Об экстраполяционных свойствах статистического классификатора*

Б. М. Гавриков, Н. В. Пестрякова

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

Аннотация. Рассматривается проблема определения способности к экстраполяции статистического классификатора, предназначенного для оценивания состояния здоровья человека по параметрам периферической крови. Используется численный подход в исследовании характеристик множества, полученного из обучающего в процессе постепенно увеличивающегося искажения. Описаны экстраполяционные свойства классификатора и динамика генерируемых им вероятностных оценок.

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

DOI 10.14357/20718632200407

Введение

Одной из характеристик качества метода распознавания или классификации является представление о его экстраполяционных свойствах. Однако идеология данного понятия чисто интуитивная и весьма расплывчатая, а точное определение отсутствует. В настоящей работе предложена и реализована методология, предназначенная для заполнения этого пробела в рамках математического моделирования.

Предметом рассмотрения является разработанный авторами классификатор состояния здоровья человека (СЗЧ) по показателям периферической крови (из пальца) [1-4]. Он построен с применением статистического подхода, базирующегося на полиномиальной регрессии и генерирующего вероятностные оценки.

Создание статистического классификатора СЗЧ по параметрам периферической крови – дело новое и необычное. Основанием для развития этого направления служит ключевая концепция крупнейших гематологов, что многие заболевания человека вносят изменения в состав его крови. При этом организм рассматривается не в виде разнородных, не связанных между собой органов, а как совокупность их систем.

При оценке СЗЧ гематологи предлагают использовать не менее пяти показателей периферической крови [5-8]. В разработанном классификаторе отдельно для мужчин и женщин по восьми параметрам крови производится деление на четыре класса, соответствующие стадиям поражения систем организма (СО): здоровые (К1), начальные и выраженные отклонения состояния здоровья (К2 и К3), тяжелое заболевание (К4) [1-4].

В данной статье представлена математическая модель, предназначенная для численного

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

исследования динамики характеристик распознавания множеств, на которых осуществлялось обучение классификатора, в процессе их постепенно увеличивающегося изменения. Каждый признак рассматривается отдельно. Дано определение экстраполяционных свойств метода распознавания (классификации). Описано поведение оценок, проанализированы признаки ухудшения используемого способа оценивания, а также правомочность использования самой функции оценки.

1. Метод классификации

Общепринятые обозначения и размерность используемых восьми показателей крови следующие: 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).

Введем вектор $\mathrm{v}\in\mathrm{R}^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови, где N=8. Отождествим 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 $Y=\{\mathrm{e}_1,\ldots,\mathrm{e}_K\}$.

Пусть существует $p_k(\mathrm{v})$ – 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:

$$\mathrm{p}_r(\mathbf{v})=\max_k\{\mathrm{p}_k(\mathbf{v})\},\ 1\le k\le K. \qquad (1)$$

Приближенные значения $\mathrm{p}_1(\mathbf{v}),\ldots,\mathrm{p}_K(\mathbf{v})$ представляются в виде конечных многочленов от координат $\mathbf{v}=(\mathrm{v}_1,\ldots,\mathrm{v}_N)$ и определяются выбором базисных мономов:

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

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

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

Представим (2) в векторном виде:

$$\mathbf{p}(\mathbf{v})=(\mathrm{p}_1(\mathbf{v}),\ldots,\mathrm{p}_K(\mathbf{v}))^{\mathrm{T}}\cong A^{\mathrm{T}}\mathbf{x}(\mathbf{v}), \qquad (3)$$

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 $\mathbf{x}(\mathbf{v})$.

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

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

Правую часть (4) получаем посредством рекуррентной процедуры [9].

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

$$\mathbf{x}=(1,\{v_i\},\{v_iv_j\},\{v_iv_jv_k\},\{v_iv_jv_kv_l\}),\quad 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8. \qquad (5)$$

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 К1 и К4 содержат по 33 набора крови, а классы К2 и К3 включают соответственно 17 и 26 элементов.

Обученный классификатор обеспечил 99,2% правильной классификации на обучающем множестве; имеется по одной ошибке на элементах К2 и К3.

2. Численное моделирование искажений

Далее при нумерации классов К1, К2, К3, К4 используется символ $c, 1\le c\le 4$; количество элементов обозначено |К1|, |К2|, |К3|, |К4|. При-

знаки П1, …, П8 перенумерованы посредством символа $b, 1 \le b \le 8$; в Табл. 1 даны названия, обозначения и размерность соответствующих им параметров крови.

В Табл. 2 для К1, К2, К3, К4, приведены диапазоны, которым принадлежат используемые первичные признаки П1,…,П8 ($D_b^c \equiv [l_b^c, r_b^c], 1 \le c \le 4, 1 \le b \le 8$). 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 К1 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

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

На рисунках с литерой «а» для каждого из рассматриваемых классов по отдельности показано, как нарастает число модифицированных элементов, у которых значение фиксированного признака П1, …, П8 вышло за границы отрезка, определяемого этими номерами класса и признака ($D_b^c, 1 \le c \le 4, 1 \le b \le 8$), а именно, за правый его конец (число элементов N_{b+}^c) при увеличении признака, обозначенного П1+, …, П8+ (Ряд1) или за левый (N_{b-}^c) при его уменьшении: П1-, …, П8- (Ряд3). Также показано, как при этом изменяется количество ошибок N_e (Ряд2 и Ряд4 соответственно).

Будем оценивать *экстраполяционные свойства* метода следующим образом. Если на каком-то интервале изменения t имеем: $N_e < N_{b+}^c$ ($N_e < N_{b-}^c$), то считаем, что наблюдается *хорошая экстраполяция*, напротив, при выполнении $N_e > N_{b+}^c$ ($N_e > N_{b-}^c$) *экстраполяция плохая*. Равенство соответствующих величин означает *посредственное* качество.

Это деление условное, но оно является информативным и позволяет внести ясность при исследовании динамики искажений. На основе предложенных к рассмотрению величин нетрудно ввести числовые характеристики экстраполяционных свойств метода.

Пусть некоторое количество модифицированных элементов данного класса (c_0) на шаге t

Табл. 1. Соответствие признаков параметрам крови

b	*наименование*	*обозначение*	*размерность*
1	эритроциты	RBC	[L^{-1}]
2	тромбоциты	PLT	[L^{-1}]
3	лейкоциты	WBC	[L^{-1}]
4	гемоглобин	HGB	[gL^{-1}]
5	лимфоциты	LIMPH	[L^{-1}]
6	лимфоциты	LIMPH	[%]
7	гранулоциты	GRAN	[L^{-1}]
8	гранулоциты	GRAN	[%]

Табл. 2. 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]

отнесено классификатором к рассматриваемому классу (правильная классификация), тогда E_w суть среднее арифметическое их оценок; если ряд таких объектов отошел к другим классам (неправильная классификация), то E_e вычисляется для них как среднее арифметическое оценок альтернативы, соответствующей классу c_0. Если вычисление средней оценки не зависит от правильности классификации, то получаем E_m; эта величина является оценкой центра масс [4] рассматриваемого класса с учетом искажения.

Рисунки с литерой «б» демонстрируют динамику оценки центра масс E_m (Ряд1), а также средних оценок: правильной классификации E_w (Ряд2) и неправильной E_e (Ряд3) при росте значения признака П1, …, П8, а с литерой «в» – аналогичные величины при его падении.

В анализе поведения оценок отражены следующие аспекты. Выполнение условий E_e=255 или E_e<≈255 соответствует фатальному нарушению оценок. Если на некотором отрезке изменения t наблюдается: $\min_{[t_0,t_1]} E_w \le \max_{[t_0,t_1]} E_e$, то разделение оценок по диапазонам отсутствует. Уменьшение величины $(E_w\text{-}E_e)$ говорит об ухудшении разделения оценок.

Используемое в литературе свойство монотонности оценок (уменьшение количества ошибок при повышении оценки, представляющее интерес в области высоких оценок) обозначим *e-monotony*.

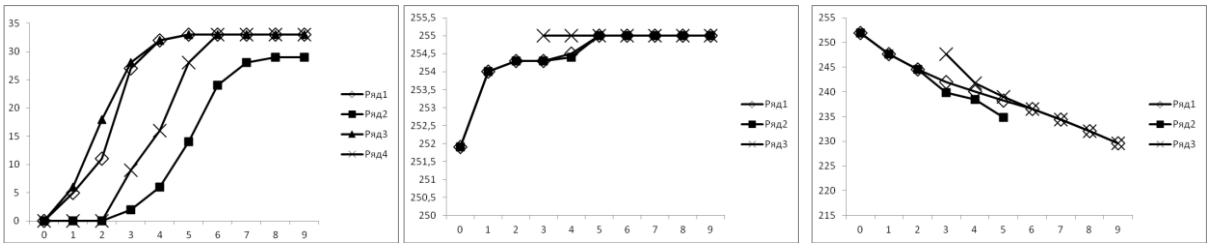
3. Класс 1

Ограничимся пока рассмотрением признаков П1, П2, П3, П4, П6, чтобы исключить вырожденные случаи. На конкретном примере покажем динамику введенных нами величин.

На Рис. 1 изображены зависимости, полученные для класса К1 по признаку П1. При неискаженном К1 количество ошибок нулевое (N_e=0), как и на двух начальных шагах изменения П1 (Рис. 1, а) в обоих направлениях (*устойчивость к небольшим искажениям*). Далее N_e растет, но остается ниже числа модифицированных объектов, у которых П1 вышел за пределы D_1^1 (*хорошая экстраполяция*): $N_e < N_{1+}^1$, $N_e < N_{1-}^1$. Также внутри отрезка [0,4] по оси абсцисс имеет место неравенство: $N_{1-}^1 > N_{1+}^1$, однако различие этих стремительно нарастающих величин незначительное. С середины интервала искажения все значения П1 оказываются за нижней и верхней границами D_1^1 ($N_{1-}^1 = N_{1+}^1$ =|К1|=33); здесь же при нарастании этого признака N_e в два раза меньше, чем при его убывании. Со следующего шага N_e≡|К1| при понижении П1, а при его повышении скорость роста Ne падает, и на последнем отрезке количество ошибок стабилизируется (N_e≡29).

При увеличении П1 изначально высокие оценки (252) правильного распознавания E_w и центра масс E_m нарастают, а с пятого шага выходят на верхний предел 255; оценка неправильного распознавания постоянна: E_e≡255, что говорит о нарушении *e-monotony* (Рис. 1, б). Столь высокий уровень и равенство E_w и E_e делает невозможным использование функции оценки согласно исходной постановке и требует ее переработки.

При уменьшении П1 (Рис. 1, в) все оценки понижаются – в этом нет противоречий с методикой распознавания. Однако на отрезке [3, 5] по оси абсцисс выполняется: $E_e > E_w$, причем начало этого отрезка соответствует высокому уровню оценки E_e (248), что свидетельствует об ухудшении *e-monotony* метода.



а) П1+:(N_{1+}^1 -*N*e), П1-:(N_{1-}^1 -*N*e) *б)* П1+:(E_m, E_w, E_e) *в)* П1-:(E_m, E_w, E_e)

Рис. 1. Признак 1, класс 1

Было показано [10], что для класса К1 на начальных шагах обоих способов изменения каждого из признаков наблюдается устойчивая безошибочная классификация – в этой зоне нет проблем ни с экстраполяцией, ни с оценками. Далее число ошибок увеличивается, в ряде случаев до максимума |К1|. Однако N_e остается ниже стремительно нарастающих N^1_{b+} и N^1_{b-} вплоть до достижения двумя последними указанными величинами значения |К1| и установления на нем (*хорошая экстраполяция*). Различие N^1_{b+} и N^1_{b-} небольшое.

Заметим, что оценка центра масс E_m суть интегральная характеристика, которая вычисляется описанным выше способом по генерируемым классификатором оценкам; именно они определяют выбор того или иного класса здоровья из перечня альтернатив, а характер их изменения при соответствующем искажении отражается на динамике N_e. Уменьшение (увеличение) E_m приводит к росту (падению) N_e. Однозначных зависимостей здесь нет, однако следует учитывать обозначенные корреляции.

Приведенный пример демонстрирует два типа поведения величины E_m при искажении. Опишем соответствующую им динамику N_e по всем восьми признакам для обоих способов модификации вне зоны устойчивой безошибочной классификации. Приведем также обобщенные результаты по оценкам E_w и E_e.

Первый тип для класса К1 обозначим TD(К1), где D - down: изменение признака сопровождается падением E_m. В результате этого N_e либо достигает |К1| (П1-, П2-, П3+, П3-, П4+, П6-), либо нарастает без замедления темпа в конце интервала (П2+). Адекватность оценок вызывает сомнение, поскольку либо диапазоны оценок разделены незначительно (П4+, П6-), либо разделение отсутствует (П1-, П2+, П2-, П3+, П3-), причем в ряде случаев в части интервала искажения выполняется: $E_e > E_w$ (П1-, П2+, П2-). Высокий уровень оценки E_e (П1-, П2+, П2-, П4+) свидетельствует об ухудшении *e-monotony* метода.

Второй тип для класса К1 обозначим TU(К1), где U - up: при изменении признака изначально большая оценка E_m (252) немного увеличивается и выходит на уровень 255 (П1+, П6+) или приближается к нему (П4-). При этом N_e нарастает и устанавливается (или затормаживается), не достигая |К1|; в рассмотренных вариантах не выполняется свойство *e-monotony* (E_e=255). Поскольку оценки изменяются в незначительном диапазоне вблизи максимума, необходимо модифицировать оценочную функцию.

Для вырожденных случаев (П5, П7, П8) по классу К1 [11] при сколь угодно малом отклонении указанных признаков от неискаженного значения N^1_{5-} и N^1_{5+}, N^1_{7-} и N^1_{7+}, N^1_{8-} и N^1_{8+} устанавливаются на максимуме |К1|=33.

Закономерности, имеющиеся в К1 для признаков П1, П2, П3, П4 и П6, сохраняются для П5, П7 и П8, а именно, N_e остается ниже величин N^1_{b+} и N^1_{b-}, при малейшем искажении нарастающих до |К1| и устанавливающихся на этом значении (*хорошая экстраполяция, в ряде случаев наблюдается на малом интервале, полученном дроблением начального отрезка*).

Более того, при увеличении П5 имеется значительная область, где сохраняется классификация без ошибок; при понижении П7 и повышении П8 это распространяется на весь исследуемый интервал искажения. В указанных случаях оценки E_w и E_m с первого шага стабилизируются на максимуме 255, то есть проблемы с экстраполяцией и с оценками отсутствуют.

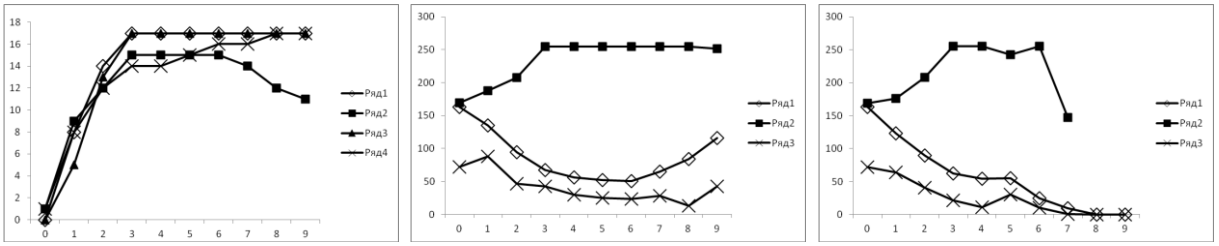
С уменьшением П5 зона правильной классификации занимает только начальную часть первого отрезка. E_m убывает, а затем устанавливается на низком уровне. Оценки адекватные, а их диапазоны разделены.

При повышении П7 и понижении П8 также в начале первого отрезка классификация безошибочная. Сначала E_m падает вплоть или практически до нуля, а затем нарастает. Наблюдается незначительное разделение диапазона оценок.

4. Класс 2

В качестве примера приведем результаты, полученные для класса К2 по признаку П2 (Рис. 2).

Немодифицированный К2 классифицируется с одной ошибкой (N_e=1). На третьем шаге обоих направлений изменения П2 стремительно нарастающие N^2_{2-} и N^2_{2+} достигают максимального значения: $N^2_{2-} = N^2_{2+} = |К2| = 17$



а) П2+:(N^2_{2+} -*N*e), П2-:(N^2_{2-} -*N*e) *б)* П2+:(E_m, E_w, E_e) *в)* П2-:(E_m, E_w, E_e)

Рис. 1. Признак 2, класс 2

(Рис. 2, а). Внутри этого интервала $N^2_{2-} < N^2_{2+}$ при их незначительном различии. На первом шаге $N_e > N^2_{2+}$, $N_e > N^2_{2-}$ (*плохая экстраполяция*), а со второго шага $N_e < N^2_{2+}$, $N_e < N^2_{2-}$. При уменьшении П2 значение N_e в целом растет (при наличии участков локальной стабилизации) с замедлением темпа, и с предпоследнего шага $N_e \equiv |K2|$. При увеличении этого параметра N_e ведет себя немонотонно: сначала быстро повышается, с третьего по шестой шаг не меняется ($N_e \equiv 15$), а на трех последних шагах понижается до значения $N_e = 11$.

С ростом П2 оценка E_w на первых трех отрезках увеличивается от 169 до 255, далее не изменяется, а на последнем шаге незначительно уменьшается до 252 (Рис. 2, б). Оценка E_e имеет локальный максимум на первом шаге, затем постепенно уменьшается; на последних трех шагах наблюдаются колебания (при итоговом росте). Прослеживается взаимное соответствие динамики числа ошибок и оценок. При увеличении П2 сначала N_e повышается (E_m падает, E_w растет), затем N_e стабилизируется с третьего шага – здесь E_w устанавливается на 255, падение E_m почти прекращается; на трех последних интервалах N_e понижается – на них происходит подъем E_m, а также колебание и увеличение E_e, которая перед этим уменьшалась монотонно и постепенно. Локальный максимум E_e на первом шаге приходится на область стремительного роста N_e, а также E_w. Оценки адекватные, причем $\max\limits_{[0,9]} E_e < \min\limits_{[0,9]} E_w$ (выполняется условие разделения диапазона оценок).

При уменьшении П2 (Рис. 2, в) оценка E_w на первых трех интервалах быстрого роста *N*e стремительно увеличивается от 169 до 255 (E_m падает), далее происходит замедление как роста N_e, так и понижения E_m (у последней величины имеется небольшой локальный максимум), а E_w изменяется незначительно до седьмого шага, где резко уменьшается до 147 (далее $N_e \equiv |K2|$, $E_m \equiv E_e \equiv 0$). Оценка E_e постепенно уменьшается, исключением является локальный максимум на пятом шаге (в такт со скачкообразным изменением темпа роста N_e перед выходом на *N*e≡|K2|). Проблемы с оценками не отмечены, причем $\max\limits_{[0,9]} E_e < \min\limits_{[0,7]} E_w$ (диапазоны оценок разделены).

В [10, 11] для класса К2 по обоим способам вариации каждого из признаков показано, как число ошибок меняется от единственной, имеющейся в исходном состоянии. Нарастающие N^2_{b+} и N^2_{b-} не превышают N_e (*плохая экстраполяция*) либо до достижения $N_e = |K2|$ и установления на этом значении (П1+, П3+, П3-, П4+, П5+, П5-, П8+), либо только в первой части интервала модификации, а далее соотношение указанных величин меняется на противоположное (П1-, П2+, П2-, П4-, П6+, П6-, П7+, П7-, П8-). Различие N^2_{b+} и N^2_{b-} как существенное, так и незначительное, в зависимости от признака; обе эти величины увеличиваются и принимают значение |К2|.

Опишем типы поведения E_m при искажении, а также соответствующую им динамику N_e. Обобщим полученные данные по оценкам E_w и E_e.

Первый тип обозначим TD(K2): при изменении признака величина E_m, имеющая достаточно низкое начальное значение (163), падает, вначале быстро, а затем замедленно (или стабилизируется). В результате этого N_e нарастает, а далее устанавливается, при этом либо дости-

гает |К2| (П1+, П2-, П3+, П3-, П4+, П5-, П8+), либо не поднимается до уровня |К2| (П1-, П4-, П6-, П8-). В рассмотренных вариантах оценки представляются адекватными, диапазоны оценок разделены. Лишь в двух случаях разделение незначительное (П6-, П8-), причем для них наблюдается некоторое нарушение в монотонности уменьшения E_m и связанные с этим колебания N_e с сохраняющимся в целом указанным характером динамики этой величины.

Второй тип обозначим TDU(К2): при изменении признака E_m понижается, а затем повышается (П2+, П6+, П7-). Вследствие этого N_e увеличивается, а далее уменьшается. В случае П2+ оценки адекватные, их диапазоны разделены, что не выполняется для П6+ и П7-.

В случае увеличения пятого признака (П5+) E_m падает, а затем растет, однако остается на достаточно низком уровне. При этом N_e повышается, быстро достигает |К2| и устанавливается. Следовательно, имеется значительное сходство с TD(К2). Разделение оценок небольшое.

При увеличении седьмого признака (П7+) на первом этапе E_m растет при аналогичном, но небольшом изменении N_e; на втором этапе E_m стабилизируется вблизи 255 и затем понижается, что сопровождается существенным повышением N_e. Итак, данная динамика величин также соответствует TD(К2). Разделение оценок отсутствует. Высокий уровень оценки E_e говорит об ухудшении *e-monotony*.

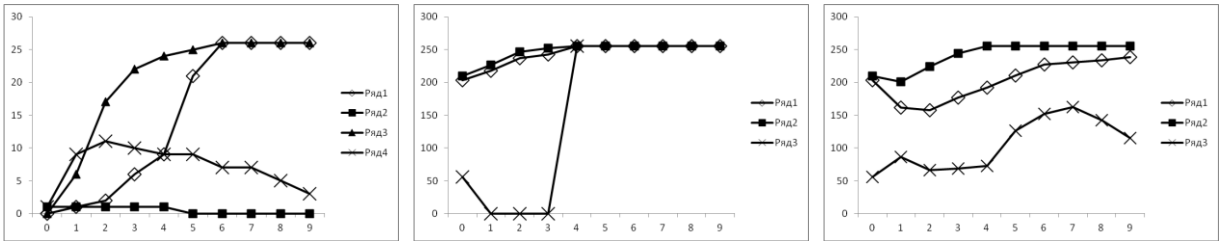
5. Класс 3

Рассмотрим результаты, полученные для класса К3 по признаку П3 (Рис. 3).

В немодифицированном классе К3 имеется одна ошибка (N_e=1). При увеличении П3 вплоть до четвертого шага Ne≡1, а далее число ошибок падает до нуля (при этом E_m увеличивается и стабилизируется на 255 с четвертого шага) (Рис. 3, а,б). Значение N^3_{3+} нарастает и устанавливается на максимуме |К3|=26 с шестого шага, причем на первом шаге Ne= N^3_{3+}, а далее Ne< N^3_{3+} (*хорошая экстраполяция*). Исходная оценка E_w (209) монотонно нарастает и с четвертого шага сохраняется на верхнем уровне 255 (Рис. 3, б). E_e, имеющая неискаженное значение 56, с первого по третий шаг равна нулю, а на четвертом шаге подскакивает до 255, что говорит о нарушении *e-monotony*.

На первых двух отрезках уменьшения П3 наблюдается рост Ne до 11 (сопровождается уменьшением E_m), далее Ne падает (при увеличении E_m); Ne=3 на последнем шаге (Рис. 3, а). N^3_{3-} монотонно увеличивается и с шестого шага сохраняется на максимальном значении, причем на первом шаге Ne> N^3_{3-} (*плохая экстраполяция*), а далее Ne< N^3_{3-}. Внутри отрезка [0,6] выполняется соотношение $N^3_{3-} > N^3_{3+}$, разница между ними значительная. E_w почти монотонно повышается и с четвертого шага устанавливается на 255 (Рис. 3, в). E_e от неискаженного значения 56 сначала в целом очень медленно нарастает до 73 на четвертом шаге, далее темп роста E_e резко увеличивается, а после максимума (E_e=163) на седьмом шаге на двух последних интервалах падает до 115. Выполняется неравенство: $\max\limits_{[0,9]} E_e < \min\limits_{[0,9]} E_w$. На первом шаге E_e и E_w имеют соответственно локальный максимум и минимум, вызванные резким увеличением Ne. Оценки адекватные.

В [10, 11] для класса К3 и обоих способов изменения каждого из восьми признаков продемонстрирован характер динамики количества



а) П3+:(N^3_{3+} -Ne), П3-:(N^3_{3-} -Ne) *б)* П3+:(E_m, E_w, E_e) *в)* П3-:(E_m, E_w, E_e)

Рис. 3. Признак 3, класс 3

ошибок от одной, имеющейся в неискаженном состоянии. Нарастающие величины N^3_{b+} или N^3_{b-} не превышают N_e (*плохая экстраполяция*) либо вплоть до достижения N_e=|К3| и установления на этом значении (П1-, П4+, П7+, П7-), либо только в первой части интервала модификации, а затем соотношение величин становится противоположным (П1+, П2+, П3-, П5+, П6+, П6-, П8+). В ряде случаев наблюдается *хорошая экстраполяция*: с первого до последнего шага N_e не превышает N^3_{b+} или N^3_{b-} (П2-, П3+, П4-, П5-, П8-). В рассмотренных вариантах N^3_{b+} и N^3_{b-} различаются в зависимости от признака по-разному, как существенно, так и незначительно; обе эти величины увеличиваются и достигают значения |К3|, за исключением П2+, П2-, П4+.

Проанализируем типы поведения E_m при искажении, соответствующую им динамику N_e, а также оценок E_w и E_e.

Первый тип обозначим TD(К3): при изменении признака величина E_m, имеющая начальное значение 203, падает, вначале быстро, а затем замедленно (или стабилизируется). В результате этого N_e нарастает, а далее устанавливается, при этом либо достигает |К3| (П1-, П4+, П7-), либо не поднимается до этого уровня (П1+, П2+, П2-, П5+, П6+). В рассмотренных вариантах оценки представляются адекватными, поскольку диапазоны оценок разделены. Это условие нарушено лишь в случае (П1-) в локальном максимуме E_e на первом шаге при невысоком значении данной величины (119).

Второй тип обозначим TDU(К3): при изменении признака E_m понижается, а затем повышается (П3-, П6-). Вследствие этого N_e увеличивается, а далее уменьшается. Диапазоны оценок разделены, оценки адекватные.

Третий тип обозначим TU(К3): при изменении признака E_m повышается до значений, близких или равных максимальному (П3+, П4-, П5-, П8-), при этом N_e стабилизируется на очень низком уровне. Для П3+ E_e принимает значение 255 (нарушение *e-monotony*), оценки неадекватные. В остальных случаях (П4-, П5-, П8-) разделение диапазонов оценок небольшое, E_e поднимается до высокого уровня, возможны проблемы с оценками.

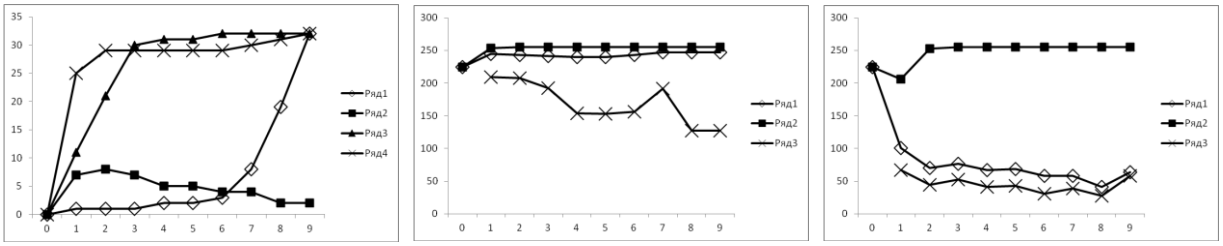
При увеличении седьмого признака (П7+) сначала наблюдается полное соответствие TD(К3); на второй половине интервала искажения N_e устанавливается на максимуме |К3|, а E_m после резких колебаний падает до нуля. В целом данная динамика величин может быть отнесена к TD(К3). Разделение оценок ничтожно мало. Высокий уровень оценки E_e говорит об ухудшении *e-monotony*.

Схожая ситуация наблюдается в случае увеличения восьмого признака (П8+): E_m падает, а на последней трети интервала искажения (где N_e устанавливается ниже уровня |К3|) колеблется при итоговом росте, но остается на низком уровне. Следовательно, имеется значительное сходство с TD(К3). Разделение оценок небольшое.

6. Класс 4

Основные закономерности поведения числа ошибок и оценок для класса К4 сходны с изложенными при рассмотрении К1, К2 и К3. На конкретном примере (К4, П7) продемонстрируем, насколько существенным может быть различие в динамике N^4_{7+} и N^4_{7-}; аналогичная картина наблюдалось также для ряда признаков в классах К2 и К3.

При неискаженном К4 число ошибок нулевое (N_e=0). Увеличение П7 характеризуется тем, что N^4_{7+} к шестому шагу постепенно повышается до 3, затем резко нарастает до 32 на последнем шаге (максимальное значение |К4|=33) (Рис. 4, а). *N*e поднимается до 8 на втором шаге, а далее медленно опускается до 2 в конце интервала искажения. При этом E_m на первом отрезке изменяется от 224 до 244, и после незначительного понижения к последнему шагу нарастает до 247 (Рис. 4, б). До шестого шага *N*e> N^4_{7+} (*плохая экстраполяция*), а далее *N*e< N^4_{7+}. Оценка E_w увеличивается от исходного значения 224 и со второго шага стабилизируется на уровне 255 (Рис. 4, б). E_e в целом падает от 210 на первом шаге до 128 в конце интервала искажения; нарушение гладкости E_e на четвертом шаге, а также локальный максимум на седьмом шаге вызваны изменением *N*e при низком уровне этой величины.



а) П7+:(N^4_{7+} -*N*e), П7-:(N^4_{7-} -*N*e) *б)* П7+:(E_m, E_w, E_e) *в)* П7-:(E_m, E_w, E_e)

Рис. 4. Признак 7, класс 4

С уменьшением П7 вначале N^4_{7-} стремительно растет и на третьем шаге достигает значения 30; далее повышается до 32 на шестом шаге и стабилизируется на этом уровне. В динамике исследуемых величин можно выделить два этапа. Первый: *N*e стремительно увеличивается до 29 на втором шаге (сопровождается понижением E_m от 224 до 70 со схожими темпами), при этом *N*e> N^4_{7-} (*плохая экстраполяция*). Второй: *N*e≡29 до шестого шага, затем число ошибок линейно нарастает до 32 при максимальном искажении (сопутствует колебательная динамика E_m, в целом с понижением до диапазона 41-64 в конце), *N*e< N^4_{7-}. E_w на первом отрезке падает от исходного значения 224 до 206, затем поднимается и с третьего шага устанавливается на 255 (Рис. 4, в). E_e в целом очень незначительно понижается от 67 на первом шаге до 58 на последнем, совершая небольшие колебания.

В каждом из указанных способов изменения П7 выполняется условие разделения диапазонов оценок: $\max\limits_{[1,9]} E_e < \min\limits_{[0,9]} E_w$. При уменьшении П7 оценки адекватные, а при увеличении этого признака разделение небольшое из-за появления на первом шаге ошибок с большими значениями оценок (E_e=210), что говорит об ухудшении *e-monotony*.

В [10, 11] для класса К4 и обоих способов изменения каждого из восьми признаков продемонстрирован характер динамики количества ошибок, отсутствующих в неискаженном состоянии. Нарастающие величины N^4_{b+} или N^4_{b-} не превышают N_e (*плохая экстраполяция*) вплоть до достижения N_e=|К4| и установления на этом значении (П6-, П8-). Наблюдается *плохая экстраполяция* $N^4_{b+} \le N_e$ ($N^4_{b-} \le N_e$) в первой части интервала модификации, а затем соотношение величин становится противоположным (П1+, П1-, П2+, П3+, П4-, П5+, П7+, П7-); в эту градацию включим также случаи с *посредственной экстраполяцией* вначале: для П2- на первом шаге $N^4_{2-}=N_e+1$, а на втором $N^4_{2-}=N_e-1$, а для П8+ до третьего шага $N^4_{8+} \equiv N_e$. Имеется ряд случаев с *хорошей экстраполяцией*: с первого до последнего шага N_e не превышает N^4_{b+} или N^4_{b-} (П3-, П4+, П5-, П6+). В рассмотренных вариантах N^4_{b+} и N^4_{b-} увеличиваются и достигают значения |К4|, за исключением П1+, П1-, П7+, П7-, П8+, П8-.

При рассмотрении признака П3 различие N^4_{3+} и N^4_{3-} также существенное.

Аналогичная картина в классе К2 для N^2_{b+} и N^2_{b-} наблюдается по признакам П3, П5, П6, П7, а в К3 для N^3_{b+} и N^3_{b-} по признакам П2 и П6, в меньшей степени по П3 и П7 [10, 11]. Этот способ деления N^c_{b+} и N^c_{b-} визуальный и весьма приблизительный, однако нетрудно ввести соответствующую интегральную оценку степени сходства.

Приведем общие для класса К4 результаты по типам динамики E_m, N_e, а также E_w и E_e.

Первый тип обозначим TD(К4): при изменении признака величина E_m, имеющая начальное значение 224, падает. В результате этого N_e нарастает, при этом либо достигает |К4| (П6-), либо не поднимается до этого уровня (П2+, П4-, П7-, П8-). В рассмотренных вариантах оценки адекватные, поскольку диапазоны оценок существенно разделены, E_e принимает низкие значения.

Второй тип обозначим TDU(К4): при изменении признака E_m понижается, а затем повышается (П1+, П2-, П3+, П5+). Вследствие этого N_e увеличивается, а далее уменьшается. Для П1+ и П2- диапазоны оценок существенно разделены, E_e находится на низком уровне, оценки адекватные. В двух оставшихся случаях есть проблемы с оценками, поскольку для П3+ разделение оценок небольшое, а для П5+ оно отсутствует, а E_e поднимается до высоких значений.

Третий тип обозначим TU(К4): при изменении признака E_m увеличивается от исходной величины 224 до значений, близких или равных максимальному, при этом N_e стабилизируется на очень малых величинах (П1-, П4+, П7+) или не достигает больших (П3-, П5-, П6+, П8+). Для П1- и П4+ диапазоны оценок существенно разделены, E_e находится на низком уровне, оценки адекватные. В остальных случаях имеются проблемы с оценками: разделение диапазонов оценок либо незначительное (П5-, П7+), либо вообще отсутствует (П3-, П6+, П8+), E_e поднимается до высокого уровня. Для П6+ E_e принимает значение 255 (нарушении *e-monotony*), оценки неадекватные.

Заключение

Предложена методика и разработана математическая модель численного исследования экстраполяционных способностей метода распознавания (классификации). Введены новые понятия, позволяющие проследить за динамикой качества распознавания объектов, представляющих собой заданную модификацию элементов обучающего множества при увеличении различия между исходными и полученными векторными величинами.

Способ деления экстраполяционных свойств классификатора на «хорошее» и «плохое» позволил формализовать интуитивное представление о них, а также с использованием Ne, N^c_{b+} и N^c_{b-} описать процесс искажений. На основе этих величин нетрудно получить иные числовые характеристики качества метода.

В рамках данной модификации исследовалось поведение оценки центра масс E_m и средних оценок правильной и неправильной классификации (E_w, E_e).

Показано, что, несмотря на разнообразие результатов для наборов комбинаций класс&признак, наблюдается сходство динамики исследуемых величин, в основе которого лежит следующая корреляционная зависимость: уменьшение (увеличение) E_m приводит к росту (падению) N_e. Это позволило объединить полученные решения в три типа. Уменьшению E_m соответствует увеличение N_e: TD(К1), TD(К2), TD(К3), TD(К4); E_m падает, а затем нарастает, при этом для N_e наблюдается противоположная динамика: TDU(К2), TDU(К3), TDU(К4); повышение E_m при низком уровне N_e: TU(К1), TU(К3), TU(К4).

В классе К1 для TU(К1) полученные данные несколько смазаны, поскольку изменение оценок E_m происходит в весьма узком диапазоне вблизи максимума 255. Высокий уровень оценок делает невозможным использование функции оценки согласно исходной постановке и требует ее переработки. Это чисто техническая проблема вызвана тем, что мы искусственным образом обрезаем «хвосты» реального расчетного интервала вероятностных оценок, выходящие за пределы отрезка [0,1]. Собственно при классификации и распознавании это значения не имеет, но для модифицированных объектов ситуация меняется, и требуется расширить диапазон оценок.

Динамика E_m более гладкая, чем для E_w и E_e, поскольку набор объектов в первом случае (E_m) фиксированный, постепенно меняется только значение одного признака. Гладкость E_m свидетельствует о непрерывности функции оценки. Совокупность объектов при вычислении E_w и E_e меняется скачкообразно (элементы синхронно переходят из нераспознанных в распознанные или наоборот). Отсюда следует нарушение гладкости E_w и E_e.

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. Б.М. Гавриков, И.М. Лебеденко, Н.В. Пестрякова, Р.В. Ставицкий. Об одном статистическом методе оцени-

вания состояния здоровья человека. // Труды ИСА РАН, 2016. Т. 66. № 2. С. 54-59.
2. Гавриков Б.М., Пестрякова Н.В. О построении признакового пространства в задаче обучения // Информационные технологии и вычислительные системы. 2018. №1. С. 22-29. DOI: 10.14357/20718632180104
3. 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
4. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Статистический метод распознавания на основе нелинейной регрессии. // Математическое моделирование. 2020. Т.32. №4. С.116-130. DOI: 0.20948/mm-2020-04-09
5. 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.
6. 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.
7. 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.
8. Ю.А.Цыбульская, Р.В.Ставицкий, И.М.Лебеденко, С.В.Смердин, И.В..Шутихина, Л.С.Коков, О.В.Батурин. Количественный подход к диагностике поражения костно-суставной системы при туберкулезном спондилите // Медицинский вестник Северного Кавказа. 2015. №3. С.212-217.
9. Schürmann J. Pattern Classification. — New York: John Wiley&Sons, Inc. 1996.
10. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Об устойчивости статистического классификатора состояний систем организма человека к искажениям // Препринт ИПМ им. М.В. Келдыша. 2020. №32. 40 с. https://doi.org/10.20948/prepr-2020-32
11. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Об устойчивости статистического классификатора состояний систем организма человека к искажениям в вырожденных случаях // Препринт ИПМ им. М.В. Келдыша. 2020. №49. 26 с. https://doi.org/10.20948/prepr-2020-49.

Гавриков Борис Михайлович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Аспирант. Окончил Московский инженерно-физический институт (МИФИ) в 2014г. Количество печатных работ: 21 (в т.ч.3 монографии). Область научных интересов: вычислительная математика, распознавание образов, медицинская физика. E-mail: bmgavrikov@gmail.com

Пестрякова Надежда Владимировна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Ведущий научный сотрудник, доктор технических наук. Окончила Московский физико-технический институт (МФТИ) в 1983г. Количество печатных работ: более 100 (в т.ч.1 монография). Область научных интересов: вычислительная математика и физика, распознавание образов. E-mail: pestryakova@isa.ru

On Extrapolation Properties of the Statistical Classifier

B. M. Gavrikov, N. V. Pestryakova

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

Abstract. The problem of determining the ability to extrapolate a statistical classifier intended for assessing the state of human health by the parameters of peripheral blood is considered. A numerical method is used to study the characteristics of the set obtained from the training in the process of gradually increasing distortion. The extrapolation properties of the classifier and the dynamics of the probabilistic estimates generated by it are described.

Keywords: human health state, body system, peripheral blood, classification, polynomial regression, learning set.

DOI 10.14357/20718632200407

References

1. 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.
2. 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.
3. 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
4. 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.
5. 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.
6. 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.
7. 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.
8. IU.A.Tsybulskaia, R.V.Stavitskii, I.M.Lebedenko, S.V.Smerdin, I.V.Shutikhina, L.S.Kokov, O.V.Baturin. Kolichestvennyi podkhod k diagnostike porazheniia kostno-sustavnoy sistemy pri tuberkuleznom spondilite // Meditsinskii vestnik Severnogo Kavkaza. 2015. №3. S.212-217.
9. Schürmann J. Pattern Slassification. — New York: John Wiley&Sons, Inc. 1996.
10. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. Ob ustoychivosti statisticheskogo klassifikatora sostoyaniy sistem organizma cheloveka k iskazheniyam // Preprint IPM im. M.V. Keldysha. 2020. №32. 40 s. https://doi.org/10.20948/prepr-2020-32.
11. Gavrikov B.M., Gavrikov M.B., Pestryakova N.V. Ob ustoychivosti statisticheskogo klassifikatora sostoyaniy sistem organizma cheloveka k iskazheniyam v vyrozhdennykh sluchayakh // Preprint IPM im. M.V. Keldysha. 2020. №49. 26 s. https://doi.org/10.20948/prepr-2020-49.

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

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