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

Препринты ИПМ • Препринт № 76 за 2022 г.

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

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

Прототип системы поддержки принятия решений в медицинской диагностике на основе статистического подхода

Рекомендуемая форма библиографической ссылки: Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Прототип системы поддержки принятия решений в медицинской диагностике на основе статистического подхода // Препринты ИПМ им. М.В.Келдыша. 2022. № 76. 23 с.
https://doi.org/10.20948/prepr-2022-76
https://library.keldysh.ru/preprint.asp?id=2022-76

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

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

Прототип системы поддержки принятия решений в медицинской диагностике на основе статистического подхода

Москва — 2022

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

Прототип системы поддержки принятия решений в медицинской диагностике на основе статистического подхода

Предложен статистический подход для предварительной диагностики заболеваний человека по параметрам крови в качестве системы поддержки принятия врачебных решений. Авторами разработан классификатор, основанный на полиномиальной регрессии и генерирующий вероятностные оценки. Реализованы приложения, позволяющие как проводить диагностику состояния пациента от практически здорового до максимальной степени поражения (онкология) для отдельной системы организма, так и определять область локализации опухоли при онкологическом заболевании.

Ключевые слова: онкологическое заболевание, система организма, периферическая кровь, классификация, полиномиальная регрессия, обучающее множество

Boris Mikhailovich Gavrikov, Mikhail Borisovich Gavrikov, Nadejda Vladimirovna Pestryakova

Prototype of a decision support system in medical diagnostics based on a statistical approach

A statistical approach is proposed for the preliminary diagnosis of human diseases by blood parameters as a system for supporting medical decision-making. The authors have developed a classifier based on polynomial regression and generating probabilistic estimates. Applications have been implemented that allow both diagnosing the patient's condition from practically healthy to the maximum degree of damage (oncology) for a particular body system, and determining the area of tumor localization in oncological disease.

Key words: cancer, body system, peripheral blood, classification, polynomial regression, training set

Оглавление

Введение

В последние годы существенно вырос масштаб информатизации медицины. Одно направление – организационное, позволяющее на базе современных технологий решать проблемы «учета и контроля» потоков пациентов: удаленная запись к специалистам, электронная медицинская карта и пр. Имеющиеся здесь проблемы носят чисто технический характер.

Наблюдается также интерес медицинского сообщества к появлению систем искусственного интеллекта (ИИ), которые оказывали бы помощь практикующим врачам. Эта проблематика требует тщательной проработки.

Проведенная 28.07.2022 ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава России научно-практическая конференция «Повышение доверия к системам искусственного интеллекта в здравоохранении» является свидетельством актуальности данного направления в медицине.

Подлежали обсуждению вопросы доверия к системам ИИ с точки зрения разработчиков и инвесторов. Отмечалось, что один из ключевых барьеров для более широкого внедрения и повышения спроса на ИИ-системы – необходимость формировать доверие к их безопасности и эффективности. Однако устоявшегося понимания – как именно это должно быть сделано, сейчас в практическом здравоохранении и у регуляторов нет.

Предлагались практические рекомендации по полноте и структуре отчетов о проведении машинного обучения, а также тестирования и валидации полученных моделей. Целью является повышение прозрачности процессов машинного обучения. Результат – повышение доверия регуляторов, экспертов и лиц, принимающих решения в сфере применения и тиражирования ИИ-систем.

Таким образом, выстраивается линия взаимодействия руководства медицинских организаций с разработчиками программного продукта.

В этой системе общественного разделения труда мы занимаем особую нишу: проводим численный эксперимент в рамках новых оригинальных направлений в медицинской диагностике. Мы руководствуемся идеями, основанными на опыте авторитетных представителей медицинского сообщества, и формализуем их в виде математических моделей. Объем фактических данных, имеющихся в нашем распоряжении, весьма ограничен, поэтому полученные результаты носят либо качественный, либо рекомендательный характер. Однако они представляют интерес как для врачей-диагностов, так и для более могущественных разработчиков ИИ-систем.

Объект классификации. Системы организма и классы здоровья

Описываемая в настоящей работе система поддержки врачей при постановке диагноза основывается на концепции крупнейших гематологов, что многие заболевания человека вносят изменения в состав его крови [1].

Проведенные нами статистические исследования баз параметров крови служат подтверждением правомочности этой концепции [2-8].

При оценке состояния здоровья человека (СЗЧ) гематологи предлагают использовать не менее пяти показателей периферической крови (из пальца) [1]. Организм представляется в виде совокупности систем органов (пищеварения, дыхания и пр.). Мужчины и женщины рассматриваются по отдельности, поскольку диапазоны вариации показателей крови среди множества людей существенно зависят от пола.

Предварительная диагностика, о которой идет речь в данной работе, должна удовлетворять таким требованиям, как общедоступность, быстрота, дешевизна. Использование разработанного авторами классификатора соответствует всем этим условиям. Полученные с его помощью результаты позволят врачу оперативно сориентироваться и назначить при необходимости дорогостоящие уточняющие диагностические процедуры и исследования с применением высокотехнологичной аппаратуры.

В серии работ мы показали, что статистический метод классификации может успешно применяться для оценивания СЗЧ по показателям периферической крови [2-8].

СЗЧ включает четыре градации – от практически здорового состояния до максимальной степени поражения организма: здоровые, начальные отклонения состояния здоровья, выраженные отклонения, тяжелое заболевание (онкология).

По каждой системе организма (СО) проводилось исследование СЗЧ при помощи соответствующего классификатора, обученного на профильной выборке, включающей восемь параметров периферической крови. Базы показателей крови были созданы с использованием верифицированных диагнозов [1].

В качестве следующего шага мы предложили способ предварительной диагностики онкологических заболеваний человека различной локализации. Обучение статистического классификатора проводилось на наборах показателей периферической крови пациентов с опухолями в ряде СО.

Эти два подхода, замкнутые воедино, могут использоваться в качестве прототипа системы поддержки принятия решений в медицинской диагностике.

Используется восемь показателей крови. Их общепринятые обозначения и размерность: 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).

Рассмотрим четыре СО для мужчин: пищеварительная система, органы дыхания, опорно-двигательный аппарат, урологическая система. В таблице 1 приведены данные о том, как именно элементы обучающих последовательностей различных СО количественно распределены по классам:

«1» – практически здоровые, «2» – начальные отклонения состояния здоровья, «3» – выраженные отклонения, «4» – тяжелые заболевания (онкология).

Таблица 1

Распределение объема базы по классам здоровья

Пищеварительная система - 1				
Распределение объема базы по классам здоровья				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	17	26	33	109
Органы дыхания - 2				
Распределение объема базы по классам здоровья				
«1»	«2»	«3»	«4»	«1-2-3-4»
32	11	12	21	76
Опорно-двигательный аппарат - 3				
Распределение объема базы по классам здоровья				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	3	7	33	76
Урологическая система - 4				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	18	26	32	109

Данные, приведенные в таблице 1, соответствуют объектам классификации двух видов.

Во-первых, они используются для обучения классификатора по каждой отдельно взятой СО. При этом классы, которые отражают СЗЧ, обозначены «1», «2», «3», «4» с указанным количеством элементов.

Во-вторых, был построен и обучен классификатор, определяющий локализацию онкологии для четырех приведенных СО. Поклассовая нумерация приведена в таблице 1 рядом с названием СО, а число элементов каждого класса содержится в столбце «4».

Метод классификации

Изложим постановку задачи в двух вариантах. **Вариант 1** – классификатор, позволяющий определить класс здоровья; строится отдельно по каждой СО. **Вариант 2** предназначен для нахождения СО, в которой локализована онкологическая опухоль.

Вариант 1. Рассматриваем определенную СО человека. Вводим вектор $\mathbf{v}\in\mathbf{R}^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови, причем N=8.

Нормировка на отрезок [0,1] проводится следующим образом. По обучающей выборке 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.

$$\begin{aligned} v_i^{\min} &= \min_j \{v_i^j\},\ j=1,\ldots,J, \\ v_i^{\max} &= \max_j \{v_i^j\},\ j=1,\ldots,J, \end{aligned} \tag{1}$$

где J – объем выборки по данной СО.

Затем выполняем следующее преобразование:

$$v_i \rightarrow (v_i - v_i^{min})/(v_i^{max} - v_i^{\min}). \tag{2}$$

Отождествляем k-й элемент множества классов СЗЧ с базисным вектором $\mathbf{e}_k$=(0…1…0) (здесь 1 находится на k-м месте, $1\le k\le K$, причем K=4) из $\mathbf{R}^K$. Обозначаем $Y=\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$.

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

$$p_r(\mathbf{v}) = \max_k \{p_k(\mathbf{v})\},\ 1\le k\le K. \tag{3}$$

Вариант 2. Рассматриваем K определенных перенумерованных СО человека, $1\le k\le K$. Вводим вектор $\mathbf{v}\in\mathbf{R}^N$, i-я компонента которого – отнормированная на отрезок [0,1] величина i-го показателя крови онкобольных, где N=8.

Нормирование проводим следующим образом. Рассмотрим все четыре обучающие выборки из столбцов «4» исследуемых СО. Для каждого i-го показателя крови находим (1) минимальное и максимальное значение $v_i^{\min}$, $v_i^{\max}$, где i = 1,..., N. В (1) J – суммарный объем выборок из столбцов «4» по совокупности рассматриваемых СО. Затем выполняем преобразование (2).

Отождествляем k-й элемент множества СО с базисным вектором из $\mathbf{R}^K$: $\mathbf{e}_k$=(0…1…0), причем 1 находится на k-м месте, $1\le k\le K$. Обозначим $Y=\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$.

Пусть существует $p_k(\mathbf{v})$ – вероятность того, что набор отнормированных показателей крови онкобольных соответствует k–му элементу СО, где $1 \le k \le K$. Искомый элемент СО будет иметь порядковый номер r, получивший максимальное значение вероятности (3).

Вариант 1, 2. 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:

$$p_k(\mathbf{v}) \cong c_0^{(k)} + \sum_{i=1}^{N} c_i^{(k)} v_i + \sum_{i,j=1}^{N} c_{i,j}^{(k)} v_i v_j + \ldots,\ 1 \le k \le K. \qquad (4)$$

Представим упорядоченные базисные мономы из (4) в виде вектора размерности L:

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

Тогда (4) можно записать в векторном виде:

$$\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}),\ldots,p_K(\mathbf{v}))^{\mathrm{T}} \cong A^{\mathrm{T}}\mathbf{x}(\mathbf{v}), \qquad (5)$$

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 (4) (с совпадающим верхним индексом), упорядоченных так же, как в векторе $\mathbf{x}(\mathbf{v})$.

Значение A вычисляется приближенно в процессе обучения с использованием базы данных: $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$. Здесь используются следующие обозначения.

Вариант 1: $\mathbf{v}^{(j)}$ – набор параметров крови, соответствующий элементу СЗЧ с номером k ($1 \le k \le K$).

Вариант 2: $\mathbf{v}^{(j)}$ – набор параметров крови, соответствующий элементу СО с номером k ($1 \le k \le K$).

Вариант 1, 2: $\mathbf{y}^{(j)}$= (0…1…0) – базисный вектор, где 1 стоит на k-м месте, $1 \le j \le J$).

Итак,

$$A \cong \left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathrm{T}}\right)^{-1}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{y}^{(j)})^{\mathrm{T}}\right). \qquad (6)$$

Поскольку проблема обращения заполненной матрицы большой размерности до сих пор не решена [9], правую часть (6) получаем посредством рекуррентной процедуры [10].

Использовалась различные модификации вектора $\mathbf{x}(\mathbf{v})$. Указаны типы классификатора (**Вариант 1** и **Вариант 2**), а также в первом случае СО, для которых при $\mathbf{x}(\mathbf{v})$ данного вида были получены наилучшие результаты и более

сложные модификации не использовались. Точность классификации приведена в случаях, когда она меньше 100%.

Далее выражения в фигурных скобках соответствуют цепочкам элементов вектора, вычисляемым по всем показателям крови из имеющегося набора.

1). **Вариант 1.** *Опорно-двигательный аппарат.*

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

Длина полинома 77. Имеются мономы степенного вида первого, второго, третьего, четвертого, пятого и шестого порядка. Перекрестные произведения используются в качестве мономов второго порядка, а для более высоких порядков отсутствуют.

2). **Вариант 1.** *Органы дыхания.*

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

Длина полинома 165. Имеются мономы первого, второго и третьего порядка. Перекрестные произведения используются для мономов второго и третьего порядка.

3). **Вариант 1.** *Пищеварительная система (98,2%). Урологическая система (97,3%).*

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\},\{v_i v_j v_k v_l v_m\}),\\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8,\ l\le m\le 8. \qquad (9)$$

Длина полинома 1287. Имеются мономы первого, второго, третьего, четвертого и пятого порядка. Перекрестные произведения используются для мономов второго, третьего, четвертого и пятого порядка.

4). **Вариант 2.**

$$\mathbf{x}=(1,\{v_i\},\{v_i v_j\},\{v_i v_j v_k\},\{v_i v_j v_k v_l\},\{v_i v_j v_k v_l v_m\},\\ \{v_i v_j v_k v_l v_m v_n\}),\\ 1\le i\le 8,\ i\le j\le 8,\ j\le k\le 8,\ k\le l\le 8,\ l\le m\le 8,\ m\le n\le 8. \qquad (10)$$

Длина полинома 3003. Имеются мономы первого, второго, третьего, четвертого, пятого и шестого порядка. Перекрестные произведения используются для мономов второго, третьего, четвертого, пятого и шестого порядка.

Классификатор обеспечил точность 93,3% на обучающем множестве из 119 элементов (сохраняется восемь ошибок).

Итак, исследуются база периферической крови мужчин по четырем СО. Построены и обучены классификаторы двух типов.

Вариант 1 для каждой из этих СО по отдельности имеет четыре класса, соответствующие градациям СЗЧ: «1» – C^1, «2» – C^2, «3» – C^3, «4» – C^4.

Вариант 2 относится к базам крови онкологических больных по четырем СО для мужчин: пищеварительная система – C^1, органы дыхания – C^2, опорно-двигательный аппарат – C^3, урологическая система – C^4.

Далее **Вариант 1** и **Вариант 2** будут рассмотрены в рамках проводимого сравнительного анализа структуры обучающего множества $C^1UC^2UC^3UC^4$.

Расстояние между «своими» и «чужими» элементами

Для каждого из четырех рассматриваемых классов C^1, C^2, C^3, C^4 в отдельности найдем минимальное, максимальное и среднее расстояние между своими векторами (принадлежащими данному классу). Для множества векторов k-го класса определяем их следующим образом.

Минимальное расстояние:

$$U\_k_{\min} = \min_{V^k} \{\|\mathbf{v}^k - \mathbf{u}^k\|\},\ \mathbf{v}^k \in V^k,\ \mathbf{u}^k \in V^k,\ \mathbf{v}^k \neq \mathbf{u}^k. \tag{11}$$

Максимальное расстояние:

$$U\_k_{\max} = \max_{V^k} \{\|\mathbf{v}^k - \mathbf{u}^k\|\},\ \mathbf{v}^k \in V^k,\ \mathbf{u}^k \in V^k, \tag{12}$$

где $\mathbf{v}^k$ и $\mathbf{u}^k$ – пары различных векторов, принадлежащих множеству элементов k-го класса V^k.

Среднее расстояние определим с приведением алгоритма нахождения этой величины:

$$U\_k_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=j+1}^{J_k} \| \mathbf{w}^{k,j} - \mathbf{w}^{k,j1} \| / (J_k\ (J_k\ - 1)/2),\ \mathbf{w}^{k,j} \in V^k,\ j = 1, \ldots, J_k, \tag{13}$$

где $\{\mathbf{w}^{k,j},\ j = 1, \ldots, J_k\} = V^k$ – представление совокупности элементов k-го класса в виде множества перенумерованных векторов.

Аналогично получим соответствующие значения для пар свой–чужой по каждому из классов. Чужой вектор – не принадлежащий рассматриваемому классу. Для обучающего множества, содержащего элементы всех четырех классов, $V = \{ C^1UC^2UC^3UC^4\}$.

Минимальное расстояние:

$$U\_kz_{\min} = \min_{V} \{\|\mathbf{v}^k - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{v}^k \in V^k,\ \mathbf{u}^{\neg k} \in V^{\neg k}. \qquad (14)$$

Максимальное расстояние:

$$U\_kz_{\max} = \max_{V} \{\|\mathbf{v}^k - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{v}^k \in V^k,\ \mathbf{u}^{\neg k} \in V^{\neg k}, \qquad (15)$$

где $\mathbf{v}^k$ и $\mathbf{u}^{\neg k}$ – пары векторов, из которых $\mathbf{v}^k$ принадлежит множеству элементов k-го класса V^k, а $\mathbf{u}^{\neg k}$ принадлежит множеству чужих элементов $V^{\neg k}$ классов, отличных от k-го: $V^{\neg k} = V \setminus V^k$.

Среднее расстояние:

$$U\_kz_{\text{ср}} = \sum_{j=1}^{J_k} \sum_{j1=1}^{J_{\neg k}} \| \mathbf{w}^{k,j} - \mathbf{w}^{\neg k,j1} \| / (J_k\ J_{\neg k})\ ,\ \mathbf{w}^{k,j} \in V^k,\ j = 1, \ldots, J_k\ , \quad \mathbf{w}^{\neg k,j1} \in V^{\neg k},\ j1 = 1, \ldots, J_{\neg k}\ , \qquad (16)$$

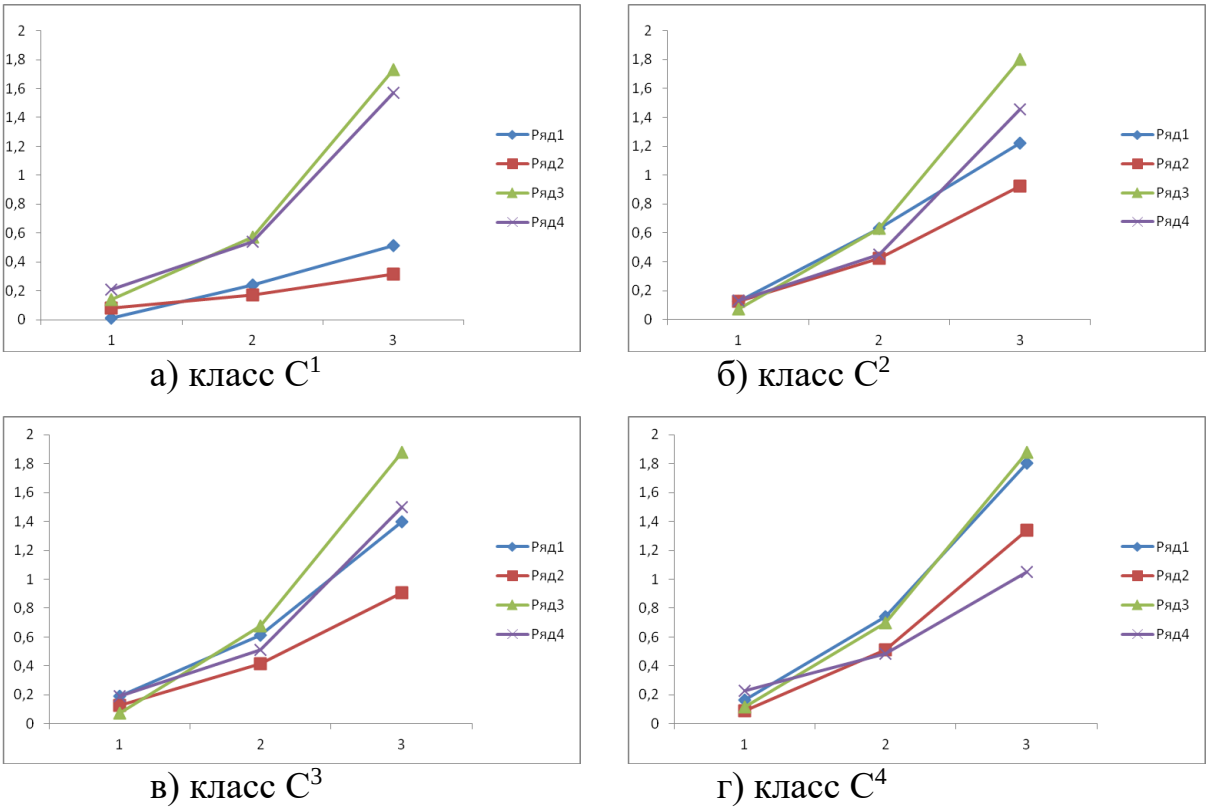
где $\{\mathbf{w}^{k,j},\ j = 1, \ldots, J_k\} = V^k$ – представление совокупности своих элементов k-го класса в виде множества перенумерованных векторов, аналогично для множества чужих элементов классов, отличных от k-го: $\{\mathbf{w}^{\neg k,j1},\ j1 = 1, \ldots, J_{\neg k}\} = V^{\neg k}$, $V^{\neg k} = V \setminus V^k$.

Продемонстрируем, какие значения принимают перечисленные величины. Расстояние между векторами определяем в метрике L_2.

Для классов C^1, C^2, C^3 и C^4 соответственно представлены (рис.1-5) минимальное, среднее и максимальное расстояние (значения ординат для точек 1, 2, 3 по оси абсцисс) между своими векторами (Ряд 1), аналогичные величины для пар свой–чужой (Ряд 3).

Вариант 1. ***Пищеварительная система.*** Минимальное, среднее и максимальное расстояние между своими векторами (Ряд 1) класса C^1 (рис. 1, а) меньше соответствующих расстояний между парами свой–чужой (Ряд 3). Для других СО результаты по классу C^1 аналогичны (рис. 2-4, а).

Порядок, характерный для класса C^1 (Ряд 1 и Ряд 3), не выполняется для классов C^2, C^3, C^4, (рис. 1, б, в, г). Например, для этих трех классов по минимальному расстоянию (а для класса C^4 и по среднему) Ряд 1 превышает аналогичное значение Ряда 3. Указанные средние величины примерно одинаковы для класса C^2 (рис. 1б).

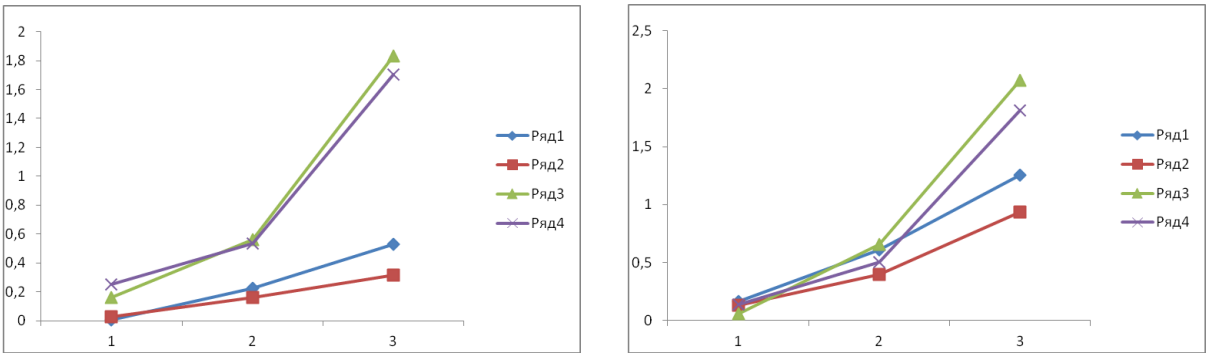


а) класс C^1 б) класс C^2

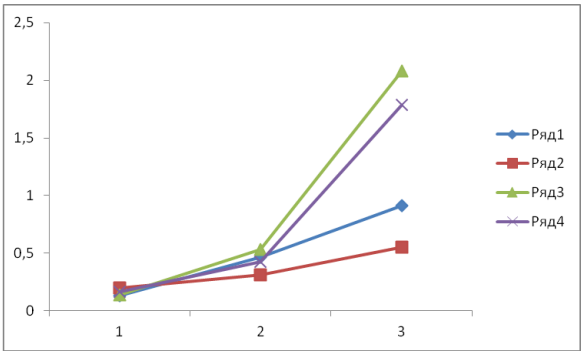
в) класс C^3 г) класс C^4

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

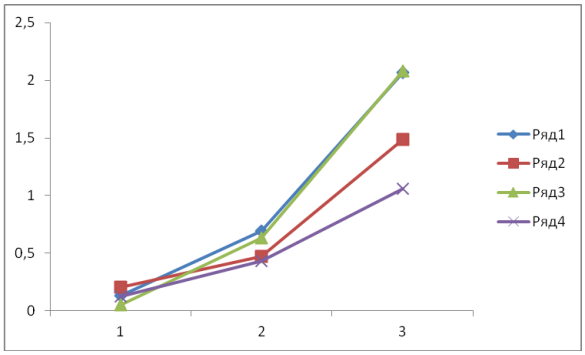
Вариант 1. *Органы дыхания.* Для классов C^2 и C^4 (рис. 2, б, г) Ряд 1 превышает Ряд 3 по минимальным значениям; аналогично и для средних величин по классу C^4. Указанные минимальные величины примерно одинаковы для класса C^3 (рис. 2, в), как и средние для класса C^2 (рис. 2, б), а также максимальные для класса C^4 (рис. 2, г).



а) класс C^1 б) класс C^2



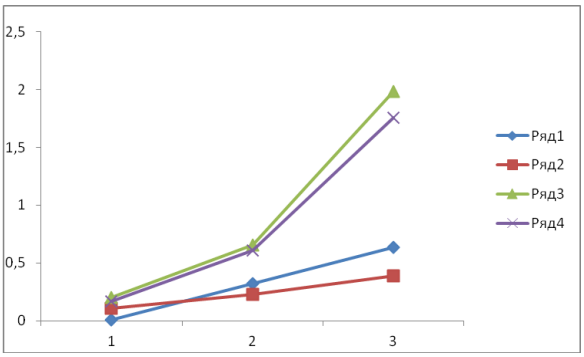
в) класс C^3



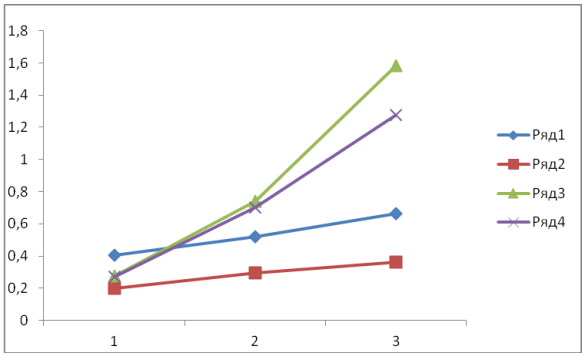
г) класс C^4

Рис. 2. Вариант 1. *Органы дыхания.* Минимальное, максимальное и среднее расстояние между парами векторов: свой–свой, свой–чужой, центр масс–свой, центр масс–чужой.

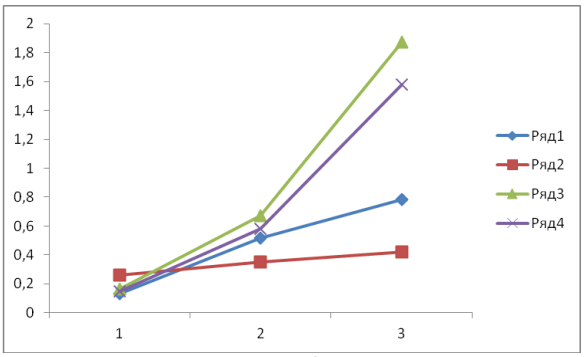
Вариант 1. *Опорно-двигательный аппарат.* Для класса C^2 (рис. 3, б) Ряд 1 превышает Ряд 3 по минимальным значениям, а для класса C^4 (рис. 3, г) – как по средним, так и по максимальным. Указанные минимальные величины для класса C^3 примерно одинаковы (рис. 3, в), как и для класса C^4 (рис. 3, г).



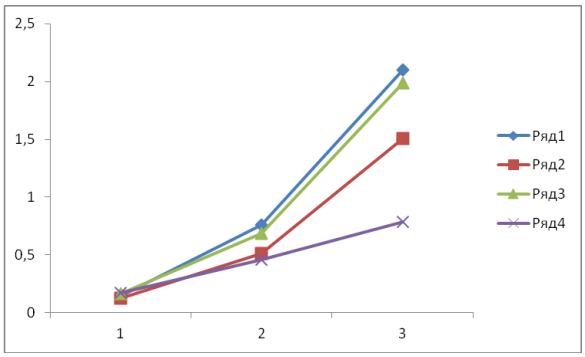
а) класс C^1



б) класс C^2



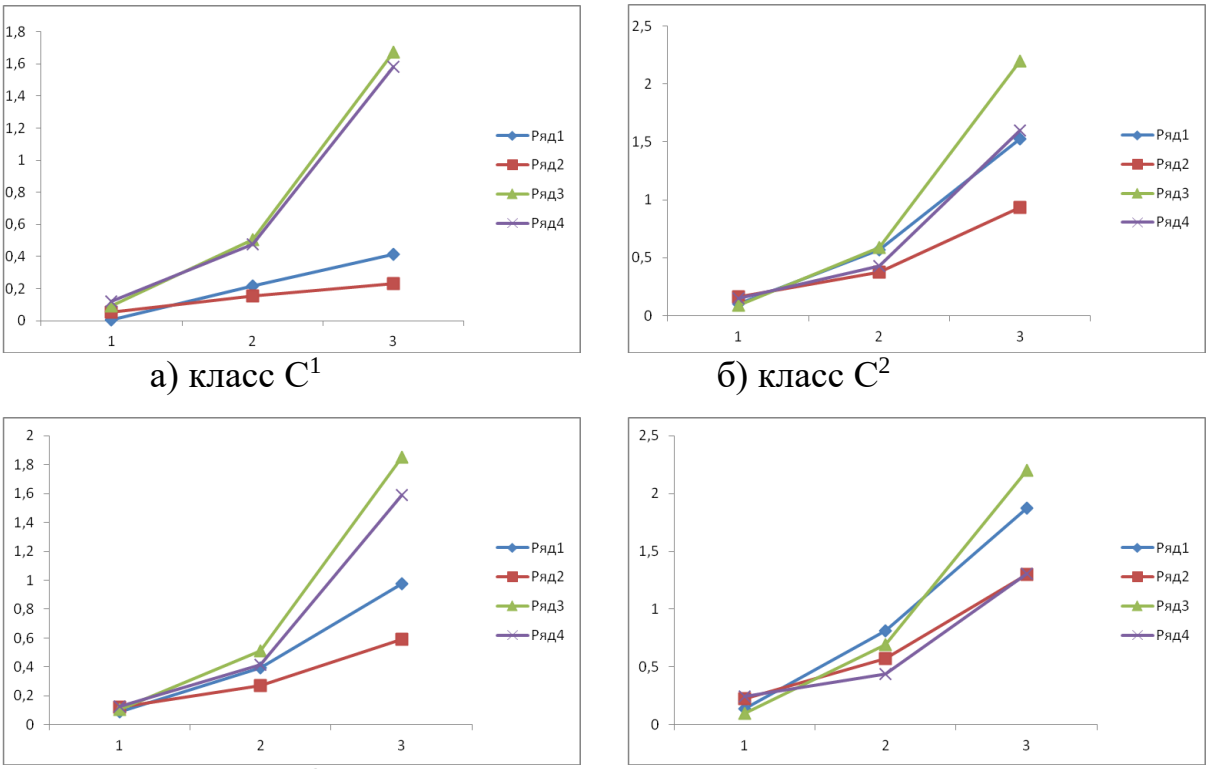
в) класс C^3



г) класс C^4

Рис. 3. Вариант 1. *Опорно-двигательный аппарат.* Минимальное, максимальное и среднее расстояние между парами векторов: свой–свой, свой–чужой, центр масс–свой, центр масс–чужой.

Вариант 1. ***Урологическая система.*** Для класса C^4 (рис. 4, г) Ряд 1 превышает Ряд 3 по минимальным и средним значениям. Указанные минимальное и среднее расстояние для класса C^2 (рис. 4, б), а также минимальное расстояние для класса C^3 (рис. 4, в) приблизительно одинаковы.

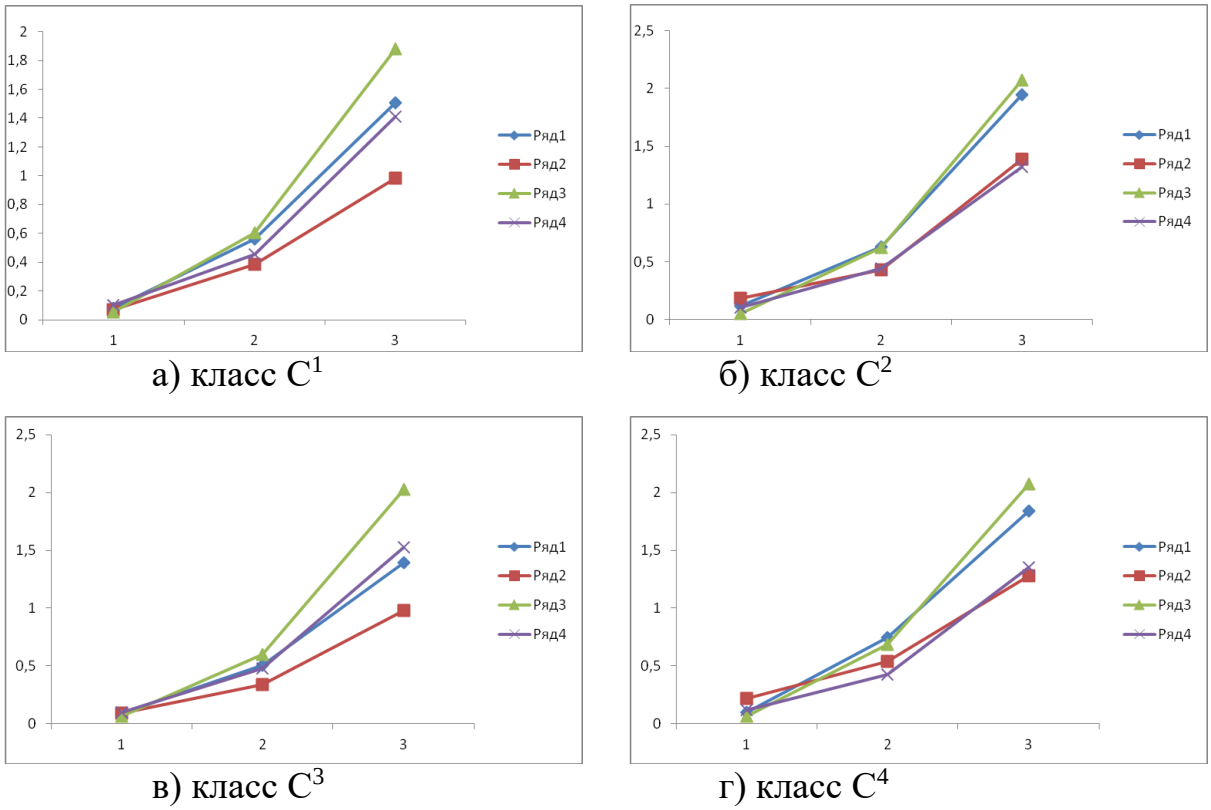


а) класс C^1 б) класс C^2

в) класс C^3 г) класс C^4

Рис. 4. Вариант 1. *Урологическая система.* Минимальное, максимальное и среднее расстояние между парами векторов: свой–свой, свой–чужой, центр масс–свой, центр масс–чужой.

Вариант 2. Для всех четырех классов (рис. 5) Ряд 1 немного превышает Ряд 3 по минимальным значениям (причем обе малы); кроме того, первая величина меньше второй для максимальных значений. Для средних имеются оба варианта различия, но оно небольшое. Следовательно, результаты, полученные для этой базы онкобольных, более однотипные по всем классам, чем по классам СЗЧ.



а) класс C^1 б) класс C^2

в) класс C^3 г) класс C^4

Рис. 5. Вариант 2. Минимальное, максимальное и среднее расстояние между парами векторов: свой–свой, свой–чужой, центр масс–свой, центр масс–чужой.

Отклонение от центра масс своих и чужих элементов

Для каждого из четырех рассматриваемых классов C^1, C^2, C^3, C^4 в отдельности получим среднестатистический вектор длины 8, принадлежащий исходному векторному пространству $\mathbf{R}^8$. Иногда такой вектор называют центром масс.

Для центра масс k-го класса значение i–го параметра крови равно среднему арифметическому значений i-х параметров крови по всем J_k имеющимся в базе наборам показателей крови, относящихся к данному классу:

$$v_i^{k,cp} = (\sum_{j=1}^{J_k} v_i^{k,j}) / J_k, \qquad (17)$$

где $\mathbf{v}^{k,j}$ – перенумерованные элементы k-го класса: $\{\mathbf{v}^{k,j}=(v^{k,j}{}_1,\ldots,v^{k,j}{}_N)$, $j = 1, \ldots, J_k\} = V^k$.

Для каждого из классов C^1, C^2, C^3 и C^4 найдем минимальное, максимальное и среднее расстояния между центром масс и своими векторами.

Указанные величины для множества векторов k-го класса определяем следующим образом. Минимальное расстояние:

$$D\_k_{\min} = \min_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\},\ \mathbf{u}^k \in V^k. \tag{18}$$

Максимальное расстояние:

$$D\_k_{\max} = \max_{V^k} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^k\|\},\ \mathbf{u}^k \in V^k, \tag{19}$$

где $\mathbf{u}^k$ – вектор, принадлежащий множеству элементов k-го класса V^k, $\mathbf{v}^{k,cp}$ – среднестатистический вектор этого класса**.**

Среднее расстояние определим более детально с приведением алгоритма нахождения этой величины:

$$D\_k_{\text{ср}} = \sum_{j=1}^{J_k} \|\mathbf{w}^{k,j} - \mathbf{v}^{k,cp}\| / J_k\ ,\ \mathbf{w}^{k,j} \in V^k, j = 1, \ldots, J_k, \tag{20}$$

где $\{\mathbf{w}^{k,j}, j = 1, \ldots, J_k\} = V^k$ – представление совокупности элементов k-го класса в виде множества перенумерованных векторов.

Аналогично получим соответствующие значения по каждому из классов между центром масс и чужими векторами. Эти результаты зависят от количества классов, входящих в обучающее множество.

Минимальное расстояние:

$$D\_kz_{\min} = \min_{V^{\neg k}} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{u}^{\neg k} \in V^{\neg k}. \tag{21}$$

Максимальное расстояние:

$$D\_kz_{\max} = \max_{V^{\neg k}} \{\|\mathbf{v}^{k,cp} - \mathbf{u}^{\neg k}\|\}\ ,\ \mathbf{u}^{\neg k} \in V^{\neg k}, \tag{22}$$

где $\mathbf{u}^{\neg k}$ – вектор, принадлежащий множеству чужих элементов $V^{\neg k}$ классов, отличных от k-го: $V^{\neg k} = V / V^k$, $\mathbf{v}^{k,cp}$ – среднестатистический вектор k-го класса**.**

Среднее расстояние:

$$D\_kz_{\text{ср}} = \sum_{j=1}^{J_{\neg k}} \|\mathbf{v}^{k,cp} - \mathbf{w}^{\neg k,j}\| / J_{\neg k}\ ,\ \mathbf{w}^{\neg k,j} \in V^{\neg k}, j = 1, \ldots, J_{\neg k}\ , \tag{23}$$

где $\{\mathbf{w}^{\neg k,j}, j = 1, \ldots, J_{\neg k}\} = V^{\neg k}$, $V^{\neg k} = V \setminus V^k$ – представление совокупности чужих элементов классов, отличных от k-го в виде множества перенумерованных векторов**.**

Для каждого класса соответственно (рис. 1-5) из объединения $C^1UC^2UC^3UC^4$ представлено минимальное, среднее и максимальное расстояние (значения ординат для точек 1, 2, 3 по оси абсцисс) между центром масс и своими векторами (Ряд 2), аналогично между парами центр масс–чужой вектор (Ряд 4).

Вариант 1. ***Пищеварительная система.*** Минимальное, среднее и максимальное расстояние между центром масс и своими векторами (Ряд 2) класса C^1 (рис. 1, а) меньше соответствующих расстояний между центром масс и чужими векторами (Ряд 4). Для других СО класс C^1 (рис. 2-4, а) рассматривать не будем ввиду однотипности.

Взаимоположение Ряда 2 и Ряда 4, имеющее место в классе C^1, либо в значительной степени нарушается для других классов (рис. 1, б, в, г), либо различие между указанными величинами становится меньше. Для класса C^4 Ряд 2 превышает Ряд 4 по максимальной величине, а соответствующие средние величины примерно одинаковы; последнее имеет место как для минимальных, так и для средних значений по классу C^2 (рис. 1, б).

Вариант 1. ***Органы дыхания.*** Для класса C^4 (рис. 2, г) Ряд 2 превышает Ряд 4 по минимальным, средним и максимальным значениям, причем для среднего – различие незначительное. Для класса C^2, как и для класса C^3, соответствующие минимальные величины приблизительно одинаковы (рис. 2, б, в).

Вариант 1. ***Опорно-двигательный аппарат.*** Для класса C^4 (рис. 3, г) Ряд 2 превышает Ряд 4 по средним и максимальным значениям, а для класса C^3 (рис. 3, в) аналогично по минимальным величинам, причем для среднего значения (класс C^4) различие незначительное.

Вариант 1. ***Урологическая система.*** Для класса C^4 (рис. 4, г) Ряд 2 превышает Ряд 4 по средним значениям. Для класса C^4 (рис. 4, г) соответствующие минимальные и максимальные величины приблизительно одинаковы; аналогично минимальные – для класса C^2 (рис. 4, б) и класса C^3 (рис. 4, в).

Вариант 2. Для класса C^1 (рис. 5, а) Ряд 4 всюду превышает Ряд 2, причем для минимальных и средних значений разница небольшая, а для максимальных – существенная. Аналогичная картина наблюдается для класса C^3, исключая практически равные минимальные значения (рис. 5, в). Для C^2 соотношения противоположные (при малом различии) для минимальных и максимальных значений, а средние – почти равны (рис. 5, б). В классе C^4 (рис. 5, г) Ряд 2 немного превышает Ряд 4 для минимальных и средних значений, а для максимальных – ситуация противоположная при незначительном различии.

Итак, результаты, полученные для базы онкобольных, по всем классам варьируются меньше, чем для СЗЧ.

Распределение числа своих и чужих элементов при удалении от центра масс

Диапазон расстояний между центром масс k-го класса СО и векторами этого же класса («своими», $\mathbf{v}^k \in V^k$) по рассматриваемой базе, согласно формулам (18), (19), находится на отрезке $[D\_k_{\min},\ D\_k_{\max}]$. Диапазон расстояний между центром масс k-го класса СО и векторами всех других классов («чужими», $\mathbf{z}^k \in \{V \setminus V^k\}$), согласно формулам (21), (22), – на отрезке $[D\_kz_{\min},\ D\_kz_{\max}]$. Пусть

$$\begin{aligned} Dk_{\min} &= \min(D\_k_{\min}, D\_kz_{\min}), \\ Dk_{\max} &= \max(D\_k_{\max}, D\_kz_{\max}). \end{aligned} \qquad (24)$$

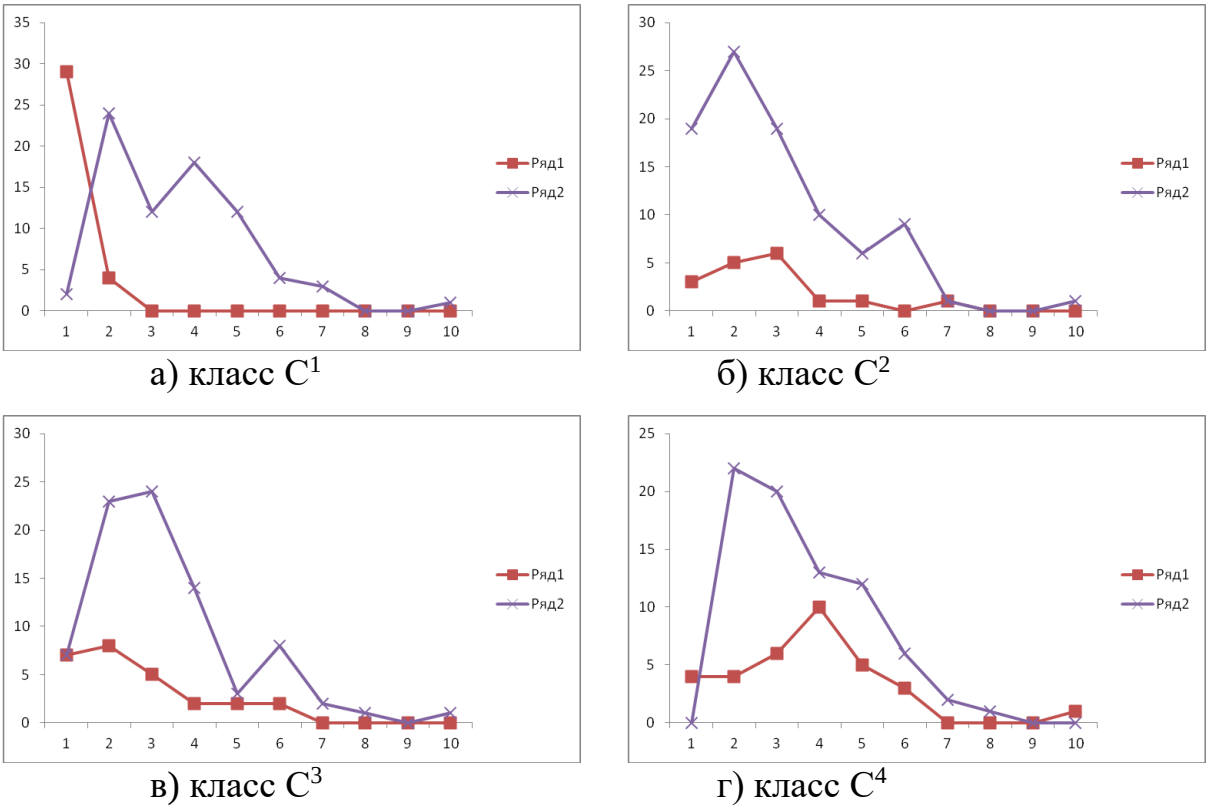
Делим отрезок $[Dk_{\min}, Dk_{\max}]$ (оси абсцисс на рис. 2, а, б, в, г) на десять равных по длине частей – один отрезок и девять полуинтервалов: $[Dk_{\min}, Dk_{\min} + d]$, $(Dk_{\min} + d, Dk_{\min} + 2d]$, … , $(Dk_{\min} + 9d, Dk_{\min} + 10d]$, где $d = (Dk_{\max} - Dk_{\min})/10$. Определим, какое количество своих векторов попало в каждый такой участок (аналогично для чужих векторов). Затем рассмотрим распределение числа своих (чужих) векторов на отрезке $[Dk_{\min}, Dk_{\max}]$.

Для каждого класса (рис. 6-10) соответственно из объединения $C^1UC^2UC^3UC^4$ представлено распределение количества своих (Ряд 1) и чужих (Ряд 2) элементов на отрезке $[Dk_{\min}, Dk_{\max}]$.

Вариант 1. ***Пищеварительная система.*** Вблизи центра масс класса C^1 имеется небольшая окрестность, в которой находятся все его элементы (рис. 6, а), причем их число убывает при удалении от центра масс (Ряд 1). В то же время в этой окрестности чужих элементов мало, а подавляющее их количество располагается вне ее (Ряд 2); соответствующая функция распределения сначала нарастает при удалении от центра масс, а затем имеет тенденцию к убыванию, и ее максимум находится на удалении от этой окрестности, т.е. там, где отсутствуют элементы класса C^1.

Взаиморасположение Ряда 1 и Ряда 2 класса C^1 для прочих СО аналогично (рис. 6, а), поэтому эти случаи (рис. 7-9, а) обсуждать не будем.

Для C^2, C^3, C^4 (рис. 6, б, в, г) Ряд 1 и Ряд 2 в целом нарастают, затем убывают, максимумы смещены от центра масс. В отличие от C^1, C^2, C^3 (рис. 6, а, б, в), для C^4 (рис. 6, г) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$. При этом чужие элементы в конце этого отрезка отсутствуют (рис. 6, г).

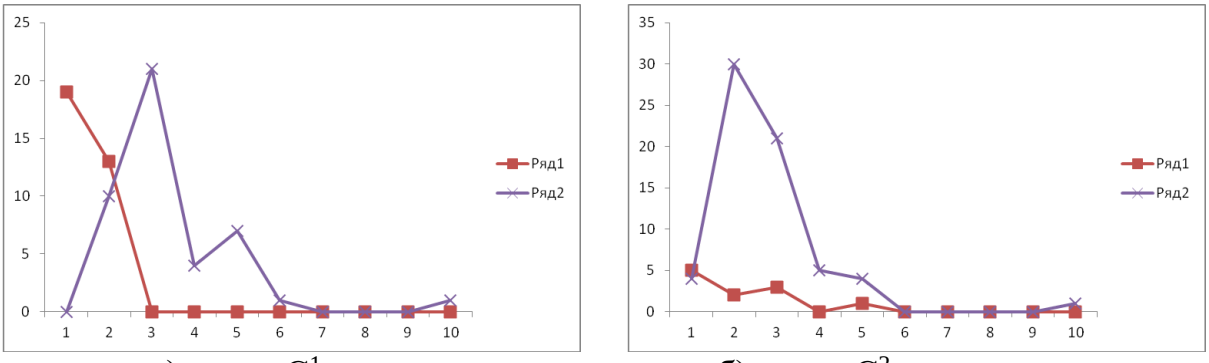


а) класс C^1 б) класс C^2

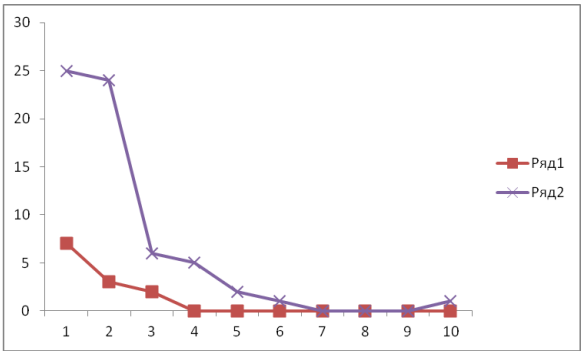
в) класс C^3 г) класс C^4

Рис. 6. Вариант 1. *Пищеварительная система.* Распределение числа своих и чужих элементов при удалении от центра масс

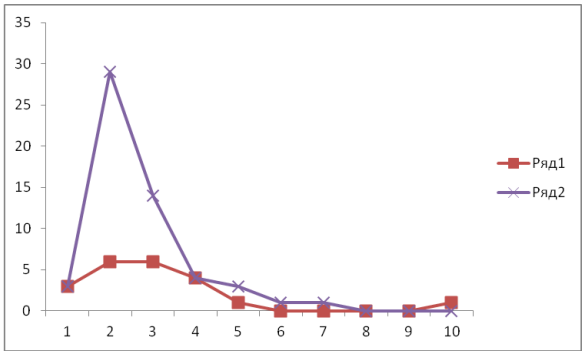
Вариант 1. *Органы дыхания.* Для C^3 (рис. 7, в) монотонно убывает Ряд 1 и в целом Ряд 2. Для C^4 (рис. 7, г) нарастает, затем убывает Ряд 2 и в целом Ряд 1; максимумы смещены от центра масс. Для C^2 (рис. 7, б) Ряд 1 и Ряд 2 немонотонны, причем в целом Ряд 1 убывает, а Ряд 2 схож с C^4. В отличие от C^1, C^2, C^3 (рис. 7, а, б, в), для C^4 (рис. 7, г) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$. При этом чужие элементы в конце этого отрезка отсутствуют (рис. 7, г).



а) класс C^1 б) класс C^2



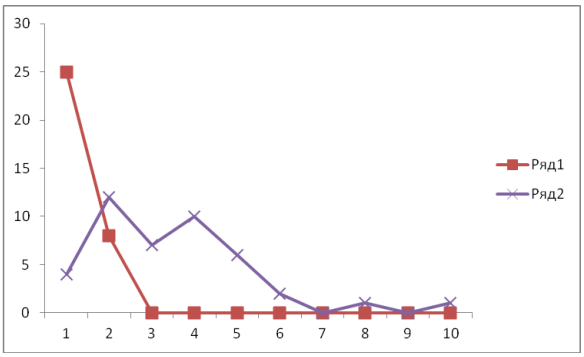
в) класс C^3



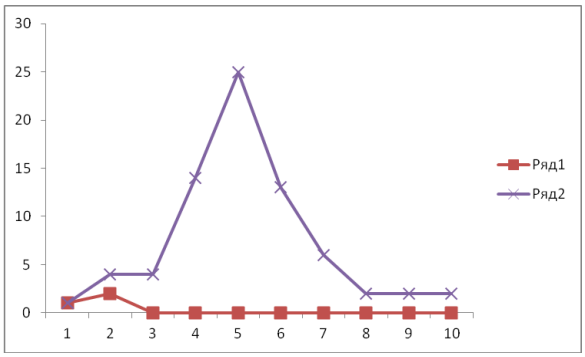
г) класс C^4

Рис. 7. Вариант 1. *Органы дыхания.* Распределение числа своих и чужих элементов при удалении от центра масс

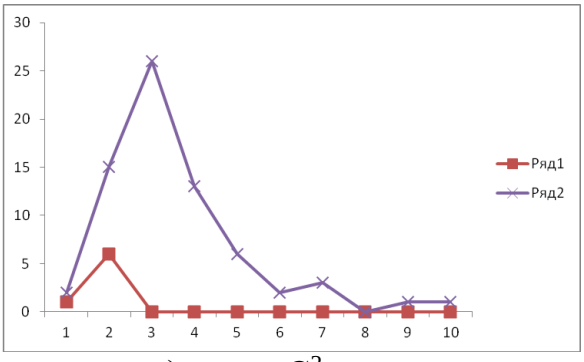
Вариант 1. *Опорно-двигательный аппарат.* Для C^2, C^3, C^4 (рис. 8, б, в, г) Ряд 1 и Ряд 2 в целом нарастают, затем убывают, максимумы смещены от центра масс. В отличие от C^1, C^2, C^3 (рис. 8, а, б, в), для C^4 (рис. 8, г) свои элементы имеются до конца отрезка $[Dk_{\min}, Dk_{\max}]$. При этом чужие элементы в конце этого отрезка отсутствуют (рис. 8, г).



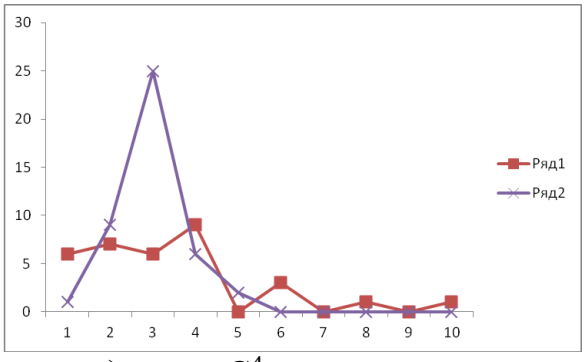
а) класс C^1



б) класс C^2



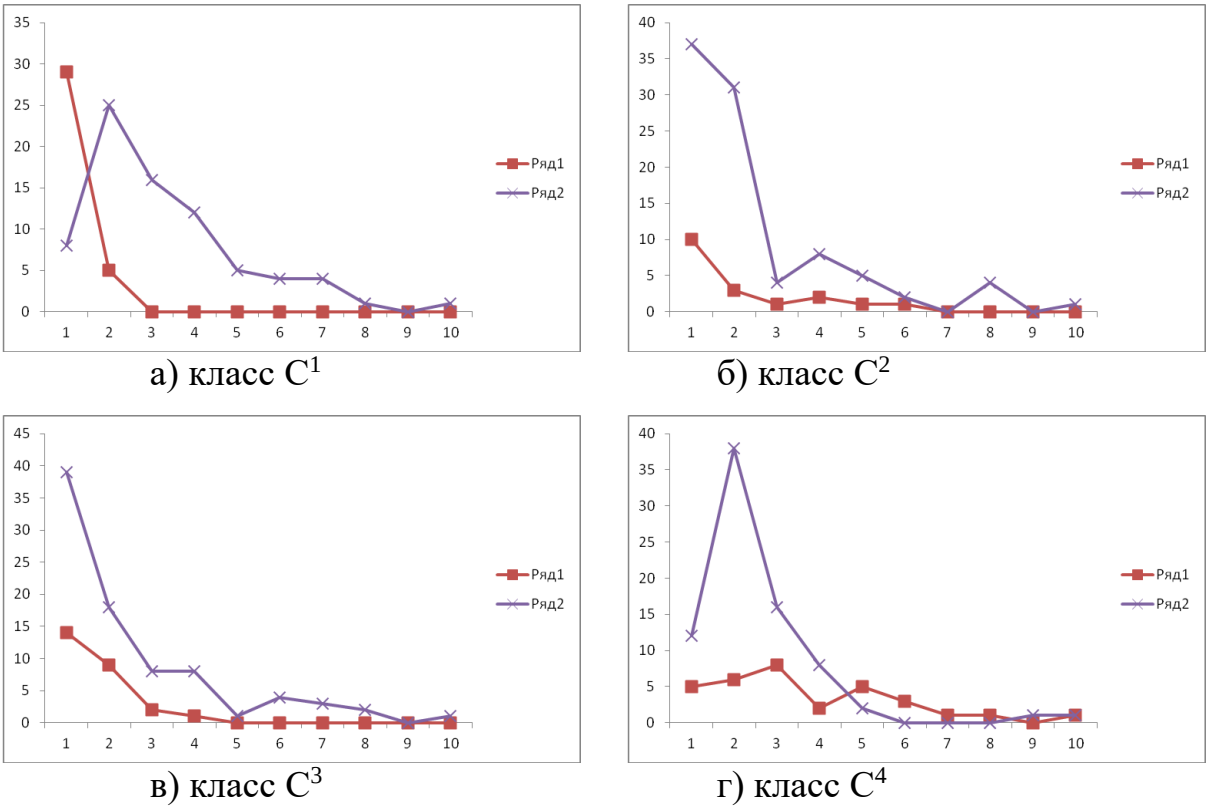
в) класс C^3



г) класс C^4

Рис.8. Вариант 1. *Опорно-двигательный аппарат.* Распределение числа своих и чужих элементов при удалении от центра масс

Вариант 1. ***Урологическая система.*** Для C^2, C^3 (рис. 9, б, в) в целом монотонно убывают Ряд 1 и Ряд 2. Для C^4 (рис. 9, г) Ряд 1 и Ряд 2 в целом нарастают, затем убывают, максимумы смещены от центра масс. В отличие от C^1, C^2, C^3 (рис. 9, а, б, в), для C^4 (рис. 9, г) до конца отрезка $[Dk_{min}, Dk_{max}]$ имеются свои элементы. Также в конце этого отрезка есть и чужие элементы (рис. 9, г).

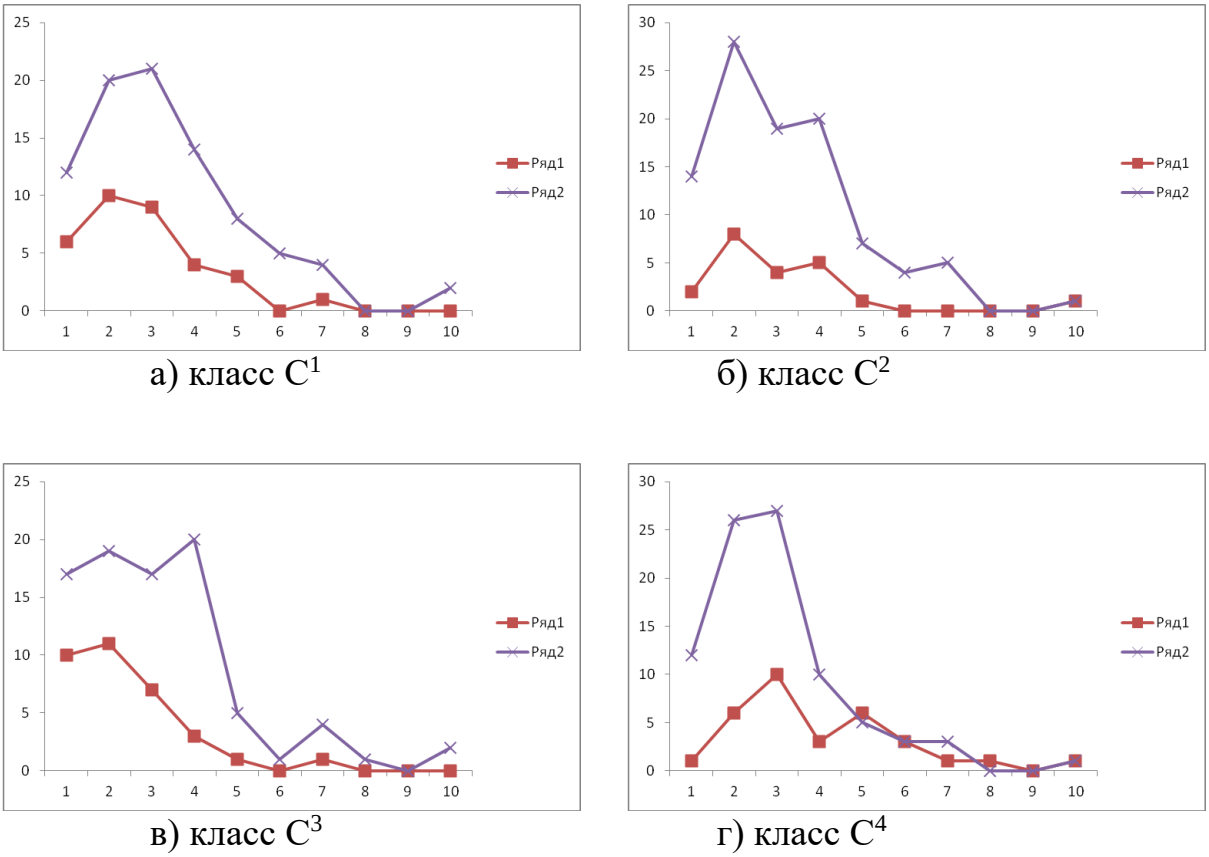


а) класс C^1 б) класс C^2

в) класс C^3 г) класс C^4

Рис. 9. Вариант 1. *Урологическая система*. Распределение числа своих и чужих элементов при удалении от центра масс

Вариант 2. Для C^1, C^2, C^3, C^4 (рис. 10, а, б, в, г) Ряд 1 и Ряд 2 в целом нарастают, затем убывают, максимумы смещены от центра масс. В характере полученных распределений числа своих (чужих) элементов имеется структурное сходство между всеми четырьмя классами. Эта же аналогия наблюдается в отношении своих–чужих элементов. В отличие от C^1 и C^3 (рис.10, а, в), для C^2, C^4 (рис.10, б, г) свои элементы имеются до конца отрезка $[Dk_{min}, Dk_{max}]$. Там же присутствуют и чужие элементы (рис.10, б, г).



а) класс C^1 б) класс C^2

в) класс C^3 г) класс C^4

Рис. 10. Вариант 2. Распределение числа своих и чужих элементов при удалении от центра масс

Заключение

С использованием статистического метода распознавания, основанного на полиномиальной регрессии, разработан классификатор, предназначенный для системы поддержки принятия врачебных решений при предварительной диагностике. Классификатор реализован в виде двух компонент этой системы. **Вариант 1** служит для оценивания состояния здоровья пациента по четырем градациям – от практически здорового до максимальной степени поражения организма (онкология). В этом случае каждая СО исследуется по отдельности. **Вариант 2** позволяет уточнить область локализации онкологической опухоли по показателям периферической крови.

Рассмотрены четыре СО для мужчин: пищеварительная система, органы дыхания, опорно-двигательный аппарат, урологическая система. Проведено исследование структуры обучающего множества, состоящего из четырех классов (C^1, C^2, C^3, C^4). Для **Варианта 1** они соответствуют градациям СЗЧ. **Варианту 2** сопутствует анализ наборов крови онкобольных по каждой из этих СО. Для каждого из четырех классов C^1, C^2, C^3, C^4 в отдельности найдено минимальное, максимальное и среднее расстояние между своими векторами (принадлежащими данному классу).

Также получены соответствующие значения для пар свой–чужой (элемент, не относящийся к рассматриваемому классу).

Для классов C^1, C^2, C^3, C^4 вычислили среднестатистический вектор, принадлежащий исходному векторному пространству $\mathbf{R}^8$ (центр масс). Найдено минимальное, максимальное и среднее расстояние между центром масс и своими (чужими) векторами.

Получено распределение количества своих и чужих элементов на отрезке их нахождения при удалении от центра масс.

Проведенный анализ позволил выявить сходство и различие в структуре множеств C^1, C^2, C^3, C^4 для **Варианта 1** и **Варианта 2**.

Библиографический список

1. Ставицкий Р.В., Лебедев Л.А., Лебедев А.Л., Смыслов А.Ю. Количественная оценка гомеостатической активности здоровых и больных людей. - М.: ГАРТ. 2013. 131 с.

2. Гавриков Б.М., Лебеденко И.М., Пестрякова Н.В., Ставицкий Р.В. Об одном статистическом методе оценивания состояния здоровья человека // Труды ИСА РАН, 2016. Т. 66. № 2. С. 54-59.

3. 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

4. Гавриков Б.М., Пестрякова Н.В., Ставицкий Р.В. О свойствах обучающих множеств // Информационные технологии и вычислительные системы. 2018. №4. С.97-107. DOI: 10.14357/207186321804010

5. 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

6. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. О способности статистического классификатора к обобщениям// Информационные технологии и вычислительные системы. 2021. № 4. С. 38-50. DOI: 10.14357/20718632210404.

7. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. О структуре базы обучения классификатора для оценивания состояния здоровья человека // Препринты ИПМ им. М.В.Келдыша. 2018. № 126. 18 с. DOI:10.20948/prepr-2018-126

8. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В., Ставицкий Р.В. Структура базы обучения статистического классификатора состояний систем

организма человека // Препринты ИПМ им. М.В.Келдыша. 2018. № 255. 40 с. DOI:10.20948/prepr-2018-255

9. Гавриков М.Б., Локуциевский О.В. Начала численного анализа. — М.: Янус, 1995.

10. Schürmann J. Pattern Classification. — New York: John Wiley&Sons, Inc., 1996.