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

Препринты ИПМ • Препринт № 8 за 2017 г.

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

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

Метод оценивания состояния здоровья человека

Рекомендуемая форма библиографической ссылки: Метод оценивания состояния здоровья человека / Б.М.Гавриков [и др.] // Препринты ИПМ им. М.В.Келдыша. 2017. № 8. 18 с. doi:10.20948/prepr-2017-8
URL: http://library.keldysh.ru/preprint.asp?id=2017-8

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

Метод оценивания состояния здоровья человека

Москва — 2017

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

Метод оценивания состояния здоровья человека

Целью настоящего исследования является разработка метода классификации состояний здоровья человека для каждой системы организма по четырем различным категориям на основании данных о параметрах периферической крови, полученных в результате лабораторного анализа. Используется вероятностный способ классификации, в основе которого лежит полиномиально-регрессионный подход. Для мужчин и женщин строятся и используются различные классификаторы.

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

Boris Mikhailovich Gavrikov, Mikhail Borisovich Gavrikov, Irina Matveevna Lebedenko, Nadejda Vladimirovna Pestryakova, Roman Vladimirovich Stavitskii

The method of estimating the state of human health

The purpose of this study is to develop a method of classifying states of human health for each system of the body to four specific categories, based on their parameters of peripheral blood, obtained as the result of laboratory analysis. There was used a probabilistic classification method, which is based on polynomial regression approach. Various classifiers are built and used for men and women.

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

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

Оглавление

Введение

Оптимизация диагностического процесса включает предварительный выбор направления исследований до начала проведения диагностических процедур. Необходимость этого обусловлена в первую очередь двумя причинами – улучшением качества диагностики и удешевлением длительных и многоступенчатых диагностических процедур за счёт заранее определённого направления действий.

Основным подходом к оптимизации является поиск некоторого принципа, общего для различных людей, подвергающихся диагностике, по которому принимается то или иное решение. В процессе поиска такого принципа нужно обработать огромное количество различных биофизических данных, относящихся ко множеству людей.

Отсюда вытекает необходимость развития методов предварительной оптимизации, основанных на искусственном интеллекте и технологиях распознавания образов, поскольку в настоящее время эти подходы являются наиболее эффективными при решении задач такого рода. Также эти методы можно использовать не только для оптимизации диагностики, но и для предсказания результатов лечения.

В наши дни непосредственное применение таких методов ограничивается преимущественно совершением выбора из многих возможных вариантов. Простейший выбор (да/нет) описан в статье [1]. В этой работе исследовались результаты выслушивания детских сердечных шумов с помощью электронного стетоскопа.

Раннее распознавание сердечных заболеваний является одной из важных задач педиатрии. Но усилия по разработке недорогого сканирующего устройства, способного помочь различить «невинные» и патологические сердечные шумы, имели ограниченный успех. Искусственные нейронные сети (ИНС) – это ценный инструмент, используемый в задачах распознавания и классификации сложных образов. Они «обучаются» сложным взаимозависимостям между входными и выходными данными и идентифицируют соотношения между ними, которые могут не быть заметны человеческому анализу. Целью этого исследования было научить ИНС эффективно различать «невинные» и патологические шумы.

С помощью электронного стетоскопа были записаны сердечные шумы 69 пациентов (37 патологических и 32 – нет). У каждого пациента шумы были записаны в двух позициях электронного стетоскопа, при этом каждая запись состояла из примерно 8 сердечных циклов. Из этих двух записей впоследствии был отобран, обработан и загружен в ИНС один звуковой образец из 3 циклов. С помощью анализа Фурье был получен нормализованный энергетический спектр звуковых данных. Были испробованы варианты с различным спектральным разрешением (1, 3 и 5 Гц) и различными диапазонами частот (0 – 90, 0 – 150, 0 – 210, 0 – 255 и 0 – 300 Гц) на входе в ИНС.

При обучении ИНС использовался метод Джека-Найфинга (Jack-Knifing), представляющий собой итеративный процесс, в котором каждый элемент исходной базы используется для проверки работы всей системы. Классификатор обучен на остальных данных и тестируется с помощью этой одной точки. Тем самым обеспечивается независимая проверка, так как классификатор не видит эту точку при обучении. Один за другим, каждый доступный пример берётся для тестирования. Такой подход проверяет весь метод классификации, а не какой-нибудь отдельный классификатор. Таким образом, каждый из 69 наборов данных был взят для тестирования, что привело к созданию для каждого набора входных данных (с уникальной комбинацией спектрального разрешения и диапазона частот) 69 раздельно обученных ИНС.

При спектральном разрешении 1 Гц и спектре от 0 до 210 Гц с помощью разработанной системы классификации удавалось добиться чувствительности и специфичности, равных 100%. Как правило, увеличение разрешения и расширение спектра приводили к улучшению результатов. Тенденция улучшения точности классификатора при более широком энергетическом спектре может быть объяснена тем, что с уширением спектра у классификатора появляется больше рабочей информации. Тенденция к улучшению производительности с лучшим разрешением преобразования Фурье была подтверждена путём уменьшения разрешения до 5 Гц. При таком разрешении не удавалось добиться точности, сопоставимой с уровнями 1 Гц или 3 Гц.

Однако увеличенная вычислительная нагрузка ИНС может привести и к обратному результату. Широко известно, что схема распознавания, которая может обрабатывать больший набор признаков, работает лучше, если этот набор не содержит избыточной информации.

Хотя все 69 образцов были применены для проверки, в этой процедуре каждый раз использовался только один элемент с получающимся в итоге набором из 69 ИНС. Возникает вопрос: может ли этот набор быть эффективным на практике? Для того чтобы это проверить, имеющиеся данные были разделены на два набора. Набор из 54 примеров был обозначен как учебный для разработки 54 ИНС по схеме Джека-Найфинга, а набор из 15 элементов был зарезервирован как набор для проверки работы получившейся системы в реальных условиях. При этом решение принималось путём простого голосования ИНС. В результате система из 54 ИНС смогла правильно распознать 7 из 9 патологических примеров и 5 из 6 «невинных». Оба патологических примера, распознанных неправильно, относились к заболеваниям, мало представленным в обучающих данных. Когда эти неправильно классифицированные патологические и нормальный примеры были переведены в учебный набор данных, мог быть установлен порог, по отношению к которому результирующая классифицирующая система из 57 ИНС была полностью точной при проверке на оставшихся 12 примерах.

Таким образом, был продемонстрирован значительный потенциал ИНС при использовании их в качестве точного диагностического инструмента для

классификации сердечных шумов. Основная проблема, стоящая перед исследователями этой и подобных тем, – сбор достаточного количества данных. Ожидается, что в конечном итоге эта работа приведёт к созданию автоматического сканирующего устройства с дополнительной возможностью предсказания состояний сердца.

Аналогичный подход применялся и в работе [2], где с помощью искусственных нейронных сетей исследовались выживаемость пациентов, получающих хирургическое лечение по поводу рака пищевода.

Всего в исследовании приняло участие 418 человек. При этом в качестве входных данных использовался ряд факторов, относящихся как к пациенту, так и к опухоли: 5 демографических факторов, 6 клинических, 8 параклинических, 19 факторов, соответствующих сопутствующим заболеваниям, 34 категории первичного и сопутствующего лечения, 15 патологических особенностей, 7 параметров, относящихся к стадированию и 41 биологический или генетический параметр.

Аналогично предыдущему случаю, на выходе результат был равен либо 0 (пациент мёртв), либо 1 (пациент жив) для временных интервалов в 1 год и в 5 лет. Используемые при этом нейронные сети состояли из 1 входного слоя из 199 нейронов, 1 или 2 скрытых слоёв и 1 выходного слоя (1 нейрон). Для уменьшения структурного смещения ИНС было обучено 14 разных структур (с 1 скрытым слоем: n-2-1, n-3-1, n-4-1, n-5-1 и n-6-1; с 2 скрытыми слоями: n-2-2-1, n-2-4-1, n-2-6-1, n-4-2-1, n-4-4-1, n-4-6-1, n-6-2-1, n-6-4-1, и n-6-6-1; n – количество входных переменных).

В начале обучения исследуемая популяция была разделена на 3 группы: учебный набор (53% пациентов), набор для подтверждения (27%) и тестовый набор (20%). В течение обучения величина ошибки набора подтверждения отслеживалась для того, чтобы избежать переобучения и определить оптимальное время для прекращения обучения. Для минимизации переобучения использовалось несколько математических методов: обратное распространение с импульсом, регуляризация затухания веса (фактор затухания = 0,01; фактор масштабирования = 1), добавление гауссовского шума (отклонение = 0,3) и случайное смешивание порядка представления пациентов.

Предполагалось, что будет существовать оптимальное количество информативных переменных в наборе данных, которые приведут к максимальной точности выхода. Любое дальнейшее уменьшение переменных привело бы тогда к уменьшению точности. Для 1-летней выживаемости максимальная точность предсказаний (ТП) составила 0,883 и была получена с набором данных из 65 переменных. Чувствительность и специфичность составили 78,1% и 84,7% соответственно. Максимальная ТП для 5-летней выживаемости составила 0,884 для набора данных из 60 переменных. При этом чувствительность и специфичность составили 80,7% и 86,5% соответственно.

В любом анализе с применением ИНС могут представлять опасность переобучение и неустойчивость модели.

Для устранения неустойчивости текущей модели был сформирован «предсказательный комитет» из множества ИНС. Оказалось, что ТП малых комитетов (<40 ИНС) для всех структур была неустойчивой и относительно низкой, но ТП больших комитетов (>40 ИНС) стабилизировались.

Очевидно, что бинарный (да/нет) выход сильно уменьшает возможности решения многих задач. В работе [3] на выходе нейронной сети получается 5 результатов. Авторы решали задачу определения того, относится ли данный образец ткани к одному из типов опухолей (нейробластома, рабдомиосаркома, лимфома Бёркитта или опухоль семейства Юинга) или нет, а если относится, то к какому.

В качестве материала для анализа использовался генетический материал комплементарной ДНК. В остальном методы, применяемые авторами данной статьи, схожи с методами предыдущих авторов, например, использовалось определение вклада каждого гена в классификацию и, по сути, ранжирование генов, а также разделение исходной выборки на обучающую и тестовую группы.

Оценивание состояния здоровья человека по анализу крови

Поскольку организм человека является сложной биологической системой, то разработка методологии оценивания его состояния – большая проблема, решением которой занимались многие поколения врачей. В медицине за долгие века ее существования и развития, несомненно, был накоплен и систематизирован огромный фактический материал по диагностике заболеваний. Однако зачастую при общении с пациентом врач полагается в основном на самого себя, и гарантом правильной постановки диагноза являются его квалификация и опыт. Высокий уровень технического оснащения современной медицины имеет существенное значение, но именно врач должен принять безошибочное решение относительно направления дальнейших исследований состояния различных органов и систем, а затем уж стратегии и тактики лечения пациента.

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

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

Показатели крови здорового человеческого организма варьируются в некоторых известных диапазонах (табл.1). Значительные отклонения от нормы

могут быть характерными проявлениями определенных заболеваний, выявить которые не составляет большого труда. Но во многих случаях ситуация неоднозначная, и врачу необходима помощь в принятии решения о состоянии той или иной системы организма.

Таблица 1

Средние показатели периферической крови здоровых мужчин и женщин России

% колебаний		Пол		Основные показатели крови
М	Ж	М	Ж	
6,5	12	4,3- 4,6	3,9-4,2	Эритроциты RBC 10^{12}1/л
1,5	7	148-149	130-138	Гемоглобин HGB г/л
7,5	11	9±2	8,0±1,5	Ретикулоциты RET до 20 % от RBC
10	15	198±16	214±28	Тромбоциты PLT 10^9 1/л
39		4,46-7,28		Лейкоциты WBC 10^9 1/л
21		48,6-63,4		Нейтрофилы NEUT 10^9 1/л
--		2,39-4,39		------//------- NEUT%%
44		1,34-2,38		Лимфоциты LIMPH 10^9 1/л
--		2,47-3,89		------//------- LIMPH %%
--		0,32-1,26		Базофилы BASO %%
88		0,035-0,3		Эозинофилы EOS 10^9 1/л
--		0-5,9		------//------- EOS %%
107		0,02-0,14		Палочкоядерные P 10^9 1/л
--		0,24-2,36		------//------- P %%
46		2,31-4,31		Сегметоядерные S10^9 1/л
--		48,6-63,4		------//------- S %%
42		0,3-0,52		Моноциты MONO 10^9 1/л
--		5,72-8,62		------//------- MONO %%
5,0		1,0-2,0		Гранулоциты GRAN 10^9 1/л

Академик РАН Роберт Нигматулин в интервью изданию «Аргументы недели», опубликованном 23.04.2015 под названием «Власть «послушных», сказал: «В наше время эффективных информационных систем нельзя управлять “на глазок”». То же самое можно сказать и о системе диагностики в условиях современных технологий.

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

Как отмечается в монографии [4], при оценке гомеостаза и его динамики (гомеостатической активности) в клинической практике применяется комплексный подход. А именно, производится сопоставление ряда измеряемых параметров (температуры тела, пульсации сердца, отдельных показателей периферической крови и др.). Этот способ оценивания крайне неточен, поскольку некоторые измеряемые величины нестабильны, они могут изменяться под влиянием внешних факторов (режима питания, физических нагрузок и т.д.). В качестве более объективного способа определения уровня гомеостаза и гомеостатической активности предлагается использовать не менее пяти показателей периферической крови.

Общая постановка задачи

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

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

По каждой системе организма (СО) – пищеварения, дыхания и пр. – проводится самостоятельное исследование СЗЧ посредством отдельного классификатора. При его построении используется обучающая выборка для рассматриваемой СО (табл.2).

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

1 класс – здоровые – 0–20%;

2 класс – начальные отклонения состояния здоровья – 21–40%;

3 класс – выраженное отклонение состояния здоровья – 41–70 %;

4 класс – тяжелое заболевание – 71–100%.

При обучении используются выборки, полученные в результате детального обследования пациентов большой группой специалистов из различных областей медицины. Для определенной СО из рассматриваемого перечня к каждой из четырех возможных градаций относится список заболеваний, соответствующих этой СО (табл.2). База показателей крови практически здоровых людей одинакова для всех СО.

Идея использовать при решении описанной задачи подход, основанный на полиномиальной регрессии [5-8], основывалась на том обстоятельстве, что данный метод хорошо зарекомендовал себя при распознавании столь сложных объектов, как печатные и рукопечатные символы. Он является точным, быстрым, генерирует монотонные (надежные) оценки, имеющие вероятностную природу. Представилась уникальная возможность адаптировать

этот подход для классификации объектов принципиально иного происхождения.

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

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

RBC [L^{-1}] – эритроциты,
HGB [g L^{-1}] – гемоглобин,
PLT [L^{-1}] – тромбоциты,
WBC [L^{-1}] – лейкоциты,
LIMPH [L^{-1}], [%] – лимфоциты,
GRAN [L^{-1}], [%] – гранулоциты
(GRAN = NEUT + EOS + BASO),
NEUT [L^{-1}], [%] – гранулофилы,
EOS [L^{-1}], [%] – эозинофилы,
BASO [L^{-1}], [%] – базофилы.

Таблица 2

Наименование основных систем организма и классов заболеваний

<u>Пищеварительная система</u>
Полость рта, глотка, пищевод, желудок, тонкая кишка, толстая кишка, поджелудочная железа.
1 класс – здоровые;
2 класс – гастрит, геморрой, аппендицит, грыжа пищевого отверстия, диафрагмы, эзофагит, хронический колит, гастроитоз, диспенсия, дуоденит, язвенный колит, эрозия желудка и двенадцатиперстной кишки;
3 класс – язва желудка, язва двенадцатиперстной кишки, панкреатит,

энтерит, неспецифический язвенный колит, желудочно-кишечное кровотечение, туберкулез; *4 класс* – онкологические заболевания.
<u>Органы дыхания</u> Гортань, легкие, трахея, бронхи, диафрагма. *1 класс* – здоровые; *2 класс* – ларингит, трахеит, бронхит, гайморит, плеврит; *3 класс* – бронхиальная астма, туберкулез, инфаркт легкого, пневмокониоз, пневмония, спонтанный пневмоторакс; *4 класс* – онкологические заболевания.
<u>Опорно-двигательный аппарат</u> Кости, связки, сухожилия, суставы, мышцы, фасции. *1 класс* – здоровые; *2 класс* – артрозы, бурситы, тендовагиниты, вывихи, растяжения, радикулит, остеохондроз, артрит, миозиты; *3 класс* – подагра, ревматизм, миопатия, ревматоидный артрит, туберкулез; *4 класс* – онкологические заболевания.
<u>Урологическая система</u> Мочевыделительная система, половые органы. *1 класс* – здоровые; *2 класс* – цистит, простатит, водянка яичка, аденома предстательной железы, аднексит, бартолинит; *3 класс* – пиелонефрит, макрогематурия, почечная колика, мочекаменная болезнь, амлоидоз, эндометрит, хроническая почечная недостаточность, киста почек, туберкулез; *4 класс* – онкологические заболевания.
<u>Гинекологическая система</u> *1 класс* – здоровые; *2 класс* – эрозия шейки матки, гонорея; *3 класс* – маточное кровотечение миомы, крауроз вульвы, туберкулез; *4 класс* – онкологические заболевания.
<u>Эндокринная система</u> Щитовидная, поджелудочная железы, надпочечники, гипофиз. *1 класс* – здоровые; *2 класс* – зоб, гипертиреоз, сахарный диабет 2-го типа, несахарный диабет, ожирение, аллергия; *3 класс* – акромегалия, тиреотоксикоз, сахарный диабет 1-го типа, гипсерпаратериоз, микседема, болезнь Иценко-Кушинга, надпочечная недостаточность, туберкулез; *4 класс* – онкологические заболевания.
<u>ЦНС, органы чувствительности</u> Головной мозг, спинной мозг, зрение, обоняние, вкусовые железы,

осязание, периферические нервы. *1 класс* – здоровые; *2 класс* – невралгии, невриты, дистония, воспаление среднего уха, тугоухость, хронический отит; *3 класс* – энцефалиты, менингиты, инсульт, миопатии наследственные, туберкулез; *4 класс* – онкологические заболевания.
Грудные железы (мужские и женские) *1 класс* – здоровые; *2 класс* – фиброаденомы, мастопатии, мастодении; *3 класс* – острый мастит, туберкулез; *4 класс* – онкологические заболевания.
Печень и желчевыводящие пути *1 класс* – здоровые; *2 класс* – дискинезия желчевыводящих путей, жировой гепатоз, синдром Жильбера, ротора, Дубина–Джонсона; *3 класс* – гепатиты, холецистит, желчекаменная болезнь, абсцесс печени, холангиты, механическая желтуха, туберкулез; *4 класс* – онкологические заболевания.
Общее состояние всего организма

В отношении обучающего множества следует заметить, что для печатных и рукопечатных символов имелось количественное расхождение. А именно, выборка рукопечатных цифр в сотни раз превосходила по объему множество печатных цифр. Это было вполне обоснованным, поскольку рукопечатное написание более вариативное, чем печатное. В то же время в каждой из этих выборок доли изображений различных символов были достаточно соразмерными и отличались не очень существенно.

Однако ввиду объективных трудностей составления обучающих выборок анализов крови принцип соразмерности не выполняется в необходимой степени. Для ряда СО некоторые из градаций СЗЧ недостаточно заполнены или вообще отсутствуют. Градация, соответствующая здоровым людям (класс «1»), наиболее обширная и ввиду универсальности используется для всех СО.

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

Еще одна особенность заключается в том, что имеется очевидный изначальный порядок расположения градаций СЗЧ от здорового до максимальной степени поражения организма. Однако при решении

поставленной задачи классификации в полученном перечне альтернатив он может нарушаться.

Метод оценивания СЗЧ и полученные результаты

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

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

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

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

Затем выполняем следующее преобразование:

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

Отождествляем k-й элемент множества градаций СЗЧ с базисным вектором $\mathbf{e}_k=(0\ldots1\ldots0)$ (здесь 1 находится на k-м месте, причем $1\le k\le K$) из $\mathbf{R}^K$. Обозначаем $Y=\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$.

Пусть можно найти $p_k(\mathbf{v})$ – вероятность того, что набор (отнормированных) показателей крови соответствует k–му элементу СЗЧ, где $1\le k\le K$. На выходе имеем искомый элемент СЗЧ с порядковым номером r, где

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

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

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

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

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

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

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

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

Значение A вычисляется приближенно в процессе обучения, используя содержащиеся в некоторой базе данных наборы пар векторов $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$ ($\mathbf{v}^{(j)}$ набор параметров крови, соответствующий элементу СЗЧ с каким-либо номером k ($1 \le k \le K$) и его базисный вектор $\mathbf{y}^{(j)} = (0\ldots1\ldots0)$, где 1 стоит на k-м месте, $1 \le j \le J$):

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

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

$$\begin{aligned} &A_j = A_{j-1} - \alpha_j G_j \mathbf{x}^{(j)}[A_{j-1}^{\mathrm{T}}\mathbf{x}^{(j)} - \mathbf{y}^{(j)}]^{\mathrm{T}}, \quad \alpha_j = 1/J, \\ &G_j = \frac{1}{1-\alpha_j}\Big[G_{j-1} - \alpha_j \frac{G_{j-1}\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathrm{T}}G_{j-1}}{1+\alpha_j((\mathbf{x}^{(j)})^{\mathrm{T}}G_{j-1}\mathbf{x}^{(j)} - 1)}\Big], \qquad 1 \le j \le J \\ &G_j \equiv D^{-1},\ D = \mathrm{diag}\,(\,\mathrm{E}\{x_1^2\},\mathrm{E}\{x_2^2\},\ldots,\mathrm{E}\{x_L^2\}). \end{aligned} \qquad (5)$$

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

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

$$\mathbf{x} = (1, \{v_i\},\{v_i^2\},\{v_i^3\},\{v_i^4\}\{v_i^5\},\{v_i^6\},\ \},\ \{v_i v_j\})\ \ 1 \le i \le 8,\ 1 \le j \le 8,\ i \ne j\ . \qquad (6)$$

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

Поскольку обучающие множества имели неравноценные по объему подмножества для различных градаций заболевания организма, то при получении обучающей последовательности использовалась лишь часть объема данных по классу «1» (здоровых людей), во многих случаях равная количеству

элементов класса «4» (наивысшая степень заболевания), далее добавлялись элементы классов «2» и «3», а затем проводилось перемешивание по всему объему обучающего множества. Этот прием, конечно, не позволил осуществить приблизительное выравнивание числа элементов для всех четырех классов по каждой СО. Однако вычислительная практика показала, что используемого количества элементов класса «1» достаточно для безошибочного распознавания всех элементов этого класса, в том числе не включенных в процесс обучения элементы класса «1».

Для количественной оценки качества классификации требуется ввести следующее понятие.

Точностью распознавания по базе *B* называется величина

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

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

С целью повышения точности распознавания проводилось многократное обучение на одной и той же базе с контролем точности распознавания, поскольку при неограниченном увеличении числа таких итераций точность сначала стабилизируется на некотором минимальном значении, а затем может начинать нарастать. В ряде случаев процесс имеет вид периодических колебаний с одним и тем же значением минимума количества ошибочных распознаваний, но с частично изменяющимся набором неправильно распознанных элементов, а для одинаковых элементов – с различным перечнем альтернатив.

В табл.3 приведены данные о том, как именно количественно распределены по классам элементы обучающих множеств для различных СО. Рассмотрен мужской контингент. Показаны полученные в результате расчетов параметры, относящиеся к качеству классификации. Сюда входит точность классификации на элементах обучающего множества, а также максимальная оценка ошибочного распознавания. Значение последней существенно меньше максимальной оценки (255). Это обстоятельство служит основанием для утверждения о надежности выставляемых оценок (то, что определялось понятием монотонности на множествах большого объема, например, рукопечатных цифрах).

Отметим, что для пищеварительной системы первоначально обучающая база содержала 109 элементов. Часть имеющихся данных по классу «1» не использовалась, чтобы уменьшить диспропорцию обучающего множества в

отношении преобладания доли элементов класса «1». Затем было добавлено 22 элемента этого класса с последующим перемешиванием. Набор нераспознанных элементов остался прежним.

Из табл. 3 видно, что по СО «Печень и желчевыводящие пути» нет элементов класса «3». Однако, несмотря на это, классификатор нормально обучается и обеспечивает высокую точность при распознавании обучающего множества. В наборах альтернатив распознавания этот класс отсутствует.

По ряду СО мало элементов классов «2» и (или) «3». В этих случаях наблюдается большая точность классификации. Наиболее проблемными в отношении точности классификации являются пищеварительная, урологическая и эндокринная системы. Они характеризуются высокой заполненностью классов «2» и «3». Для СО «Пищеварительная система» проведен следующий численный эксперимент. Из обучающего множества убирали 1-2 элемента класса «1». При этом набор нераспознанных элементов на первоначальном обучающем множестве не меняется. Но если обучающее множество уменьшали на элемент класса «2» или «3», то число ошибок увеличивалось.

Таблица 3

Распределение объема базы по классам основных систем организма и точность классификации

Пищеварительная система				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	17	26	33	109
Качество классификации				
количество ошибок		*доля ошибок*	*максимальная оценка ошибочного распознавания*	
17		15,6%	218	
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
55	17	26	33	131
Качество классификации				
количество ошибок		*доля ошибок*	*максимальная оценка ошибочного распознавания*	
17		13,0%	212	

Органы дыхания				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
32	11	12	21	76
ачество классификации				
количество ошибок		*доля ошибок*		*максимальная оценка ошибочного распознавания*
3		3,9%		159
Опорно-двигательный аппарат				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	3	7	33	76
Качество классификации				
количество ошибок		*доля ошибок*		*максимальная оценка ошибочного распознавания*
2		2,6%		174
Урологическая система				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
33	18	26	33	110
Качество классификации				
количество ошибок		*доля ошибок*		*максимальная оценка ошибочного распознавания*
14		12,7%		155
Эндокринная система				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
34	26	27	34	121
Качество классификации				
количество ошибок		*доля ошибок*		*максимальная оценка ошибочного распознавания*
16		13,2%		176

ЦНС, органы чувствительности				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
26	2	7	26	61
Качество классификации				
количество ошибок		*доля ошибок*	*максимальная оценка ошибочного распознавания*	
1		1,6%	147	
Печень и желчевыводящие пути				
Распределение объема базы по классам				
«1»	«2»	«3»	«4»	«1-2-3-4»
31	12	0	31	74
Качество классификации				
количество ошибок		*доля ошибок*	*максимальная оценка ошибочного распознавания*	
1		1,4%	180	

Тестирование классификатора проводилось на той же базе, которая использовалась для обучения. Достигнутая точность распознавания для различных рассматриваемых СО находится в диапазоне 84,4 – 98,6 %. Эти результаты соответствуют данным, полученным при помощи алгебраического подхода Журавлева [4].

Заключение

Рассмотрены методы предварительной оптимизации лечебных и диагностических процессов. Показана перспективность и осуществимость на практике точных методов в применении к различным клиническим случаям – сердечным шумам, раку пищевода и некоторым детским опухолям.

Для разработанного авторами статистического метода распознавания на основе полиномиальной регрессии реализовано приложение в качестве классификатора СЗЧ по показателям периферической крови из пальца для различных СО.

Подтверждена возможность применения данного метода для объектов, природа которых принципиально отлична от изображений печатных и рукопечатных символов, а структура обучающего множества значительно более сложная, чем в случае изображений символов.

Библиографический список

1. Artificial neural network-based method of screening heart murmurs in children / DeGroff C.G. [et al.]. // Circulation. 2001. №103. Pp. 2711-2716.

2. Prediction of survival in patients with esophageal carcinoma using artificial neural networks / Sato F. [et al.]. // Cancer. 2005. №103. Pp. 1596–1605.

3. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks / Khan J. [et al.]. // Nature Med. 2001. № 403. Pp. 673–679.

4. Количественная оценка гомеостатической активности здоровых и больных людей / Ставицкий Р.В. [и др.]. // М.: ГАРТ. 2013. 131 с.

5. Гавриков М.Б., Пестрякова Н.В. Метод полиномиальной регрессии в задачах распознавания печатных и рукопечатных символов. // Препринты ИПМ им.М.В.Келдыша. 2004. № 22. 12 с.

6. Об одном методе распознавания символов, основанном на полиномиальной регрессии / Гавриков М.Б. [и др.]. // Автоматика и Телемеханика. 2006. № 2. С. 119-134.

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

8. Об одном статистическом методе оценивания состояния здоровья человека / Гавриков Б.М. [и др.]. // Труды ИСА РАН. 2016. Т.66. Вып.2. С. 54-59.