). Область научных интересов: математическое моделирование, вычислительная гидродинамика, распознавание образов. E-mail: nadya_p@cs.isa.ru

Ставицкий Роман Владимирович. Г.н.с. РНЦ Рентгенорадиологии. Д.б.н., профессор. Окончил в 1953 г. Ленинградский электротехнический институт. Кол-во печатных работ: более 700 (44 монографии). Область научных интересов: радиационная физика. E-mail: nadya_p@cs.isa.ru