О теоретико-возможностных методах решения задач медицинской диагностики

В. А. Газарян, Ю. М. Нагорный,
Ю. П. Пытьев, А. К. Шаховская

Показано, что для решения многих задач медицинской диагностики естественно применение не вероятностной, а возможностной модели анализа и интерпретации данных [1, 2], полученных как с использованием современных медицинских технологий, так и отражающих самочувствие и состояние больного, профессиональный опыт и интуицию врача. Построена возможностная модель симптоматики заболевания, характеризующая нечёткую связь между симптоматийным описанием состояния больного и реальным его состоянием, в которой диагностические критерии определяются группами значений признаков (симптомов), ранжированных по значениям их возможностей при данном заболевании. В возможностной модели диагностики оптимальное решающее правило минимизирует возможность или (и) неизбежность потерь при постановке диагноза.

Введение

При моделировании медицинских объектов приходится на практике сталкиваться с нечёткостью и неопределённостью их описания, связанной с неполнотой и недостоверностью знаний, случайностью и неточностью данных, субъективным подходом к их описанию. Недостаточная эффективность вероятностных методов моделирования подобных объектов связана прежде всего с невозможностью эмпирического построения их стохастических моделей, обусловленной как изменчивостью во времени характеристик самочувствия больного и симптомов заболевания, так и неформализованным и во многих случаях субъективным характером признаков заболевания.

В работе [1] показано, что в то время как вероятностную модель стохастического объекта, непредсказуемо эволюционирующего во времени, эмпирически построить невозможно, его возможностная модель при известных ограничениях на характер эволюции вероятностной модели может быть восстановлена, причём точно и на основании почти наверное конечного числа наблюдений. Если к этому добавить, что возможностная модель и нестохастического объекта может быть построена эмпирически на основе заключений экспертов, то предпочтительность возможностного моделирования медицинской диагностики, в которой используются как «чёткие» данные, полученные в результате применения современных медицинских технологий, так и неформализованные данные, отражающие субъективную самооценку, интуицию и опыт врача, становится очевидной.

1. Возможностная модель медицинской диагностики

Задача медицинской диагностики ставится как задача классификации, в которой требуется принять решение о принадлежности объекта — больного — к одному из M классов заболеваний. Состояние больного оценивается врачом по результатам проведённых исследований, анализов, а также тестирования, при котором больной отвечает на вопросы о наличии и степени выраженности различных симптомов. В рамках задачи классификации будем считать симптомы признаками, описывающими объект. Каждый признак может принимать различные значения. Например, признак «число лейкоцитов» принимает количественные значения, признак «утомляемость» — значения в ранговой шкале, выражающие степень утомляемости.

На этапе обучения алгоритма возможностной классификации, рекомендующего тот или иной диагноз заболевания, следует найти такие значения признаков, которые являются характерными для каждого класса. По сути это диагностические критерии заболеваний.

Во многих случаях при наличии взаимозависимых данных, обладающих естественной изменчивостью и нечёткостью, возникает неоднозначность трактовки симптоматийной картины заболевания разными специалистами, и клиническая оценка заболевания сопряжена с трудностями. Такая проблема характерна для диагностики функциональных расстройств в системе пищеварения, в частности, синдрома раздражённой толстой кишки (СРТК) [3]. Из-за отсутствия органических патологий выявление чётких диагностических критериев заболевания затруднено, и диагноз зачастую ставят методом исключения. Необходимо уточнить диагностические критерии, формализовать и проанализировать симптоматику заболевания.

Как было отмечено, при неформализованном характере признаков заболевания и ограниченном размере обучающих выборок вполне естественны возможностные методы обучения и распознавания. Качество стохастических методов медицинской диагностики [4] определяется вероятностью потерь, сопутствующих ошибочному диагнозу. В возможностной модели медицинской диагностики оптимальным является такое правило постановки диагноза, которое минимизирует возможность сопутствующих ему потерь. В [1] показано, что для восстановления возможностной модели классификации, в данном случае — диагностики, достаточно восстановить только переходные возможности постановки того или иного диагноза при зарегистрированном у больного наборе симптомов, в то время как для восстановления вероятностной модели классификации достаточно восстановить соответствующие переходные вероятности. Во введении была обоснована предпочтительность возможностного моделирования медицинского объекта по сравнению с вероятностным. Алгоритм, осуществляющий постановку диагноза, носит при возможностном моделировании «советующий» характер, оценивая последствия разных вариантов ошибочной диагностики. Эти ошибки могут иметь различные последствия для больного: одни при зачислении здорового человека в класс патологии и совсем иные при отнесении субъекта с серьёзным заболеванием в класс здоровых. Руководствуясь результатами проведённых медицинских исследований и компьютерной диагностики, врач ставит пациенту окончательный диагноз и назначает лечение.

При решении задачи медицинской диагностики недостаточно руководствоваться лишь построенной моделью медицинского объекта, в данной задаче — моделью состояния больного, характеризующейся распределением возможностей пары нечётких элементов — («наблюдение, как набор зарегистрированных признаков (симптомов)», «реальное состояние больного»). Главная задача компьютерной системы предварительной диагностики — вынесение решения о состоянии больного. Потери характеризуются индикаторной функцией $l_{k,d} \in [0,1]$, определяющей возможность покрытия точки (k,d) нечётким множеством Λ. Здесь $l_{k,d}$ интерпретируется как возможность потерь, обусловленных постановкой диагноза d, в то время как на самом деле больной страдает заболеванием k. Выносимое компьютерной системой решение о состоянии больного определяется как значение нечёткого элемента δ. Нечёткое правило решения о состоянии больного по набору зарегистрированных симптомов определяется распределением переходной возможности решения δ при наблюдающемся у больного наборе признаков (симптомов). Оптимальным является решающее правило, минимизирующее возможность потерь при постановке диагноза.

Совместные с клиникой НИИ питания РАМН исследования СРТК [4] привели к созданию специального опросника, включающего вопросы о причинах и давности заболевания, о самооценке состояния как органов пищеварения, так и психоэмоциональной сферы больного, что позволило достаточно полно охарактеризовать его состояние. Исходные данные для построения возможностной модели симптоматики заболевания суть сведения о больных, содержащиеся в ответах на вопросы такой анкеты. Анкета содержит 26 вопросов по гастроэнтерологии и психосоматике. Возможны от 2 до 4 вариантов ответов в зависимости от постановки вопроса. На каждый вопрос пациент дает ответ в ранговой шкале, отражающей степень тяжести соответствующего симптома.

1.1. Правило постановки диагноза, минимизирующее риск потерь

Сопоставим каждому субъекту (пациенту) n-мерный нечеткий вектор признаков $\chi = (\chi^1, \chi^2, \ldots, \chi^n)$, принимающий значение $x \in X$, где

$$x = (x^1, x^2, \ldots, x^n), \tag{1}$$

$x^j \in X^j$ — ответ пациента на j-й вопрос анкеты, рассматриваемый как значение j-го признака (симптома) его заболевания, $j = 1 \ldots n$, $X = X^1 \times X^2 \times \ldots \times X^n$, $X^j = \{x_1^j, x_2^j, \ldots, x_{n_j}^j\}$ — множество значений j-го признака, n_j — количество значений j-го признака, $j = 1 \ldots n$, n — число признаков.

Пусть æ — нечёткий элемент, значениями которого являются номера классов (заболеваний) $k \in \{1, \ldots, M\}$. Распределение $\varphi^{\chi,æ}(\cdot,\cdot)$ характеризует нечёткую связь между симптоматийным описанием состояния больного и реальным его состоянием и определяет возможностную модель состояния больного. $\varphi^{\chi,æ}(x,k)$ — возможность равенств $\chi = x$, æ $= k$, $x \in X$, $k \in \{1, \ldots, M\}$. $l_{k,d} \in [0,1]$ — возможность потерь при отнесении субъекта класса k к классу q, $q = 1, \ldots, M$. Далее не будем делать различий между субъектом и присущим ему вектором значений признаков $x = (x^1, x^2, \ldots, x^n)$.

В задаче диагностики требуется принять решение о принадлежности больного x к одному из M классов заболеваний. Решение о состоянии больного — рекомендуемый диагноз — определим как нечёткий элемент δ, принимающий значения на множестве $\{1, \ldots, M\}$. Распределение возможностей $\pi^{\delta|\chi}$ назовём нечётким правилом решения δ о состоянии больного, основанного на наблюдаемом у больного наборе признаков $\chi = (\chi^1, \chi^2, \ldots, \chi^n)$, или нечётким правилом постановки диагноза; значение $\pi^{\delta|\chi}(d|x)$ — возможность решения о заболевании $\delta = d$, когда $\chi = x$ — наблюдающиеся у больного симптомы.

В [1] рассмотрены методы идентификации состояний нечёткой системы, которая может находиться в одном из M состояний, полностью или частично определяющих распределение возможностей значений нечёткого элемента. Рассматривая нечеткий вектор признаков χ в качестве нечёткого элемента, составляющего основу для принятия решения о состоянии системы, в данном случае — о состоянии больного и классы заболеваний (диагнозы) — в качестве состояний системы, перейдём к решению задачи диагностики как задачи идентификации. В [1] дано решение задачи идентификации нечёткой системы, минимизирующее риск потерь. Применим его для решения задачи диагностики. Возможность потерь, определяющую качество правила постановки диагноза $\pi^{\delta|\chi}$, можно определить как

$$PL(\pi^{\delta|\chi}) = \sup_{x\in X,\, k\in\{1,\dots,M\},\, d\in\{1,\dots,M\}} \min(l_{k,d}, \pi^{\delta|\chi}(d|x), \varphi^{\chi,\text{æ}}(x,k)). \quad (2)$$

Поскольку качество решения тем выше, чем меньше возможность потерь, оптимальным естественно считать такое правило $\pi^{*\delta|\chi}$, которое минимизирует возможность потерь (2):

$$PL(\pi^{*\delta|\chi}) = \min_{\pi^{\delta|\chi}} PL(\pi^{\delta|\chi}). \quad (3)$$

Значение $PL(\pi^{*\delta|\chi})$ определяет риск потерь при оптимальном правиле постановки диагноза, рекомендующем диагноз $d = \delta^*(x)$:

$$\delta^*(x) \in \left\{ d \in \{1,\dots,M\},\ \pi^{*\delta|\chi}(d|x) = \max_{d'\in\{1,\dots,M\}} \pi^{*\delta|\chi}(d'|x) \right\}, \quad x \in X.$$

Возможность потерь, сопутствующих решению о постановке диагноза $\delta = d$ больному при наличии у него симптомов $\chi = x$, $x \in X$, $d \in \{1,\dots,M\}$, определим как

$$P_d(x) = \max_{1\leqslant k\leqslant M} \min(l_{k,d},\ \varphi^{\chi,\text{æ}}(x,k)). \quad (4)$$

Тогда согласно (2), (3), (4) задача определения оптимального нечёткого правила $\pi^{*\delta|\chi}$ постановки диагноза формулируется как следующая задача на минимум:

$$PL(\pi^{\delta|\chi}) = \sup_{x\in X} \max_{1\leqslant d\leqslant M} \min(\pi^{\delta|\chi}(d|x),\ P_d(x)) \sim \min_{\pi^{\delta|\chi}(\cdot|\cdot)}. \quad (5)$$

Её решение можно получить, решая для каждого вектора значений признаков $x \in X$ следующую задачу:

$$\max_{1 \leqslant d \leqslant M} \min(\pi^{\delta|\chi}(d|x),\ P_d(x)) \sim \min_{\pi^{\delta|\chi}(\cdot|x)} \tag{6}$$

Согласно условиям (5), (6), оптимальное правило $\pi^{*\delta|\chi}$ постановки диагноза предписывает при наличии у больного симптомов $\chi = x$ ставить диагноз $\delta = d$, возможность которого тем больше, чем меньше возможность потерь, обусловленных таким решением. В [1] сформулированы следующие достаточные условия оптимальности правила $\pi^{*\delta|\chi}$. Пусть

$$D^*(x) = \left\{ d \in \{1, \dots, M\},\ P_d(x) = \min_{d'} P_{d'}(x) \right\}, \quad x \in X. \tag{7}$$

Тогда любое нечёткое правило $\pi^{*\delta|\chi}$ постановки диагноза, удовлетворяющее условиям

$$\max_{d \in D^*(x)} \pi^{*\delta|\chi}(d|x) = 1, \ \ x \in X,$$
$$\pi^{*\delta|\chi}(d|x) = 0, \ \ d \in \{1, \dots, M\} \setminus D^*(x), \ \ x \in X, \tag{7*}$$

является решением задачи (5), (6). Поскольку согласно (7*) можно выбрать $\pi^{*\delta|\chi}(d|x) = 1$, если $d \in D^*(x)$ и $\pi^{*\delta|\chi}(d|x) = 0$, если $d \in \{1, \dots, M\} \setminus D^*(x)$, $x \in X$, то в качестве решения можно использовать значение $d^*(x)$ любой «чёткой» функции, удовлетворяющей условию $d^*(x) \in D^*(x)$, $x \in X$.

Если потери невозможны при правильной постановке диагноза: $l_{k,d} = 0$ при $d = k$, и возможность потерь максимальна при любой ошибочной постановке диагноза: $l_{k,d} = 1$ при $d \neq k$, $k, d = 1, \dots, M$, то в (4)

$$P_d(x) = \max_{k \neq d} \varphi^{\chi,æ}(x, k), \ \ d = 1, \dots, M, \ \ x \in X,$$

и минимум $P_d(x)$ при фиксированном x достигается на тех $d \in \{1, \dots, M\}$, при которых $\varphi^{\chi,æ}(x, k)$ достигает максимума. Тогда (7) имеет вид

$$D^*(x) = \left\{ d \in \{1, \dots, M\},\ \ \varphi^{\chi,æ}(x, d) = \max_{1 \leqslant k' \leqslant M} \varphi^{\chi,æ}(x, k') \right\}, \quad x \in X.$$

и правило $\pi^{*\delta|\chi}$, минимизирующее возможность потерь (в данном случае ошибки) при постановке диагноза, является правилом максимальной возможности [3].

В [1] показано, что правило идентификации d^*, минимизирующее возможность потерь, можно также определить как любую функцию $d^*(\cdot)$, $X \to \{1, \dots, M\}$, удовлетворяющую условию

$$d^*(x) \in D^*(x) = \left\{ d \in \{1, \dots, M\},\ P_d(|x) = \min_{d' \in \{1, \dots, M\}} P_{d'}(|x) \right\}, \quad x \in X, \tag{8}$$

где $P_d(|x) = \max\limits_{1 \leqslant k \leqslant M} \min(l_{k,d}, \varphi^{æ|\chi}(k|x))$ — условная при условии $\chi = x$ возможность потерь, и возможность потерь при постановке диагноза d^* определяется как

$$PL(d^*) = \sup_{x \in X} \min(P_{d^*(x)}(|x),\ \varphi^{\chi}(x)), \tag{9}$$

где $\varphi^{\chi}(\cdot) : X \to [0, 1]$ — распределение возможностей нечёткого вектора признаков (симптомов), $\varphi^{\chi}(x)$ — возможность наличия у больного симптомов $\chi = x$, $x \in X$,

$$P_{d^*(x)}(|x) = \max_{1 \leqslant k \leqslant M} \min(l_{k,d^*(x)}, \varphi^{æ|\chi}(k|x)). \tag{10}$$

Таким образом, для вычисления возможности потерь (9) необходимо распределение φ^{χ} возможностей симптомов χ. В задачах медицинской диагностики распределение $\varphi^{\chi,æ}$, как правило, неизвестно, но целью является постановка нечёткого диагноза d^*, а для этого, согласно (8) и (10), достаточно восстановить только переходную возможность $\varphi^{æ|\chi}$ (возможности потерь $l_{k,d}$ задаются врачом).

Возможностная модель $\varphi^{\chi,æ}(\cdot, \cdot)$ состояния больного определяется априорным распределением возможностей $\varphi^{æ}(\cdot) : \{1, \dots, M\} \to [0, 1]$ заболеваний и распределением переходной возможности $\varphi^{\chi|æ}(\cdot|k) : X \to [0, 1]$ для каждого заболевания $k \in \{1, \dots, M\}$. Тогда

$$\varphi^{\chi,æ}(x, k) = \min(\varphi^{\chi|æ}(x|k),\ \varphi^{æ}(k)), \quad x \in X,\ k \in \{1, \dots, M\}, \tag{11}$$

— совместное распределение $\chi, æ$. Предполагая, согласно мнению врачей, что ни один из классов заболеваний не является априори более «предпочтительным», чем другие, получаем равенство априорных возможностей $\varphi^{æ}(k) = 1$, $k = 1, \ldots, M$. В этом случае возможность потерь (4), сопутствующих решению о постановке диагноза $\delta = d$ больному при наличии у него симптомов $\chi = x$, $x \in X$, $d \in \{1, \ldots, M\}$, определяется как

$$P_d(x) = \max_{1 \leqslant k \leqslant M} \min(l_{k,d}, \varphi^{\chi|æ}(x|k)) \tag{12}$$

и для построения оптимального правила постановки диагноза d^* (7*) следует восстановить распределение переходных возможностей $\varphi^{\chi|æ}(x|k)$, $k = 1, \ldots, M$. Рассмотрим алгоритм гранулирования пространства значений признаков заболевания, осуществляющий стохастическое моделирование переходных возможностей определённых наборов признаков при условии принадлежности субъекта выделенным классам на основании обучающей выборки — множества больных с верифицированным диагнозом.

2. Алгоритм гранулирования пространства значений признаков заболевания

Предположим, что векторы признаков имеют стохастическую природу. Пусть $pr(x) = \Pr(\chi^1 = x^1, \chi^2 = x^2, \ldots, \chi^n = x^n)$ — вероятность того, что субъекту свойственен вектор значений симптомов

$$x = (x^1, x^2, \ldots, x^n), \tag{13}$$

$p(x) = P(\chi^1 = x^1, \chi^2 = x^2, \ldots, \chi^n = x^n)$ — значение возможности равенств $\chi^j = x^j$, $j = 1, \ldots, n$. Экспериментальной оценкой $pr(x)$ согласно закону больших чисел является частота встречаемости вектора признаков.

Для уточнения диагностических критериев заболевания необходимо найти группы признаков, ранжированные по значениям их переходных возможностей при данном заболевании. Тем самым на этапе обучения будут найдены переходные возможности $\varphi^{\chi|æ}(x|k)$ всех классов $k = 1, \ldots, M$ и решена задача диагностики (7*).

Больные СРТК были разделены гастроэнтерологами на две группы (два класса): с замедленной моторикой (№ 1) и ускоренной моторикой (№ 2) [4] для выявления значимых различий в описании симптоматики больных этих групп и уточнения диагностических критериев заболевания.

Пусть $(\Omega, \mathcal{P}(\Omega), P)$ — возможностное пространство [5]: $\Omega = \prod_{i=1}^{n} X^j$. Элементарные события этого пространства — векторы, содержащие все возможные комбинации значений признаков x^j, $j = 1 \ldots n$. В [2] применён простой алгоритм разбиения множества значений признаков на группы значений признаков w_i^T, $i = 1, \ldots, N$, (N — число найденных групп признаков при данном разбиении T), ранжированных по значениям их переходных возможностей $p_i^T = p(w_i^T|k)$ при условии принадлежности субъекта, обладающего набором признаков w_i^T, к классу k, $k = 1, \ldots, M$:

$$p_1^T > p_2^T > \ldots > p_N^T. \tag{14}$$

Так же упорядочены апостериорные возможности $p(k|w_i^T)$ заболевания k при наличии симптомов w_i^T [2]. Наиболее существенными признаками заболевания k являются наборы $w_1^{T_1}, w_1^{T_2}, \ldots$ разбиений $T_1, T_2, \ldots$, имеющие максимальную возможность $p_1^{T_i} = p(w_1^{T_i}|k) = 1$.

Преобразование возможностного пространства $(\Omega, \mathcal{P}(\Omega), P)$ в возможностное пространство $(\Omega^T, \mathcal{P}^T(\Omega), P^T)$ с элементарными событиями w_i^T, $i = 1, \ldots, N$, возможности которых в пространстве $(\Omega^T, \mathcal{P}(\Omega^T), P^T)$ удовлетворяют строгим неравенствам (14), названо в [1] гранулированием. Перед началом процедуры гранулирования пронумеруем субъекты (векторы значений симптомов) в порядке убывания (невозрастания) вероятностей:

$$pr_1 \geqslant pr_2 \geqslant \ldots \geqslant pr_l \geqslant \ldots, \tag{15}$$

где $pr_l = \Pr((x)_l)$, $(x)_l \in \Omega$ — вектор значений признаков (13) l-ого субъекта, $l = 1, 2, \ldots$

Существуют различные способы гранулирования. Возможностное пространство $(\Omega^T, \mathcal{P}(\Omega^T), P^T)$ является максимально согласованным с вероятностным пространством $(\Omega^T, \mathcal{P}^T(\Omega), \Pr^T)$ и неравенства

$$p_1^T > p_2^T > \ldots > p_n^T > \ldots, \tag{16}$$

$p_n^T = P(w_n^T)$, выполняются при

$$pr_n^T > 1 - pr_1^T - pr_2^T - \ldots - pr_n^T, \tag{17}$$

что аналогично неравенствам

$$pr_n^T > (1 - pr_1^T - pr_2^T - \ldots - pr_{n-1}^T)/2 \tag{18}$$

Более того, в [1] показано, что неравенства (17) и (18) являются необходимыми и достаточными условиями (16). Исходя из этого и из условия (15), определим первый набор значений признаков w_1^T, имеющий максимальную возможность $1 = p_1^T = P(w_1^T)$, для которого $p_1^T > p_2^T$, если $pr_1^T > 1/2$. Получим $w_1^T = \{(x)_1, \ldots, (x)_{i_1}\}$, где $i_1 = \min\{i \,|\, pr_1 + \ldots + pr_i > 1/2\}$. Далее следуют наборы $w_2^T, \ldots, w_{k+1}^T, \ldots$, упорядоченные по значениям возможностей согласно условию (18):

$$\begin{aligned} w_2^T &= \{(x)_{i_1+1}, \ldots, (x)_{i_2}\}, \ldots, \\ w_{k+1}^T &= \{(x)_{i_k+1}, \ldots, (x)_{i_{k+1}}\}, \ldots, \text{ причём} \\ i_2 &= \min\{i \,|\, pr_{i_1+1} + \ldots + pr_i > (1 - pr_1 - \ldots - pr_{i_1})/2\}, \ldots, \\ i_{k+1} &= \min\{i \,|\, pr_{i_k+1} + \ldots + pr_i > (1 - pr_1 - \ldots - pr_{i_k})/2\}. \end{aligned}$$

Элементарные события — векторы признаков $(x)_{i_k+1}, \ldots, (x)_{i_{k+1}}$, содержащиеся в w_{k+1}^T, $k = 0, 1, \ldots$, $i_0 = 0$, являются неразличимыми с точки зрения значимости при данном заболевании. Разбиение Ω, выполненное таким образом, называется гранулированием первого приближения [1]. Выделенные гранулы $w_1^T, w_2^T, \ldots, w_{k+1}^T, \ldots$ обозначим соответственно $w_1^{(1)}, w_2^{(1)}, \ldots, w_{k+1}^{(1)}, \ldots$ Такое гранулирование не является оптимальным с точки зрения выявления значимых закономерностей заболевания. Ранжированных по возможности наборов признаков может оказаться недостаточно для создания информативных диагностических критериев. Поэтому на наборы $w_1^T, w_2^T, \ldots, w_{k+1}^T, \ldots$ также накладываются следующие условия: число элементарных событий, содержащихся в них, должно быть минимальным; вероятности $pr_1^T, pr_2^T, \ldots, pr_n^T, \ldots$ также должны быть минимальными, $pr_i^T = pr(w_i^T)$. При выполнении данных требований число ранжированных по значимости наборов признаков будет максимальным.

Алгоритм гранулирования Ω, удовлетворяющий таким условиям, назовём оптимизированным. Как и в предыдущем алгоритме, найдём $i_1 = \min\{i \,|\, pr_1 + \ldots + pr_i > 1/2\}$. Если $i_1 = 1$, то первый набор в любом приближении будет $w_1^{(\cdot)} = w_1^{(1)} = \{(x)_1\}$, если же $i_1 > 1$, то найдём $k_1 = \max\{k \,|\, pr_1 + \ldots + pr_{i_1-1} + pr_{i_1-1+k} > 1/2\}$ и определим первый набор второго приближения: $w_1^{(2)} = \{(x)_1, \ldots, (x)_{i_1-1}, (x)_{i_1-1+k_1}\}$. В наборе $w_1^{(2)}$ содержится столько же элементарных событий, сколько и в первом наборе первого приближения $w_1^{(1)}$, а вероятность $pr_1^{(2)} = pr(w_1^{(2)})$ меньше вероятности $pr_1^{(1)} = pr(w_1^{(1)})$.

Теперь перейдём к построению первого набора третьего приближения $w_1^{(3)}$. Рассматривается случай $i_1 > 1$. Если при этом $k_1 = 1$, первый набор в любом приближении будет $w_1^{(\cdot)} = w_1^{(2)}$, если же $k_1 > 1$, то найдём $k_2 = \max\{k \,|\, pr_1 + \ldots + pr_{i_1-1+k_1-1} + pr_{i_1-1+k_1-1+k} > 1/2\}$ и определим $w_1^{(3)} = \{(x)_1, \ldots, (x)_{i_1-1}, (x)_{i_1-1+k_1-1}, (x)_{i_1-1+k_1-1+k_2}\}$. В наборе $w_1^{(3)}$ содержится столько же элементарных событий, сколько и в $w_1^{(2)}$, при этом $pr_1^{(3)} < pr_1^{(2)}$, $pr_1^{(3)} = pr(w_1^{(3)})$, $pr_1^{(2)} = pr(w_1^{(2)})$. Аналогично строятся наборы $w_1^{(4)}, w_1^{(5)}, \ldots$ любого приближения, имеющие максимальную возможность.

Рассмотрим теперь алгоритм построения второго по степени значимости набора, взяв за основу, например, первый набор второго приближения $w_1^{(2)}$. Для того, чтобы первый и второй наборы были различимы по возможности: $1 = p_1^{(2)} > p_2^{(2)}$, согласно (18) необходимо выполнение условия $pr_1^{(2)} > 1/2$, которое накладывается на вероятность только первого набора. Для формирования набора $w_2^{(2)}$ определим новое пространство элементарных событий $\Omega \setminus w_1^{(2)}$, перенумеруем заново его точки, сохранив их упорядоченность по вероятностям, принятую в (15). Тогда в случае $i_1 > 1$ $\Omega \setminus w_1^{(2)} = \{(x)_{i_1}, \ldots, (x)_{i_1-1+k_1-1}, (x)_{i_1-1+k_1+1}, \ldots\} = \{(x)_1^{(2)}, (x)_2^{(2)}, \ldots\}$. Определим вероятности событий $\{(x)_1^{(2)}\}, \{(x)_2^{(2)}\}, \ldots$ как условные при условии $\Omega \setminus w_1^{(2)}$ и, учитывая, что $(x)_k^{(2)} \in \Omega \setminus w_1^{(2)}$, получим: $pr_k^{(2)} = \Pr(\{(x)_k^{(2)}\} \,|\, \Omega \setminus w_1^{(2)}) = \dfrac{\Pr(\{(x)_k^{(2)}\})}{\Pr(\Omega \setminus w_1^{(2)})}$, $k = 1, 2, \ldots$ Теперь второй набор второго приближения $w_2^{(2)}$ определится как подмножество $\Omega \setminus w_1^{(2)}$ точно так же, как набор $w_1^{(2)}$ был определён как подмножество Ω:

$$w_2^{(2)} = \{(x)_1^{(2)}, (x)_2^{(2)}, \ldots, (x)_{i-1}^{(2)}, (x)_{i-1+k_1}^{(2)}\}, \tag{19}$$

где

$$i_1 = \min\{i \,|\, pr_1^{(2)} + \ldots + pr_i^{(2)} > 1/2\} \tag{20}$$

и если $i_1 = 1$, $w_2^{(2)} = \{(x)_1^{(2)}\}$, если же $i_1 > 1$, то

$$k_1 = \max\{k \,|\, pr_1^{(2)} + \ldots + pr_{i_1-1}^{(2)} + pr_{i_1-1+k}^{(2)} > 1/2\}. \tag{21}$$

Тогда второй набор второго приближения обладает большей значимостью, чем третий (ещё не сформированный): $p_2^{(2)} > p_3^{(2)}$, так как $\Pr(\{w_2^{(2)}\} \,|\, \Omega \setminus w_1^{(2)}) = \dfrac{\Pr(\{w_2^{(2)}\})}{\Pr(\Omega \setminus w_1^{(2)})} > 1/2$ и во втором приближении получим: $1 = p_1^{(2)} > p_2^{(2)} > p_3^{(2)}$. Третий по значимости набор $w_3^{(2)}$ определяется как подмножество $(\Omega \setminus w_1^{(2)}) \setminus w_2^{(2)}$ аналогично определению $w_2^{(2)}$ в качестве подмножества $\Omega \setminus w_1^{(2)}$, согласно соотношениям (19), (20), (21). С помощью описанного алгоритма осуществляется построение наборов, имеющих всё меньшую и меньшую возможности. Получается необходимая для описания симптоматики упорядоченность возможностей наборов признаков (симптомов).

Следует отметить, что в задаче поиска наборов признаков, ранжированных по значимости при определённом заболевании, алгоритм гранулирования пространства значений признаков заболевания используется при неизвестных вероятностях $pr_1, pr_2 \ldots$ В [1] показано, что для каждого $n = 1, 2, \ldots$ гранулы $w_1^T, w_2^T, \ldots w_n^T$, обеспечивающие выполнение условий $pr_1^T > pr_2^T > \ldots > 0$, $pr_1^T + pr_2^T + \ldots = 1$, могут быть построены почти наверно безошибочно на основе конечного числа наблюдений, в то время как вероятности гранул и вероятности $pr_1, pr_2, \ldots$ при этих же условиях будут оценены с точностью, вообще говоря, недостаточной для того, чтобы их оценки можно было бы использовать в качестве истинных вероятностей гранул.

3. Диагностика заболеваний — обучение, распознавание. Компьютерный эксперимент

Наиболее существенными признаками заболевания k в алгоритме гранулирования являются наборы $w_1^{(1)}, w_1^{(2)}, w_1^{(3)} \ldots$, представляющие гранулы единичной возможности разных приближений, причём с ростом порядка приближения растёт информативность входящих в гранулу значений признаков (симптомов). Так как нечасто приходится наблюдать больных с одинаковым набором симптомов (и их значений), особенно при работе с подробными опросниками (в случае СРТК, например), вероятности $pr_l = \Pr((x)_l)$, $l = 1, 2, \ldots$, векторов значений признаков, аппроксимированные частотами и ранжированные в (15), малы и одни и те же значения частот присущи различным векторам значений признаков. Поэтому существует несколько наборов каждого приближения, имеющих максимальную возможность.

По обучающей выборке в каждом классе k находятся множества $\{w\}_k$, $k = 1, 2, \ldots, M$ всех наборов признаков максимального приближения, имеющих ненулевую возможность. Предъявляется субъект $x = (x^1, x^2, \ldots, x^n)$ (1), который следует отнести к одному из M классов. Выявляются все наборы $(w_s)_k$, $k = 1, \ldots, M$, из множества $\{w\}_k$, присущие субъекту. Пусть найдено S_k таких наборов в классе k: $s = 1, \ldots, S_k$. В каждом классе k определяется набор w_{s_k}, имеющий максимальную переходную возможность:

$$p(k \,|\, w_{s_k}) = \max_s p(k \,|\, w_s), \quad s = 1, \ldots, S_k.$$

Далее в (12) значение $p(x \,|\, k)$ используется в качестве $\varphi^{\chi|æ}(x|k)$, причем вместо вектора значений признаков x (1) берётся наиболее существенный в классе k набор w_{s_k}, присущий x. Субъект x относится к тому классу q^*, которому принадлежит набор $w_{s_{q^*}}$, имеющий минимальную возможность потерь (4), сопутствующих решению о постановке диагноза q^* больному при наличии у него набора $w_{s_{q^*}}$:

$$P_{q^*}(w_{s_{q^*}}) = \min_q P_q(w_{s_q}),$$

что в условиях данной задачи диагностики соответствует максимальной возможности $p(q^*|x)$ субъекту x принадлежать классу q^*. Тогда переходная возможность $P_{q^*}(|w_{s_{q^*}})$ потерь при отнесении субъекта x с $w_{s_{q^*}}$ к классу заболеваний q^*, определённая в (10), также будет минимальна.

Как уже было замечено, для описания заболевания необходимо конкретизировать его симптоматику в условиях небольшого числа объектов (пациентов, заполнивших опросники) относительно количества признаков (вопросов). Алгоритм применен к 53 субъектам, каждый из которых характеризуется 13 признаками. 34 субъекта принадлежит классу № 1 «Замедленная моторика», 19 — классу № 2 «Ускоренная моторика» [4]. Разработан интерфейс, демонстрирующий поиск диагностических критериев заболевания на основе обучающей выборки в диалоговом режиме.

Диапазон изменения признаков у субъектов обучающей выборки представлен на рис. 1. Здесь по горизонтали — номера $j = 1, \ldots, n$ признаков, $n = 13$, по вертикали — их значения. Чем больше значение признака, тем сильнее проявляется данный симптом у субъекта. Например, 13-й признак в обучающей выборке принимает значения от 1 до 7.

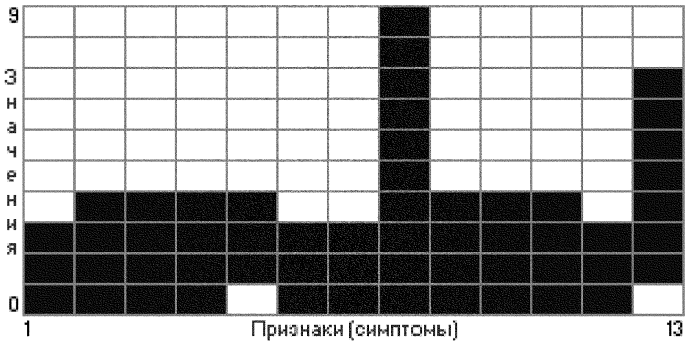

Рис. 1. Диапазон изменения всех признаков обучающей выборки. По горизонтали — признаки (симптомы) субъекта, по вертикали — их значения. Затемнённые участки соответствуют значениям признаков субъектов обучающей выборки.

Особенность реализации алгоритма гранулирования в данной задаче медицинской диагностики связана с низкими частотами элементарных событий: по результатам тестирования ни в одном классе не встретилось двух одинаковых больных. О том, чтобы восстановить вероятностную модель симптоматики в этой ситуации, не может быть и речи. Для восстановления возможностной модели проводится гранулирование, требующее перебора всех последовательностей элементарных событий с одинаковыми частотами. Переход ко всё большей степени приближения осуществляется до тех пор, пока состав гранул не перестанет изменяться. В результате в каждом классе выделяются признаки, удовлетворяющие условию (16). Во всех получившихся соотношениях вида $1 = p(w_1^T) > p(w_2^T) > \ldots$, где $w_i^T \in \Omega^T$, наиболее значимыми для данного класса заболевания являются сочетания значений признаков (гранулы) w_1^T, имеющие единичную возможность. Варианты гранулирования в классе «Замедленная моторика» приведены на рис. 2:

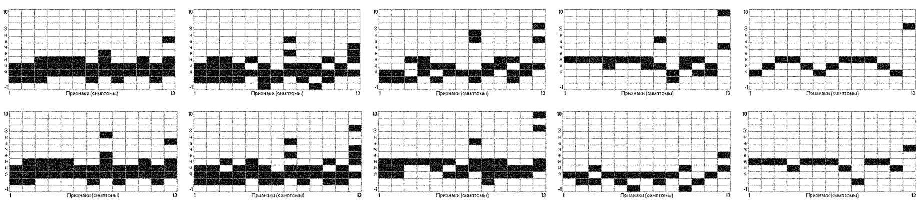

Рис. 2. Два варианта гранулирования в первом классе. В каждом варианте слева направо 5 гранул $w_1^T, w_2^T, \ldots, w_5^T$ с убывающими значениями возможности $1 = p_1^T > p_2^T > \ldots > p_5^T$. В каждой грануле по горизонтали — признаки (симптомы) субъекта, по вертикали — их значения. Затемнённые участки соответствуют значениям признаков в грануле.

Варианты гранулирования в классе «Ускоренная моторика» приведены на рис. 3:

В [2] представлен анализ нарушений по психосоматическим и гастроэнтерологическим показателям отдельно. Построенные в данной работе гранулы отражают взаимосвязь всех 13 признаков. Применение оптимизированного алгоритма позволило увидеть детализированную симптоматийную картину СРТК в целом, связь симптомов всех групп опросника, впервые осуществить анализ с использованием

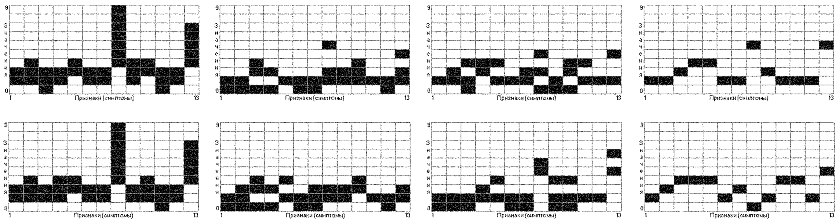

Рис. 3. Два варианта гранулирования во втором классе.

всех признаков сразу. На рисунках 3 и 4 заметно различие симптоматийных картин первого и второго классов. Например, важным по мнению врачей результатом гранулирования явилось различие диапазонов значений 8-го симптома (предпосылки болей) и 13-го симптома (факторы, предшествующие заболеванию). Класс «Ускоренная моторика» характеризуется бо́льшими значениями этих признаков, чем класс «Замедленная моторика».

Результатом применения теоретико-возможностного метода является построение возможностной модели симптоматики заболевания, определяющей нечёткую связь между симптоматийным описанием состояния больного и реальным его состоянием и выражающуюся в распределении возможности пары нечётких элементов («наблюдение, как набор зарегистрированных признаков (симптомов)», «реальное состояние больного»). Разработан и реализован алгоритм гранулирования пространства значений признаков заболевания, с помощью которого в диалоговом режиме устанавливаются диагностические критерии заболевания, характеризующиеся группами значений признаков (симптомов), ранжированных по значениям их возможностей при данном заболевании. В результате получена детализированная симптоматийная картина двух групп СРТК, отражающая взаимосвязь всех симптомов опросника. Анализ данного заболевания впервые осуществлён по всем психосоматическим и гастроэнтерологическим признакам опросника.

Список литературы

[1] Пытьев Ю. П. Возможность как альтернатива вероятности. М.: Физматлит, 2007.

[2] Газарян В. А., Илюшин В. Л., Пытьев Ю. П., Шаховская А. К. // Вестн. Моск. ун-та. Физ. Астрон. 2005. № 4. С. 3–6.

[3] Шаховская А. К., Есаулов В. И., Лоранская Т. И. Многофакторный анализ состояния здоровья и схема терапии больных с функциональными заболеваниями толстой кишки // Матер. III Российской гастроэнтерологической недели. М., 1997.

[4] Газарян В. А., Иваницкая Н. В., Пытьев Ю. П., Шаховская А. К. О задаче компьютерной диагностики заболеваний желудочно-кишечного тракта // Вестн. Моск. ун-та. Физ. Астрон. 2003. № 2. С. 12–15.

[5] Пытьев Ю. П. Возможность. Элементы теории и применения. М.: Эдиториал УРСС. 2000.