О способности статистического классификатора к обобщениям*

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

[I] Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия
[II] Федеральное государственное учреждение "Федеральный исследовательский центр "Институт прикладной математики им. М.В. Келдыша" Российской академии наук", г. Москва, Россия

Аннотация. В статье проводится изучение способности к обобщениям статистического классификатора, предназначенного для оценивания состояния здоровья человека по параметрам периферической крови. Описана и реализована математическая модель, предназначенная для численного исследования интерполяционных и экстраполяционных свойств разработанного авторами классификатора, основанного на полиномиально-регрессионном подходе и имеющего вероятностные оценки.

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

DOI 10.14357/20718632210404

Введение

Проблема компьютерной медицинской диагностики приобрела особую актуальность в последнее время из-за пандемии, вызванной коронавирусной инфекцией. Число заболевших столь велико, что стала очевидной необходимость создания программных комплексов, предназначенных для поддержки врачей, перегруженных работой при решении вопросов, связанных с прогнозом тяжести течения болезни, необходимостью госпитализации и прочее.

Подобные разработки начали появляться как внутри, так и вне России. Все они основаны на машинном обучении, а в первичные признаки включены показатели крови. Надеемся, что наша деятельность по разработке статистического классификатора для оценивания состояния здоровья человека (СЗЧ) по параметрам периферической крови пробила брешь недоверия и убедила сомневающихся в правомочности такого подхода.

Он основан 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 СЗЧ гематологи предлагают использовать не менее пяти показателей периферической крови [1-3].

Роман Владимирович Ставицкий, ушедший из жизни 30 сентября 2021 года, был инициатором создания баз данных лабораторного анализа крови, дифференцированных по тяжести заболевания. Он же стал нашим идейным вдох-

* Работа выполнена при финансовой поддержке РФФИ (гранты №18-29-26008, 18-29-26009).

новителем при разработке статистического классификатора и обучении на базах для различных систем организма, собранных благодаря его активному руководству.

Заранее не было очевидно, можно ли в принципе решить задачу классификации для такого обучающего множества. Интуитивно понятна принципиальная осуществимость распознавания визуальных объектов, для каждого из которых имеется свой набор изображений. В случае диагностики фрагменты базы по каждому классу заполнялись экспертами-медиками, исходя из их представлений о степени поражения организма [1]. Известны диапазоны, характерные для нормальных показателей крови. Однако решение, что человек практически здоров, принимали врачи, даже если параметры крови выходили за рамки этих интервалов.

Мы не пытались формализовать методику принятия решения медиками при вынесении «вердикта», не вникали, чем они его обосновывали, какие процессы протекали в их головах в это время. Также, не составляли перечень вопросов, которые они задавали пациентам, не вычисляли их частоту. Жизнь богаче «цифры», и у нас нет уверенности в идеальности медицины, которая базируется на идее именно такой повальной унификации и алгоритмизации. Результаты заполнения опросника определяют протокол лечения. Если пациент не подпал под стандартные лекала, болел чем-то другим и не выздоровел – его проблемы.

Мы построили классификатор и обучили его на базе, сформированной «по жизни» вышеуказанным способом [4-7]. Совокупность элементов этого обучающего множества, соответствующая каждому классу, может рассматриваться как некая единая случайная величина. Возникает ряд проблем: как ее описать, каковы ее свойства. Здесь изложен подход и результаты проведенного математического моделирования при изучении динамики характеристик классификации (точности, оценок) в процессе искажения элементов из фрагментов базы, соответствующих отдельным классам. В данном исследовании, в дополнение к [8], искажение рассматривается как совокупность экстраполяции и интерполяции.

1. Метод классификации

Общепринятые обозначения и размерность используемых восьми показателей крови: RBC [L^{-1}] – эритроциты, HGB [gL^{-1}] – гемоглобин, PLT [L^{-1}] – тромбоциты, WBC [L^{-1}] – лейкоциты, LIMPH [L^{-1}], [%] – лимфоциты, GRAN [L^{-1}], [%] – гранулоциты (GRAN=NEUT+EOS+BASO, где NEUT[L^{-1}],[%] – нейтрофилы, EOS[L^{-1}],[%] – эозинофилы, BASO[L^{-1}],[%] – базофилы).

Рассматриваем определенную систему организма (СО) человека. Вводим вектор $v \in R^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови, где N=8. Отождествляем k-й элемент множества градаций СЗЧ с базисным вектором из R^K: e_k=(0…1…0), причем 1 находится на k-м месте, $1 \le k \le K$, K=4. Обозначим $Y=\{e_1,\ldots,e_K\}$.

Пусть существует $p_k(v)$ – 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, получивший максимальное значение вероятности:

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

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

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

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

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

Тогда (2) можно записать в векторном виде:

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

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

Приближенное вычисление A производится при обучении на конечной последовательности: $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$. 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, $1 \le j \le J$):

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

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

В данной работе рассмотрена пищеварительная система для мужчин. В этом случае использовались модификации вектора $\mathbf{x}(\mathbf{v})$ следующего вида:

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\}),\quad 1 \le i \le 8,\ i \le j \le 8,\ j \le k \le 8,\ k \le l \le 8. \quad (5)$$

В (5) выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычисляемым по всем показателям крови из имеющегося набора. Длина полинома 495. Имеются мономы первого, второго, третьего и четвертого порядка. Перекрестные произведения используются для мономов второго, третьего и четвертого порядка.

Обучающее множество рассматриваемой СО имеет 109 элементов: классы здоровья K^1 и K^4 содержат по 33 набора крови, а классы K^2 и K^3, включают соответственно 17 и 26 элементов.

Обученный классификатор обеспечил 99,2% правильной классификации на обучающем множестве; имеется по одной ошибке на элементах K^2 и K^3.

2. Численное моделирование искажений

Ниже при нумерации классов K^1, K^2, K^3, K^4 используется символ $c, 1 \le c \le 4$, а число элементов в них обозначено $|K^1|$, $|K^2|$, $|K^3|$, $|K^4|$. Признаки Π_1, …, Π_8 перенумерованы посредством символа $b, 1 \le b \le 8$.

В Табл. 1 для K^1, K^2, K^3, K^4 приведены диапазоны, которым принадлежат используемые первичные признаки $\Pi_1,\ldots,\Pi_8$ ($D_b^c \equiv [l_b^c, r_b^c], 1 \le c \le 4, 1 \le b \le 8$). Также для каждого признака указан соответствующий отрезок по совокупности четырех классов ($D_b \equiv \bigcup_{1 \le c \le 4} D_b^c \equiv [l_b, r_b]$). Особенность K^1 заключается в том, что пятый, седьмой и восьмой признаки являются константами (вырожденные случаи).

Рассмотрим один из четырех классов (c_0). Будем исследовать объекты, полученные из элементов этого класса, входящих в обучающее множество, при модификации признака с номером b_0. На t-м шаге

$$v_{b0} \to v_{b0} \pm h_{b0} \cdot t, h_{b0} = (r_{b0} - l_{b0})/10, t = 1,\ldots,9. \quad (6)$$

Опишем, что происходит с элементами обучающего множества в процессе усиления искажения. Изначально имеется только по одной ошибке во втором и третьем классах. С нарастанием t часть элементов выходят за правую или левую границы отрезка D_{b0}^{c0}, в зависимости

Табл. 1. Интервалы признаков по классам здоровья системы организма

b	*Класс «1»*	*Класс «2»*	*Класс «3»*	*Класс «4»*	*«1-2-3-4»*
1	[437,548]	[369,574]	[330,573]	[304,586]	[304,586]
2	[170,336]	[102,217]	[61,517]	[134,504]	[61,517]
3	[439,900]	[390,1750]	[220,1380]	[467,2167]	[220,2167]
4	[1310,1630]	[1080,1770]	[860,1680]	[910,1690]	[860,1770]
5	186	[110,450]	[80,310]	[50,380]	[50,450]
6	[208,393]	[171,887]	[184,963]	[51,403]	[51,963]
7	2900	[2300,11700]	[1300,10500]	[2840,19640]	[1300,19640]
8	658	[492,763]	[437,766]	[480,906]	[437,906]

от увеличения признака (знак «+» в (6)) или его уменьшения (знак «-» в (6)).

Элементы называем внутренними (*in*), если при модификации рассматриваемый признак не покинул рамки соответствующего ему диапазона (Табл. 1); в противном случае объекты считаются внешними (*ex*).

Кроме того, при модификации может становиться противоположным статус элементов в отношении правильности классификации.

На Рис. 1, а – 5, а показано, как изменяется ряд параметров при увеличении (уменьшении) фиксированного признака $П_1,\ldots,П_8$. Здесь N^{in} (Ряд 8) – число внутренних элементов, среди которых распознано правильно – N_w^{in} (Ряд 6), неправильно – N_e^{in} (Ряд 3), причем $N^{in}=N_w^{in}+N_e^{in}$; эти величины характеризуют интерполяцию. Количество внешних элементов – N^{ex} (Ряд 1), из них распознано правильно – N_w^{ex} (Ряд 7), неправильно – N_e^{ex} (Ряд 4), причем $N^{ex}=N_w^{ex}+N_e^{ex}$; это характеристики экстраполяции. Кроме того, изображена динамика определяемого совместно экстраполяцией и интерполяцией количества распознаваний: ошибочных – $N_e=N_e^{in}+N_e^{ex}$ (Ряд 2), правильных – $N_w=N_w^{in}+N_w^{ex}$ (Ряд 5), причем $N_e+N_w=|K^c|$, $N^{in}+N^{ex}=|K^c|$.

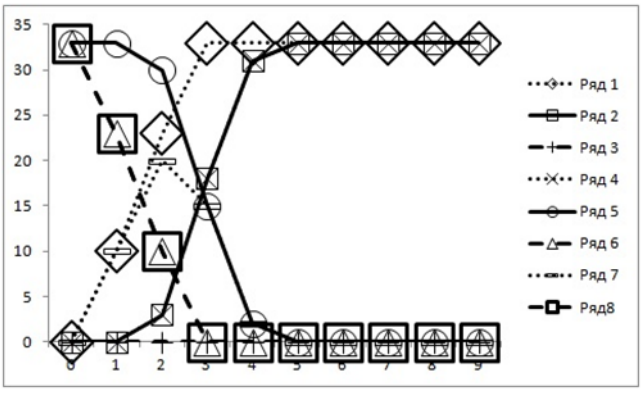
Интервал интерполяции (экстраполяции) определяется выполнением условия $N^{in}>0$ ($N^{ex}>0$).

Будем оценивать *интерполяционные (экстраполяционные) свойства* метода следующим образом. Если на каком-то интервале изменения *t* имеем: $N_e^{in}<N_w^{in}$ ($N_e^{ex}<N_w^{ex}$), то считаем это признаком *хорошей интерполяции (экстраполяции)*; напротив, при выполнении $N_e^{in}>N_w^{in}$ ($N_e^{ex}>N_w^{ex}$) *интерполяция (экстраполяция) плохая*. Равенство соответствующих величин означает *посредственное* качество. Для получения количественных характеристик имеет значение соотношение величин в левой и правой частях приведенных неравенств.

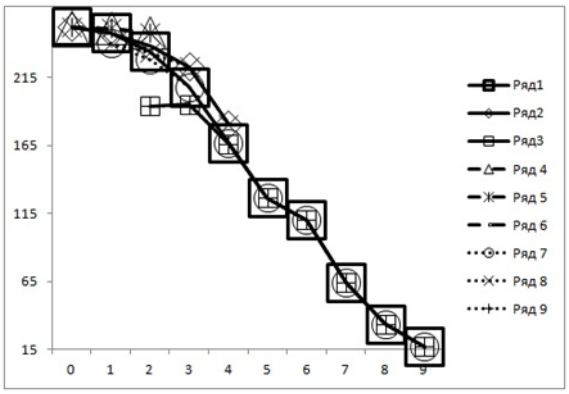
Устойчивость метода к искажению определяется наличием и протяженностью интервала, на котором $N_e<N^{ex}$. Количественные характеристики могут быть вычислены с использованием величин в левой и правой частях данного неравенства.

Рис. 1, б – 5, б демонстрируют динамику оценок при росте (падении) значения одного из признаков $П_1,\ldots,П_8$.

Для описанных типов элементов установим соответствие между обозначениями их количества и оценок. Если элементы отнесены к рассматриваемому классу (правильная классификация), то вычисляется математическое ожидание их оценок, причем N_w соответствует E_w (Ряд 2 на Рис. 1, б – 5, б), N_w^{in} – E_w^{in}



а) 1- N^{ex}, 2- N_e, 3- N_e^{in}, 4- N_e^{ex}, 5- N_w, 6- N_w^{in}, 7- N_w^{ex}, 8- N^{in}



б) 1- E, 2- E_w, 3- E_e, 4- E^{in}, 5- E_w^{in}, 6- E_e^{in}, 7- E^{ex}, 8- E_w^{ex}, 9- E_e^{ex}

Рис. 1. Класс 1. Признак 3 (увеличение)

(Ряд 5), N_w^{ex} – E_w^{ex} (Ряд 8). Если модифицированные объекты отошли к другим классам (неправильная классификация), то вычисляется математическое ожидание оценок альтернативы, соответствующей рассматриваемому классу, причем N_e соответствует E_e (Ряд 3), N_e^{in} – E_e^{in} (Ряд6), N_e^{ex} – E_e^{ex} (Ряд 9). Если рассматриваются элементы вне зависимости от правильности распознавания, то с учетом сказанного, всем $|K^c|$ элементам соответствует E (Ряд 1), N^{in} – E^{in} (Ряд 4), N^{ex} – E^{ex} (Ряд 7).

Выполнение условий E_e^{in}=255 или E_e^{ex}=255 соответствует фатальному нарушению оценок. Если на всем интервале изменения *t* выполняется $\min\limits_{0\le t\le 9} E_w \le \max\limits_{0\le t\le 9} E_e$, где в качестве E_w может фигурировать как E_w^{in}, так и E_w^{ex}, а E_e соответствует E_e^{in} или E_e^{ex}, то разделение оценок по диапазонам отсутствует.

3. Практически здоровые люди

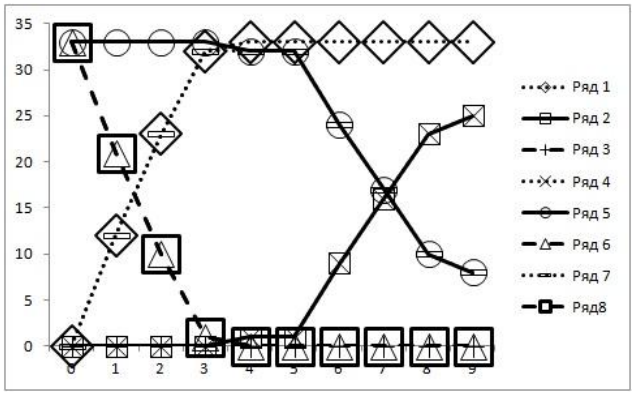
В первом классе (c_0=1) имеется 33 элемента. В неискаженном состоянии ошибок нет, $E^{in} = E_w^{in}$=252. Для рассмотренного класса есть два типа признаков: изменяющиеся в некотором диапазоне (Π_1, Π_2, Π_3, Π_4, Π_6) и принимающие константное значение (Π_5, Π_7, Π_8), соответствующие невырожденным и вырожденным случаям. Ограничимся исследованием невырожденных случаев. На конкретных примерах продемонстрируем динамику введенных нами величин.

На Рис. 1 изображены зависимости, полученные при увеличении признака Π_3. Интерполяция существует до второго шага. Число внутренних элементов N^{in} (Ряд 8) и тождественно равное ему количество безошибочных распознаваний N_w^{in} (Ряд 6) падают (Рис. 1, а) в отсутствие ошибок: $N^{in} - N_w^{in} = N_e^{in} \equiv 0$ (Ряд3). $E^{in} \equiv E_w^{in}$ (Ряд 4, Ряд 5); эти величины от изначально высокого значения 252 немного уменьшаются (Рис. 1, б). Оценки адекватные. Интерполяция отличная.

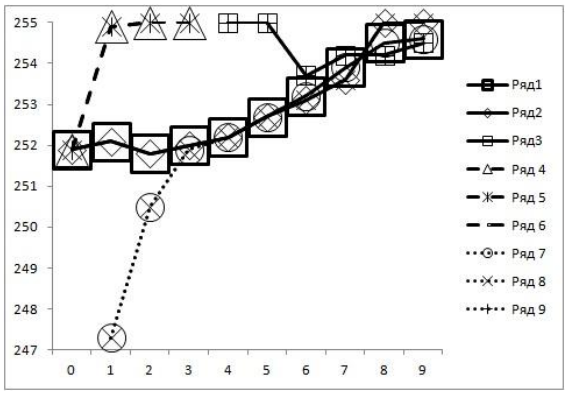
Для внешних элементов N^{ex} (Ряд 1) монотонно нарастает, как и $N_e^{ex} = N^{ex} - N_w^{ex}$ (Ряд 4) с учетом появления ошибок на втором шаге (Рис. 1, а); $N^{ex} \equiv 33$ с третьего шага, а с пятого – $N_e^{ex} \equiv 33$. Сначала N_w^{ex} (Ряд 7) увеличивается: на первом шаге в отсутствие ошибок $N^{ex} = N_w^{ex}$=10, на следующем – N_e^{ex}=3, N_w^{ex}=20; далее N_w^{ex} падает до 0 к пятому шагу. На первом шаге $E^{ex} = E_w^{ex}$=240 (Ряд 7, Ряд 8); затем эти величины понижаются (Рис. 1, б). К концу интервала искажения E_e^{ex} уменьшается от уровня $E_e^{ex} \approx 195$ (Ряд 9), имеющегося на втором и третьем шагах, на которых, как и на следующем – $E_e^{ex} < E_w^{ex}$. Однако поскольку первоначальный уровень E_e^{ex} высокий, диапазоны E_e^{ex} и E_w^{ex} не разделены. Экстраполяция до второго шага хорошая, а затем меняется на противоположную.

Устойчивость классификатора хорошая: $N_e < N^{ex}$ (Рис. 1, а) выполняется до четвертого шага, далее N_e устанавливается на максимуме.

Рис. 2 соответствует динамике величин при уменьшении Π_4. Интерполяция имеет место до третьего шага. При нулевом уровне N_e^{in} (Ряд 3) поведение N^{in} (Ряд 8) и N_w^{in} (Ряд 6) не отличается (Рис. 2, а) от наблюдавшегося на Рис. 1, а. $E^{in} \equiv E_w^{in}$ (Ряд 4, Ряд 5); эти величины от начального значения 252 (Рис. 2, б) на одном отрезке увеличиваются почти до максимума 255, а далее стабилизируются на нем. Оценки адекватные. Интерполяция отличная.



а) 1- N^{ex}, 2- N_e, 3- N_e^{in}, 4- N_e^{ex}, 5- N_w, 6- N_w^{in}, 7- N_w^{ex}, 8- N^{in}



б) 1- E , 2- E_w, 3- E_e, 4- E^{in}, 5- E_w^{in}, 6- E_e^{in}, 7- E^{ex}, 8- E_w^{ex}, 9- E_e^{ex}

Рис. 2. Класс 1. Признак 4 (уменьшение)

N^{ex} (Ряд 1) нарастает и устанавливается на максимуме 33 с четвертого шага (Рис. 2, а), на котором появляется ошибка, сохраняющаяся на одном отрезке; далее $N_e^{ex} = N^{ex} - N_w^{ex}$ (Ряд 4) повышается и достигает 25 в конце диапазона изменения Π_4. Сначала N_w^{ex} (Ряд 7) увеличивается в отсутствие ошибок: $N^{ex} = N_w^{ex}$=32 на третьем шаге; до пятого – N_w^{ex} не меняется, а затем падает до 8 к концу интервала искажения. С первого по третий шаг $E^{ex} \equiv E_w^{ex}$ (Ряд 7, Ряд 8); здесь и далее эти величины от значения 247 повышаются (Рис. 2, б), причем $E_w^{ex} \equiv$255 на последнем отрезке исследуемого диапазона. На четвертом и пятом шагах $E_e^{ex} \equiv$255 (Ряд 9) – оценки неадекватные. Экстраполяция до седьмого шага хорошая, а далее меняется на противоположную.

Устойчивость классификатора хорошая, поскольку везде $N_e < N^{ex}$ (Рис. 2, а).

Было показано [11], что во всех невырожденных случаях, как и в приведенных примерах, на этапе интерполяции ошибок нет. $E^{in} \equiv E_w^{in}$; эти оценки, вначале равные 252, далее могут повышаться или понижаться, но также имеют высокие значения. Как при увеличении, так и при уменьшении конкретного признака интерполяция отличная, диапазон ее существования один и тот же, он не превышает половину интервала искажения.

Для внешних элементов N^{ex} монотонно нарастает, как и N_e^{ex}, причем ошибки появляются позже первого шага. Сначала в отсутствие ошибок N_w^{ex} увеличивается столь же стремительно, как N^{ex}, а затем после достижения некоторого максимума N_w^{ex} падает, так что либо достигает нулевого значения до окончания интервала искажения, либо эта величина до последнего шага остается положительной. Везде экстраполяция вначале отличная, а затем становится плохой.

Исследование поведения оценок выявило варианты, когда дальнейшее искажение исключено. Во-первых, когда ошибки имеют неадекватную оценку $E_e^{ex} \equiv$255 (при увеличении Π_1, Π_2, Π_6, а также при уменьшении Π_4). Во-вторых, если на некоторых шагах $E_e^{ex} > E_w^{ex}$ (при уменьшении Π_1, Π_2).

Также возможны проблемы оценивания по следующим причинам. Во-первых, когда разделение диапазонов E_e^{ex} и E_w^{ex} отсутствует (при увеличении или уменьшении Π_3) или, во-вторых, оно является слабым (при увеличении Π_4, уменьшении Π_6).

Во всех случаях искажения на начальном этапе, а именно, до появления ошибок, проблем с оцениванием при экстраполяции нет.

Устойчивость классификатора хорошая либо на всем интервале искажения, где выполняется $N_e < N^{ex}$, либо до шага, на котором N_e устанавливается на максимуме.

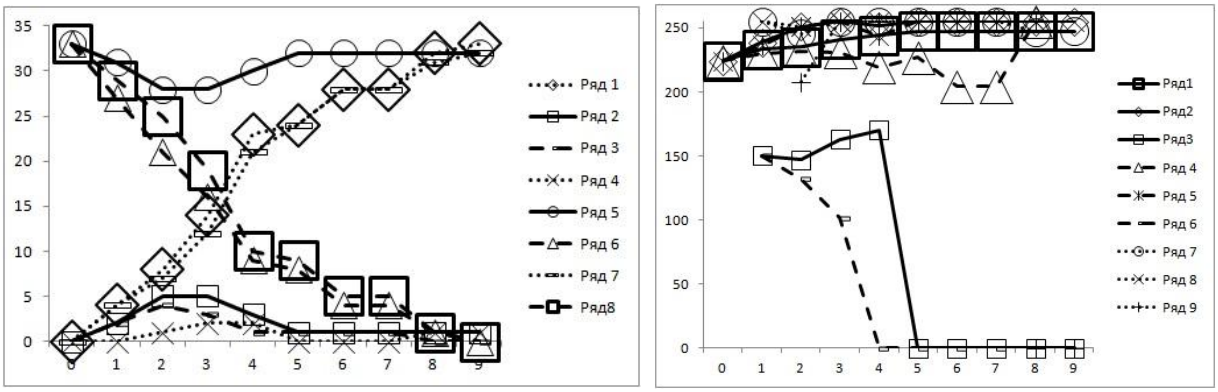
4. Онкологические больные

Приведем примеры исследования обучающего множества, соответствующего максимальному поражению рассматриваемой системы организма. Объем четвертого класса (c_0 =4) составляет 33 элемента. В отсутствие искажения ошибок нет, $E^{in} = E_w^{in}$ =224.

При увеличении Π_4 для интерполяции, имеющейся до восьмого шага, N^{in} (Ряд 8) и N_w^{in} (Ряд 6) уменьшаются: в конце $N^{in} = N_w^{in}$ =1 (Рис. 3, а). $N^{in} - N_w^{in} = N_e^{in}$ (Ряд 3) нарастает до 4 на втором шаге, а затем к восьмому – падает до нуля. Противоположно изменению отношения N_e^{in} / N_w^{in}, E^{in} (Ряд 4) колеблется, но в целом понижается от исходного значения 224 до минимума E^{in} ≡204 на отрезке между шестым и седьмым шагами (где также неизменны N_w^{in} и N_e^{in}); к последнему шагу E^{in} повышается до максимума 255 (что соответствует N_e^{in} =0) (Рис. 3, б). E_w^{in} (Ряд 5) увеличивается от 224 и с третьего шага устанавливается на 255, с незначительным отклонением. E_e^{in} (Ряд 6) между первым и четвертым шагами падает от 150 до нуля и стабилизируется на этом уровне. Оценивание обоснованное. Диапазоны E_w^{in} и E_e^{in} разделены. Интерполяция всюду хорошая.

N^{ex} (Ряд 1) и N_w^{ex} (Ряд 7) нарастают к концу интервала искажения до 33 и 32 (Рис. 3, а), причем $N^{ex} - N_w^{ex} = N_e^{ex} \le 2$ (Ряд 4). E^{ex} (Ряд 7) и E_w^{ex} (Ряд 8) с первого до последнего шага принимают значение 255 или близки к нему (Рис. 3, б). Одна ошибка с E_e^{ex} =208 появляется на втором шаге, а на двух следующих – N_e^{ex} ≡2, E_e^{ex} ≡255 (неадекватная оценка). Далее ошибки исчезают, но на последнем отрезке N_e^{ex} ≡1, E_e^{ex} ≡0. Экстраполяция хорошая везде.

$N_e < N^{ex}$ (Ряд 2, Ряд 1), устойчивость классификатора всюду хорошая (Рис. 3, а).



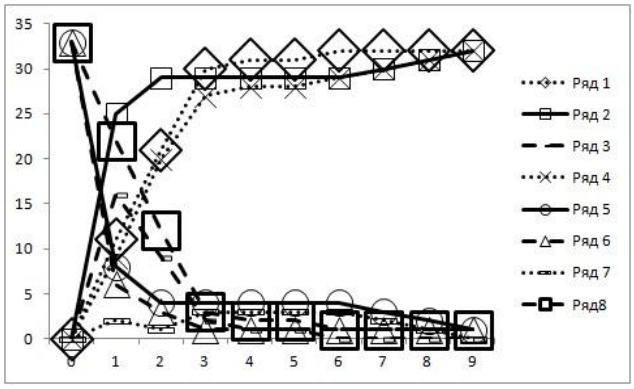
а) 1- N^{ex}, 2- N_e, 3- N_e^{in}, 4- N_e^{ex}, 5- N_w, 6- N_w^{in}, 7- N_w^{ex}, 8- N^{in}

б) 1- E, 2- E_w, 3- E_e, 4- E^{in}, 5- E_w^{in}, 6- E_e^{in}, 7- E^{ex}, 8- E_w^{ex}, 9- E_e^{ex}

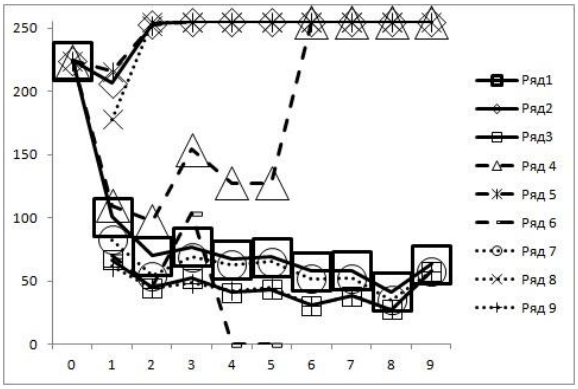
Рис. 3. Класс 4. Признак 4 (увеличение)

При уменьшении Π_7 интерполяция существует на всем интервале искажения. N^{in} (Ряд 8) и N_w^{in} (Ряд 6) уменьшаются сначала быстро: на третьем шаге N^{in}=3, N_w^{in}=1, а затем замедленно: в конце $N^{in} = N_w^{in}$=1 (Рис. 4, а). $N^{in} - N_w^{in} = N_e^{in}$ (Ряд 3) повышается до 16 на первом шаге, а к шестому – падает до нуля. Противоположно изменению отношения N_e^{in} / N_w^{in}, E^{in} (Ряд 4) сначала понижается от исходного значения 224 до 110 на первом шаге; на следующем, где N_e^{in} также существенно превышает N_w^{in}, достигается минимум E^{in}=98; далее следует участок немонотонного изменения E^{in}, где также перестраивается соотношение между N_w^{in} и N_e^{in}, находящихся на низком уровне; в итоге с шестого шага E^{in}≡255, что соответствует N_e^{in}≡0 (Рис. 4, б). E_w^{in} (Ряд 5) сначала немного уменьшается от 224, затем увеличивается, и с третьего шага E_w^{in}≡255. E_e^{in} (Ряд 6) между первым и третьим шагами в целом повышается от 70 до 104, к четвертому – падает до нуля и стабилизируется на этом уровне. Оценивание обоснованное. Диапазоны E_w^{in} и E_e^{in} разделены. Интерполяция становится плохой уже с первого шага и до третьего – остается таковой, на следующих двух шагах – посредственная, а затем хорошая.

N^{ex} (Ряд 1) и N_e^{ex} (Ряд 4) нарастают сначала стремительно (Рис. 4, а): на третьем шаге N^{ex}=30, N_e^{ex}=27; далее эти величины повышаются медленно: с шестого шага N^{ex}≡32, а N_e^{ex} достигает этого значения лишь в конце интервала искажения. $N^{ex} - N_e^{ex} = N_w^{ex}$ (Ряд 7) сначала немонотонно повышается: N_w^{ex}≡3 с третьего по шестой шаг, а затем падает до нуля к концу. E^{ex} (Ряд 7) совершает небольшие колебания в диапазоне от 83 до 35, но в целом падает (Рис. 4, б), что согласуется с динамикой N_w^{ex} / N_e^{ex}. E_e^{ex} (Ряд 9) также колеблется в диапазоне от 62 до 28, при этом немного понижается к концу. E_w^{ex}=178 (Ряд 8) на первом шаге, со второго – E_w^{ex}≡255. Оценивание обоснованное. Диапазоны E_e^{ex} и E_w^{ex} разделены. Экстраполяция плохая везде.

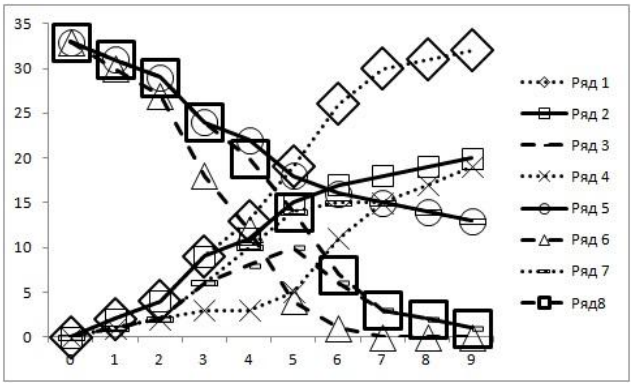


а) 1- N^{ex},2- N_e,3- N_e^{in},4- N_e^{ex},
5- N_w,6- N_w^{in},7- N_w^{ex},8- N^{in}

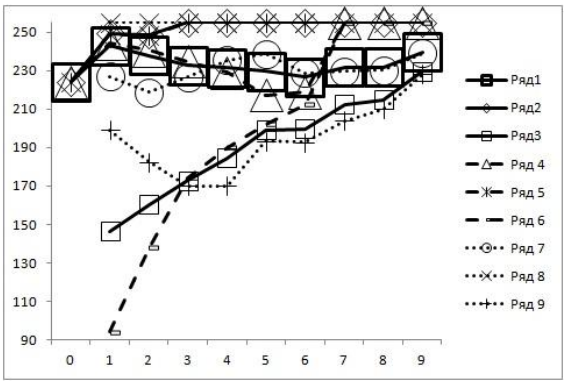


б) 1- E ,2- E_w,3- E_e,4- E^{in},5- E_w^{in},
6- E_e^{in},7- E^{ex},8- E_w^{ex},9- E_e^{ex}

Рис. 4. Класс 4. Признак 7 (уменьшение)



а) 1- N^{ex} ,2- N_e ,3- N_e^{in} ,4- N_e^{ex} ,
5- N_w ,6- N_w^{in} ,7- N_w^{ex} ,8- N^{in}



б) 1- E ,2- E_w ,3- E_e ,4- E^{in} ,5- E_w^{in} ,
6- E_e^{in} ,7- E^{ex} ,8- E_w^{ex} ,9- E_e^{ex}

Рис. 5. Класс 4. Признак 8 (увеличение)

До второго шага устойчивость классификатора плохая, а затем хорошая, а в последней точке посредственная.

При увеличении Π_8 интерполяция имеется до конца интервала искажения. N^{in} (Ряд 8) уменьшается медленнее (Рис. 5, а), чем N_w^{in} (Ряд 6): $N_w^{in} \equiv 0$ с седьмого шага. $N^{in} - N_w^{in} = N_e^{in}$ (Ряд 3) сначала повышается до 10 на пятом шаге, затем понижается: в последней точке $N_e^{in} = N^{in} = 1$. Сходно с динамикой долевого соотношения $N_w^{in} : N_e^{in}$ в N^{in}, E^{in} (Ряд 4) уменьшается от 224, причем $E^{in} \approx 218$, на пятом и шестом шагах, но с седьмого – $E^{in} \equiv 255$ при $N_w^{in} \equiv 0$ (Рис. 5, б). E_e^{in} (Ряд 6) повышается от 94 на первом шаге, и с седьмого – $E_e^{in} \equiv 255$ (неадекватная оценка). E_w^{in} (Ряд 5) почти монотонно увеличивается, и с третьего шага $E_w^{in} \equiv 255$. Интерполяция до четвертого шага хорошая, далее плохая.

N^{ex} (Ряд 1), N_e^{ex} (Ряд 4) и N_w^{ex} (Ряд 7) увеличиваются до 32, 19 и 13 к концу интервала искажения (Рис. 5, а). $E_w^{ex} \equiv 255$ (Ряд 8) с первого до последнего шага (Рис. 5, б). E^{ex} (Ряд 7) от 227 на первом шаге до 239 в конце совершает небольшие колебания, в согласии с динамикой соотношения N_w^{ex} / N_e^{ex}. $E_e^{ex} = 199$ (Ряд 9) на первом шаге, затем немного колеблется, повышаясь к концу до 228. Диапазоны E_w^{ex} и E_e^{ex} разделены, но слабо. Экстраполяция посредственная на двух участках: это отрезок от первого до второго шага, а также седьмой шаг; в промежутке между ними – хорошая, после них – плохая.

Устойчивость классификатора посредственная до третьего шага (Рис. 5, а), далее хорошая.

Было показано [12], что в рамках четвертого класса, как и в приведенных выше примерах, для внутренних элементов N^{in} убывает, как и в целом N_w^{in}. При этом N_e^{in} сначала повышается, а затем понижается, так что либо достигает нулевого значения, либо превышает его до конца интервала интерполяции.

Для внешних элементов N_e^{ex} в ряде случаев монотонно нарастает, а в других – остается на низком уровне.

Проведенное исследование показало, как изменение каждого из признаков влияет на функцию оценки (E^{in} для интерполяции и E^{ex} для

экстраполяции), которая, в свою очередь, определяет динамику долевого соотношения числа правильных распознаваний и ошибок (соответственно $N_w^{in}:N_e^{in}$ в N^{in} и $N_w^{ex}:N_e^{ex}$ в N^{ex}).

И при увеличении, и при уменьшении каждого из параметров крови совпадают интервалы интерполяции. Она, аналогично изложенному выше, является хорошей или везде, или только вначале, а затем варьируется, в том числе становится плохой, посредственной, опять хорошей…

Представленный материал позволяет сделать выводы о поведении оценок и наличии случаев, в которых дальнейшее искажение невозможно, из-за проблемы оценивания.

Устойчивость классификатора существенно варьируется, в чем можно убедиться на приведенных вариантах.

Для рассмотренного в [13] второго класса (небольшие отклонения состояния здоровья от нормы) динамика величин на этапе интерполяции схожа с имеющейся в четвертом классе. При экстраполяции либо вообще отсутствует правильная классификация, либо ошибочное распознавание становится преобладающим по отношению к правильному, которое в том числе уменьшается до нуля; исключением является противоположное соотношение указанных величин при увеличении Π_6. Учитывая наличие ошибки в неискаженном состоянии, устойчивость классификатора вначале является плохой, а затем или остается таковой до установления общего числа ошибок на максимуме, или варьируется, в том числе становится хорошей, посредственной, опять плохой…

Для третьего класса [14] (существенные отклонения состояния здоровья от нормы) динамика величин на этапе интерполяции в ряде вариантов схожа с имеющейся во втором и четвертом классах, однако имеющаяся в неискаженном состоянии одна ошибка исчезает при увеличении Π_3. Для экстраполяции наблюдается более разнообразное поведение N_w^{ex} и N_e^{ex}. То же относится и к устойчивости классификации, которая вначале является плохой ввиду наличия ошибки в неискаженном состоянии.

Динамика долевого соотношения числа N_w^{in}: N_e^{in} в N^{in} и $N_w^{ex}:N_e^{ex}$ в N^{ex} для второго и третьего класса [13, 14] аналогична имеющейся в четвертом классе.

Заключение

Проведено исследование способности классификатора к обобщениям на основе разработанной математической модели искажения исходного обучающего множества. В качестве такового использовалась реальная база параметров крови, дифференцированная по четырем классам, соответствующим различным стадиям отклонения от нормы состояния пищеварительной системы для мужчин. В рамках численного моделирования интерполяционные и экстраполяционные свойства описаны поотдельности.

Введенные в данной работе величины N^{in}, N^{ex} соответствуют числу внутренних и внешних элементов, которые в процессе постепенного увеличения или уменьшения одного из восьми параметров крови остались внутри или вышли за пределы исходных отрезков принадлежности признаков. Динамика N^{in} и N^{ex} при изменении каждого из параметров крови является исконной характеристикой самой базы. В то же время оценочная функция E^{in} и E^{ex} определяется и структурой исходной выборки, и построенным на ее основе классификатором, причем способ реализации последнего может иметь решающее значение.

Особенностью первого класса (практически здоровые) является отсутствие ошибок на этапе интерполяции; проблемы с оцениванием не возникают. Для экстраполяции ошибки появляются позже первого шага, с некоторым интервалом запаздывания, а далее их количество увеличивается и в ряде вариантов становится равным числу элементов. Количество случаев правильной классификации сначала нарастает, а затем после достижения некоторого максимума падает.

Для остальных классов картина искажения более сложная на обоих этапах искажения. Было показано, что для интерполяции характер изменения E^{in} определяет динамику долевого соотношения $N_w^{in} : N_e^{in}$ в N^{in}. Для экстраполяции также продемонстрировано, что E^{ex} влияет на $N_w^{ex} : N_e^{ex}$ в N^{ex}.

Указанные закономерности в первом классе для интерполяции не рассматриваются, поскольку нет ошибок, а для экстраполяции имеют место, за исключением случаев с неадекватными оценками E_e^{ex}=255.

Безошибочная классификация в рамках интерполяции для первого класса подтверждает, что можно использовать диапазоны параметров крови, в рамках которых человек практически здоров.

Мы разработали и реализовали способ описания обучающего множества отдельно взятого класса, как единой случайной величины и описали ряд ее свойств. Помимо этой чисто научной проблемы, предложенный подход может иметь и практическое применение.

Получаемые в реальных условиях результаты анализа крови являются лишь приблизительно верными. Забор заданного объема крови происходит не вполне точно. Каждый применяемый на практике автоматический анализатор имеет некоторую точность измерения. Да и капелька, взятая из пальца, не является полностью идентичной всему объему крови в организме человека. Поэтому необходимо, чтобы классификатор не менял принадлежность к классу в пределах совокупной погрешности. Мы показали, как можно проводить этот анализ. Однако на наших базах, весьма ограниченных по объему, не представляются практически значимыми дополнительные исследования с дроблением начального отрезка искажения, чтобы получить конкретные цифры.

Подчеркнем, что доступные нам базы включают лишь восемь параметров крови. Если представить, что количество показателей крови будет существенно увеличено, то с большой вероятностью практикующий врач в них просто запутается, а когда дочитает до конца, забудет, что было вначале. Напротив, классификатор, построенный с использованием этих данных, станет только лучше.

Рассмотренные здесь уникальные базы показателей крови являются редким предметом для подобных публикаций. Однако предложенные подходы универсальны, и при необходимости можно аналогичным образом проанализировать множества иной природы**.**

Литература

1. Р.В. Ставицкий, Л.А. Лебедев, А.Л. Лебедев, А.Ю. Смыслов. Количественная оценка гомеостатической активности здоровых и больных людей. - М.: ГАРТ. 2013. 131 с.
2. Н.Ю. Добровольская, Л.А. Лебедев, А.Л. Лебедев, Ю.Б. Новожилов, Р.В. Ставицкий. Химио-лучевая терапия рака шейки матки. Методика оценки состояния организма и его систем // Радиология-практика. 2011. №3. С.53-63.
3. И.М. Лебеденко, Т.З. Чернявская, Р.В. Ставицкий, О.Н. Плаутин. Технический контроль состояния организма и его систем в процессе химио-лучевой терапии и трансплантации костного мозга при острых лейкозах // Медицинская техника. 2014. №5. С.32-36.
4. Б.М. Гавриков, И.М. Лебеденко, Н.В. Пестрякова, Р.В. Ставицкий. Об одном статистическом методе оценивания состояния здоровья человека // Труды ИСА РАН, 2016. Т. 66. № 2. С. 54-59.
5. Б.М. Гавриков, Н.В. Пестрякова. О построении признакового пространства в задаче обучения // Информационные технологии и вычислительные системы. 2018. №1. С. 22-29. DOI: 10.14357/20718632180104.
6. Б.М. Гавриков, Н.В. Пестрякова, Р.В. Ставицкий. О свойствах обучающих множеств // Информационные технологии и вычислительные системы. 2018. №4. С.97-107. DOI: 10.14357/207186321804010.
7. Б.М. Гавриков, М.Б. Гавриков, Н.В. Пестрякова. Статистический метод распознавания на основе нелинейной регрессии // Математическое моделирование. 2020. Т.32. №4. С.116-130. DOI: 0.20948/mm-2020-04-09.
8. Б.М. Гавриков, Н.В. Пестрякова. Об экстраполяционных свойствах статистического классификатора // Информационные технологии и вычислительные системы, 2020. № 4. С. 79-90. DOI:10.14357/20718632200407.
9. М.Б. Гавриков, О.В. Локуциевский. Начала численного анализа. — М.: Янус, 1995.
10. Schürmann J. Pattern Classification. — New York: John Wiley&Sons, Inc., 1996.
11. Б.М. Гавриков, М.Б. Гавриков, Н.В. Пестрякова. О свойствах классификатора при нормальных параметрах крови // Препринты ИПМ им. М.В. Келдыша, 2021. № 36. 23 с. DOI: 10.20948/prepr-2021-36.
12. Б.М. Гавриков, М.Б. Гавриков, Н.В. Пестрякова. О свойствах классификатора при максимальных отклонениях параметров крови от нормы // Препринты ИПМ им. М.В. Келдыша, 2021. № 64. 27 с. DOI: 10.20948/prepr-2021-64.

13. Б.М. Гавриков, М.Б. Гавриков, Н.В. Пестрякова. О свойствах классификатора при незначительных отклонениях параметров крови от нормы // Препринты ИПМ им. М.В. Келдыша, 2021. № 70. 26 с. DOI: 10.20948/prepr-2021-70.

14. Б.М. Гавриков, М.Б. Гавриков, Н.В. Пестрякова. О свойствах классификатора при значительных отклонениях параметров крови от нормы // Препринты ИПМ им. М.В. Келдыша, 2021. № 31. 27 с. DOI: 10.20948/prepr-2021-31.

Гавриков Борис Михайлович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Аспирант. Количество печатных работ: 24 (в т.ч.3 монографии). Область научных интересов: вычислительная математика, распознавание образов, медицинская физика. E-mail: bmgavrikov@gmail.com

Гавриков Михаил Борисович. Федеральное государственное учреждение "Федеральный исследовательский центр "Институт прикладной математики им. М.В. Келдыша" Российской академии наук", г. Москва, Россия. Старший научный сотрудник, кандидат физико-математических наук. Количество печатных работ: более 150 (в т.ч.3 монографии). Область научных интересов: вычислительная математика и физика, распознавание образов, плазмодинамика. E-mail: mbgavrikov@yandex.ru

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 (Ответственный за переписку)

On the Ability of a Statistical Classifier to Generalize

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

[I] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
[II] Federal Research Center “Keldysh Institute of Applied Mathematics” of Russian Academy of Sciences, Moscow, Russia

Abstract. The ability to generalize the statistical classifier designed to assess the state of human health by the parameters of peripheral blood is being studied. A mathematical model is described and implemented for the numerical study of interpolation and extrapolation properties of the classifier developed by the authors, based on the polynomial-regression approach and having probabilistic estimates.
Keywords: human health state, body system, peripheral blood, classification, polynomial regression, learning set.

DOI 10.14357/20718632210404

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. N.IU.Dobrovolskaia, L.A.Lebedev, A.L.Lebedev, IU.B.Novozhilov, R.V.Stavitskii. Khimio-luchevaia terapiia raka sheiki matki. Metodika otsenki sostoianiia organizma i ego sistem // Radiologiia-praktika. 2011. №3. S.53-63.
3. I.M.Lebedenko, T.Z.Cherniavskaya, R.V.Stavitskii, O.N.Plautin. Tekhnicheskii control sostoyaniia organizma i ego sistem v protsesse khimio-luchevoi terapii i transplantatsii kostnogo mozga pri ostrykh leikozakh // Meditsinskaia tekhnika. 2014. №5. S.32-36.
4. 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.
5. 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
6. 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.
7. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. Statisticheskiy metod raspoznavaniya na osnove nelineynoy regressii. // Matematicheskoye modelirovaniye. 2020. T.32. №4. S.116-130. DOI: 0.20948/mm-2020-04-09.
8. Gavrikov B.M., Pestryakova N.V. Ob ekstrapolyatsionnykh svoystvakh statisticheskogo klassifikatora. // Informatsionnyye tekhnologii i vychislitel'nyye sistemy, 2020. № 4. S. 79-90. DOI:10.14357/20718632200407.
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.
11. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. O svoystvakh klassifikatora pri normal'nykh parametrakh krovi // Preprinty IPM im. M.V. Keldysha, 2021. № 36. 23 s. DOI: 10.20948/prepr-2021-36.
12. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. O svoystvakh klassifikatora pri maksimal'nykh otkloneniyakh parametrov krovi ot normy // Preprinty IPM im. M.V. Keldysha, 2021. № 64. 27 s. DOI: 10.20948/prepr-2021-64.
13. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. O svoystvakh klassifikatora pri neznachitel'nykh otkloneniyakh parametrov krovi ot normy // Preprinty IPM im. M.V. Keldysha, 2021. № 70. 26 s. DOI: 10.20948/prepr-2021-70.
14. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. O svoystvakh klassifikatora pri znachitel'nykh otkloneniyakh parametrov krovi ot normy // Preprinty IPM im. M.V. Keldysha, 2021. № 31. 27 s. DOI: 10.20948/prepr-2021-31.

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

Gavrikov M. B. PhD in Physics and Mathematics. Federal Research Center “Keldysh Institute of Applied Mathematics” of Russian Academy of Sciences, Moscow, Russia, e-mail: mbgavrikov@yandex.ru

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