Сравнительный анализ эффективности вероятностного и возможностного алгоритмов медицинской диагностики

Ю. П. Пытьев[1], В. А. Газарян[1,2,a], П. Б. Росницкий[3]

[1] *Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра компьютерных методов физики. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2.*
[2] *Финансовый университет при правительстве РФ, факультет прикладной математики и информационных технологий, кафедра «Теория вероятностей и математическая статистика». Россия, 125993, Москва, Ленинградский проспект, д. 49.*
[3] *Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра акустики. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2.*
E-mail: [a] *varvaragazaryan@yandex.ru*

Статья поступила 23.01.2014, подписана в печать 01.02.2014.

Для решения задач медицинской диагностики широко используются математические методы распознавания образов и построенные на их основе алгоритмы классификации заболеваний [1]. В работе [2] для классификации функциональных нарушений системы пищеварения применена алгебраическая модель алгоритма Кора. В работах [3–5] показано, что при решении многих задач медицинской диагностики более эффективными являются возможностные методы постановки медицинского диагноза. В настоящей работе приведен сравнительный анализ вероятностной и возможностной моделей постановки диагноза, алгоритмов Кора и результатов их применения к решению задачи диагностики острого аппендицита.

Ключевые слова: распознавание образов, задача идентификации, вероятностная модель диагностики, возможностная модель диагностики, гранулирование, алгоритм классификации Кора, острый аппендицит.

УДК: 519.2, 519.6. PACS: 02.70.–c, 02.50.Le.

Введение

Рассмотрим задачу медицинской диагностики как задачу идентификации, в которой требуется принять решение о принадлежности медицинского объекта, в данном случае — субъекта (пациента), к одному из M заданных врачом классов заболеваний, среди которых может быть и класс «норма», либо принять решение о том, что данный субъект не относится ни к одному из выделенных классов. При этом он может страдать заболеваниями, диагностика которых выходит за рамки настоящего исследования. Признаками, характеризующими субъект, являются симптомы заболевания, обнаруженные в результате обследования и опроса пациента, каждый симптом может принимать как количественные, так и качественные значения. Рассмотрим традиционно два этапа решения задачи идентификации — обучения и постановки предварительного диагноза. Процесс обучения состоит в определении характерных значений признаков (симптомов) заболевания в каждом из M классов по обучающей выборке объектов. Класс «норма», как правило, характеризуется значениями признаков, находящимися в пределах нормы. После этого на основании результатов обучения проводится идентификация — отнесение диагностируемого объекта к одному из M классов либо отказ от идентификации, если у данного пациента не наблюдается характерных симптомов выделенных классов[1].

В [3–5] показано, что при моделировании медицинских объектов исследователям приходится на практике сталкиваться с нечеткостью их описания, связанной со случайностью и неточностью данных, которые вызваны изменчивостью во времени, неформализованным и во многих случаях субъективным характером симптомов заболевания. Эти факторы наряду с ограниченным размером обучающих выборок приводят к принципиальным проблемам эмпирического построения стохастических моделей медицинских объектов. Если же моделируемый объект не является стохастическим, то вероятностной модели вообще не существует. Тогда неточность и нечеткость, свойственную объектам, нельзя охарактеризовать в вероятностных терминах. Однако судить о вероятностной или не вероятностной природе объектов непросто в связи с отсутствием такого критерия в теории вероятностей [6]. Ввиду неэффективности вероятностных методов при моделировании медицинских объектов естественно обратиться к невероятностным моделям случайности, нечеткости и неопределенности [7, 8]. В теории возможностей, разработанной в [6] и успешно применяемой для решения задач медицинской диагностики [3–5], показано, что в то время как вероятностную модель стохастического объекта, непредсказуемо эволюционирующего во времени, эмпирически построить невозможно, его возможностная модель, при достаточно слабых ограничениях на характер эволюции вероятностной модели, может быть восстановлена, причем точно и на осно-

[1] Далее в статье употребляется также термин «классификация», понимаемый как «идентификация», т. е. отнесение субъекта к одному из заранее определенных классов, причем «отсутствие выделенных заболеваний» рассматривается как отдельный класс.

вании конечного числа наблюдений. Таким образом, при неформализованном характере признаков заболевания, ограниченном размере обучающих выборок и непредсказуемой изменчивости вероятностных свойств симптомов возможностные методы обучения и распознавания более предпочтительны, чем вероятностные.

1. Вероятностная модель постановки медицинского диагноза

Предположим, что признаки заболеваний имеют стохастическую природу. Тогда каждый объект можно охарактеризовать n-мерным случайным вектором признаков $\boldsymbol{\chi} = (\chi^1, \chi^2, \ldots, \chi^n)$, принимающим значения $\boldsymbol{x} \in X$, где

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

$x^j \in X^j$ – значение j-го признака (симптома), $j = 1, \ldots, n$, X^j — множество значений j-го признака, n_j — количество значений j-го признака, $j = 1, \ldots, n$, n — число признаков,

$$X = X^1 \times X^2 \times \ldots \times X^n, \tag{2}$$

$\mathrm{pr}^{\chi}(\boldsymbol{x})$ — значение вероятности равенства $\boldsymbol{\chi} = \boldsymbol{x}$, $\boldsymbol{x} \in X$.

Задача идентификации — отнести предъявленный для диагностики объект к одному из классов k, $k \in \{1, \ldots, M\}$.

Обозначим κ случайный элемент, значениями которого являются номера классов $k \in \{1, \ldots, M\}$. Пусть $\mathrm{pr}^{\chi,\kappa}(\boldsymbol{x}, k)$ — вероятность равенств $\boldsymbol{\chi} = \boldsymbol{x}$, $\kappa = k$, $\boldsymbol{x} \in X$, $k \in \{1, \ldots, M\}$, характеризующая совместное распределение наблюдаемого набора симптомов $\boldsymbol{\chi}$ и класса заболеваний κ, $l_{kd} \in [0, 1]$ — вероятность потерь при отнесении объекта (пациента) класса k к классу d, $d = 1, \ldots, M$, которую следует понимать как определенную врачом вероятность неблагоприятных для здоровья пациента последствий, вызванных постановкой ему диагноза «d», в то время как на самом деле он страдает заболеванием «k», $k, d = 1, \ldots, M$. Обозначим $\mathrm{pr}^{\kappa}(k)$, $k = 1, \ldots, M$, априорную вероятность заболевания k. В рассматриваемой модели диагностики ни один класс не является более предпочтительным, чем другой, и априорные вероятности $\mathrm{pr}^{\kappa}(k)$ равны $\mathrm{pr}^{\kappa}(k) = 1/M$, $k = 1, \ldots, M$. Обозначим $\mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|k)$ — значение условной вероятности наблюдения симптомов $\boldsymbol{\chi} = \boldsymbol{x}$ у пациента, страдающего заболеванием $\kappa = k$. Тогда $\mathrm{pr}^{\chi,\kappa}(\boldsymbol{x}, k) = \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|k)\,\mathrm{pr}^{\kappa}(k)$.

Пусть решение о принадлежности пациента с набором признаков $\boldsymbol{\chi} = \boldsymbol{x}$, $\boldsymbol{x} \in X$, к классу k принимается при $\boldsymbol{x} \in X_k$, где $X_1, X_2, \ldots, X_M$ — некоторое упорядоченное разбиение множества $X = X^1 \times X^2 \times \ldots \times X^n$ (2) значений признаков. $X = \bigcup\limits_{j=1}^{M} X_j$, $X_j \cap X_i = \varnothing$, $i \neq j$, $i, j = 1, \ldots, M$. Поэтому каждое разбиение $X = \bigcup\limits_{j=1}^{M} X_j$ определяет правило постановки диагноза. По сути, множество X_k состоит из значений признаков, характерных для класса k, $k = 1, \ldots, M$.

Для правила постановки диагноза, определенного конкретным разбиением X, математическое ожидание вероятности потерь (риск потерь) [6]

$$L(X) = \sum_{j=1}^{M} \int_{X_j} S_j(\boldsymbol{x})\, d\boldsymbol{x}, \tag{3}$$

$$S_j(\boldsymbol{x}) = \sum_{k=1}^{M} l_{kj}\, \mathrm{pr}^{\chi,\kappa}(\boldsymbol{x}, k) = \sum_{k=1}^{M} l_{kj}\, \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|k)\, \mathrm{pr}^{\kappa}(k), \tag{4}$$
$$\boldsymbol{x} \in X, \quad j = 1, \ldots, M,$$

— математическое ожидание вероятности потерь при отнесении пациента $\boldsymbol{x}$ к классу j.

Рассмотрим задачу определения оптимального правила постановки диагноза (байесовского) как задачу отыскания разбиения X, минимизирующего риск L (3). В работе [6] показано, что минимум (3) достигается на любом упорядоченном разбиении $X^* = \bigcup\limits_{j=1}^{M} X_j^*$, $X_j^* \cap X_i^* = \varnothing$, $i \neq j$, $i, j = 1, \ldots, M$, удовлетворяющем условию

$$X_j^* \subset \left\{ \boldsymbol{x} \in X,\ S_j(\boldsymbol{x}) = \min_{1 \leqslant i \leqslant M} S_i(\boldsymbol{x}) \right\}, \quad j = 1, \ldots, M. \tag{5}$$

Минимальное значение риска (3)

$$L(X^*) = \sum_{j=1}^{M} \int_{X_j^*} S_j(\boldsymbol{x})\, d\boldsymbol{x}. \tag{6}$$

В [6] также показано, что правило идентификации, определенное разбиением X^*, удовлетворяющим условию (5), можно определить как решающую функцию $d^*(\cdot)\colon X \to \{1, \ldots, M\}$ такую, что $d^*(\boldsymbol{x}) = k$ для каждого $\boldsymbol{x} \in X$, если $\boldsymbol{x} \in X_k^*$, $k = 1, \ldots, M$, т. е.

$$d^*(\boldsymbol{x}) \in D^*(\boldsymbol{x}) \mathrel{\hat{=}} \left\{ d \in \{1, \ldots, M\},\ S_d(\boldsymbol{x}) = \min_{1 \leqslant i \leqslant M} S_i(\boldsymbol{x}) \right\}. \tag{7}$$

Если $l_{kj} = 1 - \delta_{kj}$, т. е. вероятность потерь равна нулю при правильном решении ($k = j$) и единице при любом ошибочном решении ($k \neq j$), то риск (3) равен ожидаемой доле ошибочных решений, или вероятности ошибки идентификации, а согласно (4)

$$S_j(\boldsymbol{x}) = \sum_{k=1}^{M} l_{kj}\, \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|k)\, \mathrm{pr}^{\kappa}(k) = \mathrm{pr}^{\chi}(\boldsymbol{x}) - \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|j)\, \mathrm{pr}^{\kappa}(j),$$
$$\boldsymbol{x} \in X, \quad j = 1, \ldots, M,$$

где $\mathrm{pr}^{\chi}(\boldsymbol{x})$ — распределение вектора симптомов $\boldsymbol{\chi}$. В этом случае решением задачи минимизации риска (3) будет вместо разбиения (5) разбиение

$$X_j^* \subset \left\{ \boldsymbol{x} \in X,\ \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|j)\, \mathrm{pr}^{\kappa}(j) \geqslant \max_{i \neq j} \mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|i)\, \mathrm{pr}^{\kappa}(i) \right\},$$
$$j = 1, \ldots, M. \tag{8}$$

Следовательно, согласно байесовскому решающему правилу, к классу j следует отнести пациентов, обладающих такими значениями симптомов $\boldsymbol{\chi} = \boldsymbol{x}$, для которых значение условной вероятности $\mathrm{pr}^{\chi|\kappa}(\boldsymbol{x}|j)$ максимально.

2. Вероятностный алгоритм типа Кора. Обучение. Распознавание

Для оценки условных вероятностей $\mathrm{pr}^{x|\kappa}(\boldsymbol{x}|j)$ и решения задачи классификации заболеваний в рамках вероятностной модели диагностики в настоящей работе применяется алгоритм классификации типа Кора. Существует несколько разновидностей алгоритма Кора [9, 10]. На этапе обучения алгебраического алгоритма строятся сочетания характерных значений признаков класса, называемые представительными наборами класса. Варьируя эмпирические параметры, которые вводятся на этапах обучения и распознавания алгебраического алгоритма, можно добиваться разных результатов диагностики, однако решение многомерной задачи оптимизации в этом случае представляет значительные трудности. В работе [2] алгоритм Кора был модифицирован на базе его алгебраической модели путем применения байесовского решающего правила.

Обучающее множество содержит N объектов — больных с верифицированным диагнозом из M непересекающихся классов заболеваний. Обучающая выборка объектов k-го класса — $\boldsymbol{\omega}_{k1},\ldots,\boldsymbol{\omega}_{kN_k}$, где N_k — число объектов обучающей выборки k-го класса, $\boldsymbol{\omega}_{kl}=(\omega_{kl}^1,\ldots,\omega_{kl}^n)$, $l=1,\ldots,N_k$, $k=1,\ldots,M$, $\sum\limits_{k=1}^{M}N_k=N$. Иными словами, ω_{kl}^j — значение j-го признака l-го объекта k-го класса. В [2] рассмотрен алгоритм построения представительных наборов классов путем сравнения на этапе обучения каждого объекта $\boldsymbol{\omega}_{kl}$ k-го класса, $l=1,\ldots,N_k$, по всем признакам $j=1,\ldots,n$ с остальными объектами обучающей выборки. Пусть $D_q^k(\boldsymbol{\omega}_{kl})$ — q-й представительный набор значений признаков k-го класса, порожденный объектом $\boldsymbol{\omega}_{kl}$: $D_q^k(\boldsymbol{\omega}_{kl})=\omega_{kl}^{j_1},\ldots,\omega_{kl}^{j_r}$, $j_1<\ldots<j_r$, $r\leqslant n$, и имеющий в k-м классе частоту не менее ν_k. Поскольку длинные представительные наборы (с большими значениями r) встречаются реже, чем короткие, у которых r меньше, необходимо задать минимальную длину представительного набора класса $r_{\min}$. В противном случае в каждом классе получим $r=1$, т. е. представительный набор будет состоять всего из одного признака, и описание класса окажется неполным. На этапе обучения по обучающей выборке находятся все представительные наборы $D_q^k(\omega_{kl})$ объектов, $l=1,\ldots,N_k$, $k=1,\ldots,M$.

На этапе распознавания предъявляется объект $\boldsymbol{x}=\left(x^1,x^2,\ldots,x^n\right)$, который следует отнести к одному из M классов. Для классификации объекта $\boldsymbol{x}$ требуется не все его описание, а только представительные наборы, которыми он обладает. Выявляются все представительные наборы всех классов, присущие объекту $\boldsymbol{x}$. Пусть найдено Q_k таких наборов в классе k: $q=1,\ldots,Q_k$. В каждом классе k определяется представительный набор $D_{q_k}^k$, имеющий минимальное математическое ожидание вероятности потерь (4) при отнесении пациента $\boldsymbol{x}$ с данным представительным набором к классу k:

$$S_k\left(D_{q_k}^k\right)=\min_q S_k\left(D_q^k\right), \tag{9}$$

$$q=1,\ldots,Q_k,\quad S_k\left(D_q^k\right)=\sum_{i=1}^{M}l_{ik}\,\mathrm{pr}\left(D_q^k|i\right)\mathrm{pr}(i).$$

Согласно решающему правилу (7), объект $\boldsymbol{x}$ относится к классу k^*, в котором математическое ожидание вероятности потерь $S_{k^*}\left(D_{q_{k^*}}^{k^*}\right)$ минимально:

$$S_{k^*}\left(D_{q_{k^*}}^{k^*}\right)=\min_k S_k\left(D_{q_k}^k\right). \tag{10}$$

В случае когда $l_{kj}=1-\delta_{kj}$, согласно (8), в каждом классе k определяется представительный набор $D_{q_k}^k$, имеющий максимальную вероятность

$$\mathrm{pr}\left(D_{q_k}^k|k\right)=\max_q \mathrm{pr}\left(D_q^k|k\right),\quad q=1,\ldots,Q_k. \tag{11}$$

Тогда объект $\boldsymbol{x}$ относится к тому классу k^*, которому принадлежит набор $D_{q_{k^*}}^{k^*}$, имеющий максимальную вероятность $\mathrm{pr}\left(k^*|D_{q_{k^*}}^{k^*}\right)$ в (8): $\mathrm{pr}\left(k^*|D_{q_{k^*}}^{k^*}\right)=\max\limits_k \mathrm{pr}\left(k|D_{q_k}^k\right)$. При равенстве априорных вероятностей $\mathrm{pr}(k)$ решение о диагнозе k^* принимается на основании условия

$$\mathrm{pr}\left(D_{q_{k^*}}^{k^*}|k^*\right)=\max_k \mathrm{pr}\left(D_{q_k}^k|k\right). \tag{12}$$

Согласно закону больших чисел, при достаточно большом объеме N обучающей выборки можно аппроксимировать вероятности в (9) и (11) частотами при неизменных вероятностных характеристиках признаков. В частности, аппроксимируя на практике вероятности $\mathrm{pr}\left(D_q^k|k\right)$ частотами встречаемости объектов, имеющих D_q^k в k-м классе, следует помнить об условиях практического применения ЗБЧ. Применим оценки Хёфдинга ошибки приближения вероятности частотой [6] для выборки объема N_k в k-м классе (как для неизменных, так и для меняющихся вероятностных характеристик объектов, что в данной задаче является актуальным из-за индивидуальных особенностей пациентов и изменчивости их состояния). Пусть $\nu_q^{(N_k)}=\nu\left(D_q^k|k\right)$ — частота q-го представительного набора, $pr_q^{(N_k)}=\frac{1}{N_k}\sum\limits_{j=1}^{N_k}\mathrm{pr}_j\left(D_q^k|k\right)$ — его «эмпирическая вероятность» в k-м классе с N_k объектами обучающей выборки. Тогда, согласно лемме Хёфдинга, вероятность отклонения частоты представительного набора от вероятности оценивается следующим образом:

$$\Pr\left(\left\{\left|\nu_q^{(N_k)}-\mathrm{pr}_q^{(N_k)}\right|>\varepsilon\right\}\right)\leqslant 2\exp(-2N_k\varepsilon^2). \tag{13}$$

Поскольку $\sum\limits_{N_k=1}^{\infty}\exp(-2N_k\varepsilon^2)<\infty$ при любом $\varepsilon>0$, то, согласно лемме Бореля–Кантелли о достаточном условии сходимости с вероятностью единица, $\nu_q^{(N_k)}-\mathrm{pr}_q^{(N_k)}\overset{\text{п. н.}}{\to}0$.

Если в Ω_k имеется N_k объектов, N_k^q из них содержат набор D_q^k, то $\mathrm{pr}\left(D_q^k|k\right)$ оценим значением $\mathrm{pr}\left(D_q^k|k\right)\approx\frac{N_k^q}{N_k-1}$. В таблице приведены результаты оценки сверху вероятности отклонения частоты представительного набора класса k от его вероятности при различных объемах обучающих выборок N_k согласно (13). Оценки показывают, что для аппроксимации вероятностей в (9) и (11) частотами требуется достаточно большой объем обучающих выборок.

Следует отметить принципиальную важность оценки тяжести последствий разных вариантов ошибочного

Оценки вероятности отклонения частоты представительного набора D_s^k от его вероятности при различных значениях ε и N_k

ε	Объем обучающей выборки класса		
	100	200	500
0.05	1	0.74	0.16
0.1	0.27	0.037	10^{-4}
0.2	$6.7 \cdot 10^{-4}$	$2 \cdot 10^{-7}$	0

диагноза в медицинской практике. В математическое ожидание потерь (9), сопутствующих постановке определенного диагноза, входит матрица потерь с заданными врачом элементами l_{kd}, характеризующими вероятность потерь при постановке диагноза d пациенту, страдающему на самом деле заболеванием k. Если при точном решении вероятность потерь равна нулю, а при любом ошибочном решении — единице, то в алгоритме Кора применяется решающее правило (12), и риск равен ожидаемой доле ошибочных решений, т. е. вероятности ошибки идентификации.

3. Возможностная модель постановки медицинского диагноза

В работе [5] подробно рассмотрено решение задачи возможностного моделирования процесса постановки медицинского диагноза, сформулировано правило постановки диагноза, минимизирующее риск потерь. По аналогии с вероятностным моделированием, при построении возможностной модели диагностики каждый объект (пациент) характеризуется n-мерным нечетким вектором признаков $\boldsymbol{\chi} = (\chi^1, \chi^2, \ldots, \chi^n)$, принимающим значения $\boldsymbol{x} \in X$ (2), где

$$\boldsymbol{x} = (x^1, x^2, \ldots, x^n), \quad \boldsymbol{x} \in X, \tag{14}$$

$x^j \in X^j$ — значение j-го признака (симптома) заболевания, $j = 1, \ldots, n$, $X = X^1 \times X^2 \times \ldots \times X^n$, X^j — множество значений j-го признака (2), n_j — количество значений j-го признака, $j = 1, \ldots, n$.

В задаче диагностики требуется принять решение о принадлежности больного $\boldsymbol{x}$ к одному из M классов заболеваний. Решение о состоянии больного определяется в [5] как нечеткий элемент δ, принимающий значения на множестве $\{1, \ldots, M\}$. Предполагая, согласно мнению врачей, что ни один из классов заболеваний не является априори более «предпочтительным», чем другие, получаем равенство априорных возможностей $\phi^{\kappa}(k) = 1$, $k = 1, \ldots, M$. В этом случае возможность потерь, определяющая качество правила постановки диагноза $\pi^{\delta|\chi}$, задается в [5] как

$$\mathrm{PL}(\pi^{\delta|\chi}) = \sup_{\substack{\boldsymbol{x} \in X, \\ k \in \{1,\ldots,M\}, \\ d \in \{1,\ldots,M\}}} \min\left(l_{kd}, \pi^{\delta|\chi}(d|\boldsymbol{x}), \phi^{\chi|\kappa}(\boldsymbol{x}|k)\right), \tag{15}$$

где κ — нечеткий элемент, значениями которого являются номера классов (заболеваний) $k \in \{1, \ldots, M\}$; $\phi^{\chi|\kappa}(\boldsymbol{x}|k)$ — переходная возможность равенства $\boldsymbol{\chi} = \boldsymbol{x}$, когда $\kappa = k$, $\boldsymbol{x} \in X$; $l_{kd} \in [0, 1]$ — возможность потерь при отнесении субъекта класса k к классу q, $q = 1, \ldots, M$; $\pi^{\delta|\chi}(d|\boldsymbol{x})$ — возможность решения о заболевании $\delta = d$, когда $\boldsymbol{\chi} = \boldsymbol{x}$ — наблюдающиеся у больного симптомы. Оптимальным является правило $\pi^{*\delta|\chi}$, минимизирующее возможность потерь (15):

$$\mathrm{PL}(\pi^{*\delta|\chi}) = \min_{\pi^{\delta|\chi}} \mathrm{PL}(\pi^{\delta|\chi}). \tag{16}$$

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

$$\delta^*(\boldsymbol{x}) \in \Big\{ d \in \{1, \ldots, M\}, \; \pi^{*\delta|\chi}(d|\boldsymbol{x}) = \max_{d' \in \{1,\ldots,M\}} \pi^{*\delta|\chi}(d'|\boldsymbol{x}) \Big\}, \quad \boldsymbol{x} \in X.$$

В [6] показано, что оптимальное правило $\pi^{*\delta|\chi}$ в (16) можно получить путем решения задачи на минимум для каждого вектора значений признаков $\boldsymbol{x} \in X$:

$$\max_{1 \leqslant d \leqslant M} \min\left(\pi^{\delta|\chi}(d|\boldsymbol{x}), P_d(\boldsymbol{x})\right) \sim \min_{\pi^{\delta|\chi}(\cdot|\boldsymbol{x})}, \tag{17}$$

$$P_d(\boldsymbol{x}) = \max_{1 \leqslant k \leqslant M} \min\left(l_{kd}, \phi^{\chi|\kappa}(\boldsymbol{x}|k)\right), \tag{18}$$

где $P_d(\boldsymbol{x})$ — возможность потерь, сопутствующих решению о постановке диагноза $\delta = d$ больному при наличии у него симптомов $\boldsymbol{\chi} = \boldsymbol{x}$, $\boldsymbol{x} \in X$, $d \in \{1, \ldots, M\}$. В [6] сформулированы следующие достаточные условия оптимальности правила $\pi^{*\delta|\chi}$. Пусть

$$D^*(\boldsymbol{x}) = \left\{ d \in \{1, \ldots, M\}, \; P_d(\boldsymbol{x}) = \min_{d'} P_{d'}(\boldsymbol{x}) \right\}, \quad \boldsymbol{x} \in X. \tag{19}$$

Тогда в качестве решения задачи (17) можно использовать значение $d^*(\boldsymbol{x})$ любой функции, удовлетворяющей условию $d^*(\boldsymbol{x}) \in D^*(\boldsymbol{x})$, $\boldsymbol{x} \in X$.

Если потери невозможны при правильной постановке диагноза $l_{kd} = 0$ при $d = k$ и возможность потерь максимальна при любой ошибочной постановке диагноза $l_{kd} = 1$ при $d \neq k$, $k, d = 1, \ldots, M$, то в (18) $P_d(\boldsymbol{x}) = \max_{k \neq d} \phi^{\chi|\kappa}(\boldsymbol{x}|k)$, $d = 1, \ldots, M$, $\boldsymbol{x} \in X$ и минимум $P_d(\boldsymbol{x})$ при фиксированном $\boldsymbol{x}$ достигается на тех $d \in \{1, \ldots, M\}$, при которых $\phi^{\chi,\kappa}(\boldsymbol{x}, k)$ достигает максимума, а правило $\pi^{*\delta|\chi}$, минимизирующее возможность потерь при постановке диагноза, является правилом максимальной возможности [6].

Из условий (17)–(19) следует, что для построения оптимального правила постановки диагноза d^* следует восстановить распределение переходных возможностей $\phi^{\chi|\kappa}(\boldsymbol{x}|k)$, $k = 1, \ldots, M$. В [5] рассмотрен алгоритм гранулирования пространства значений признаков заболевания, осуществляющий стохастическое моделирование переходных возможностей $\phi^{\chi|\kappa}(\boldsymbol{x}|k)$ на основании обучающей выборки объектов.

4. Возможностный алгоритм классификации типа Кора. Обучение и распознавание

В теории возможностей по аналогии с конструкцией и терминологией теории вероятностей [6], вероятностному пространству $(X, P(X), \Pr)$ соответствует пространство с возможностью $(X, P(X), P)$, где $X = X^1 \times X^2 \times \ldots \times X^n$ (2). Обозначим $\mathrm{pr}_i = \mathrm{pr}(x_i)$ и $p_i = p(\boldsymbol{x}_i)$ значения вероятности и соответственно возможности равенства $\boldsymbol{\chi}_i = \boldsymbol{x}_i$ для i-го объекта, $i = 1, \ldots, n$, $j = 1, \ldots, n$. Упорядочив вероятно-

сти векторов значений симптомов $x_1, x_2, \ldots$ в модели $(X, \mathrm{P}(X), \mathrm{Pr})$

$$\mathrm{pr}_1 \geqslant \mathrm{pr}_2 \geqslant \ldots \geqslant 0, \quad \mathrm{pr}_1 + \mathrm{pr}_2 + \ldots = 1, \qquad (20)$$

определим их возможностную модель $(X, \mathrm{P}(X), P)$, в которой возможности векторов значений симптомов удовлетворяют условию

$$1 = p_1 \geqslant p_2 \geqslant \ldots \geqslant 0. \qquad (21)$$

Конкретная упорядоченность в (21), содержащая в определенных местах строгие неравенства, а в остальных — равенства, определяет единственную (с точностью до эквивалентности) возможностную модель, в которой

$$p_k = p_{k+1}, \text{ если } \mathrm{pr}_k \leqslant \frac{(1 - \mathrm{pr}_1 - \mathrm{pr}_2 - \ldots - \mathrm{pr}_{k-1})}{2}, \qquad (22)$$

$$p_k > p_{k+1}, \text{ если } \mathrm{pr}_k > \frac{(1 - \mathrm{pr}_1 - \mathrm{pr}_2 - \ldots - \mathrm{pr}_{k-1})}{2}. \qquad (23)$$

При выполнении условий (22), (23) возможность является максимально согласованной с вероятностью [6] и число строгих неравенств в (21) максимально. Векторы симптомов $\boldsymbol{x}_k$ и $\boldsymbol{x}_{k+1}$ различаются по значимости, если $p_k > p_{k+1}$ (23), однако на практике разности вероятностей векторов значений признаков оказываются малыми и выполняется условие (22), следовательно, различие вероятностей не приводит в данном случае к различию возможностей. Для того чтобы «стохастические детали» стали «различимыми» для возможности, проведено гранулирование пространства X: векторы значений признаков $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, ..., удовлетворяющие условию (20), объединены в гранулы w_1^T, w_2^T, ... так, чтобы возможности этих гранул были строго упорядочены:

$$p_1^T > p_2^T > \ldots, \quad p_i^T = p(w_i^T). \qquad (24)$$

В этом случае значения признаков, объединенные в одну гранулу w_i^T (имеющие одинаковую возможность), неразличимы с точки зрения их значимости при данном заболевании — в пределах каждой гранулы вероятности отдельных векторов $\boldsymbol{x}$, входящих в нее, могут меняться во времени. Соответственно значения признаков принципиально «различны», если принадлежат гранулам, имеющим разные возможности.

Для создания наиболее информативных диагностических критериев заболевания в [5] рассмотрен алгоритм гранулирования, позволяющий получить разбиения $T_1, T_2, \ldots$, содержащие максимальное число строго упорядоченных по значимости гранул значений признаков. Наиболее характерными признаками заболевания k являются найденные методом гранулирования нечеткие представительные наборы - гранулы значений признаков, имеющие в классе k единичную возможность, а в остальных классах — меньшие возможности. На этапе обучения возможностного (нечеткого) алгоритма Кора по обучающей выборке находятся множества $\{w\}_k$, $k = 1, \ldots, M$, всех нечетких представительных наборов.

Распознавание в нечетком алгоритме Кора осуществляется следующим образом. Предъявляется объект $\boldsymbol{x} = \left(x^1, x^2, \ldots, x^n\right)$ (14), который следует отнести к одному из M классов. Для классификации объекта $\boldsymbol{x}$ требуется не все его описание (14), а только нечеткие представительные наборы, которыми он обладает. Пусть найдено S_k таких наборов в классе k: $s = 1, \ldots, S_k$. В каждом классе k определяется нечеткий представительный набор w_{s_k}, имеющий минимальную возможность потерь (18) на множестве всех представительных наборов класса k, где в качестве $\phi^{\mathrm{x}|\kappa}(\boldsymbol{x}|k)$ используется значение возможности $p\left(w_{s_k}|k\right)$:

$$P_k\left(w_{s_k}\right) = \min_s P_k(w_s),$$

$$s = 1, \ldots, S_k, \quad P_k(w_s) = \max_{1 \leqslant q \leqslant M} \min\left(l_{qk}, p(w_s|q)\right).$$

Объект $\boldsymbol{x}$ относится к классу q^*, в котором возможность потерь при постановке диагноза q^*, минимальна: $P_{q^*}\left(w_{s_{q^*}}\right) = \min_q P_q\left(w_{s_q}\right)$, $q = 1, \ldots, M$.

Если $l_{qk} = 0$ при $q = k$ и $l_{qk} = 1$ при $q \neq k$, то в каждом классе k определяется нечеткий представительный набор w_{s_k}, имеющий максимальную возможность: $p\left(w_{s_k}|k\right) = \max_s p(w_s|k)$, $s = 1, \ldots, S_k$. Объект $\boldsymbol{x}$ относится к классу q^*, в котором возможность набора $w_{s_{q^*}}$ максимальна: $p\left(w_{s_{q^*}}|q^*\right) = \max_q p\left(w_{s_q}|q\right)$. Тогда возможность $P\left(q^*|w_{s_{q^*}}\right)$ ошибки классификации объекта $\boldsymbol{x}$ с $w_{s_{q^*}}$ при отнесении его к классу q^* будет минимальна.

5. Результаты применения вероятностного и возможностного алгоритмов Кора

В [4] возможностные методы медицинской диагностики применялись для решения задачи диагностики острого аппендицита (ОА). Своевременное обнаружение ОА чрезвычайно важно для здоровья и жизни больного. Качество врачебной диагностики ОА и хирургические возможности достигли в настоящее время высокого уровня, однако и в данной области остаются еще некоторые проблемы [11, 12]. В [2] найдены типичные наборы признаков ОА как класса в целом, так и различных его форм посредством применения алгоритма поиска групп признаков, ранжированных по значениям их возможностей. Несмотря на то что удалось выделить типичные наборы признаков трех форм ОА — гангренозной (1-й класс), флегмонозной (2-й класс) и катаральной (3-й класс), результаты классификации этих форм ОА на практике не всегда оказываются удовлетворительными. В настоящей работе для классификации ОА и трех его форм применяются рассмотренные во 2-м и 4-м разделах вероятностный и возможностный алгоритмы Кора.

Обучающая выборка состоит из 28 объектов 1-го класса, 25 — второго и 26 — третьего. Четвертый класс — «неподтвержденный диагноз» (НД) — содержит 24 объекта обучающей выборки. Каждый объект характеризуют 8 выделенных врачами признаков-симптомов, которые могут принимать от 2 до 4 значений в ранговой шкале в зависимости от степени тяжести симптома [4].

Рассмотрим результаты применения вероятностного алгоритма Кора. На этапе обучения алгоритма следует задать минимальную длину представительного набора $r_{\min}$ и порог ν_k по частоте встречаемости представительного набора в k-м классе, $k = 1, \ldots, 4$. В выборку для классификации входят как объекты обучающей

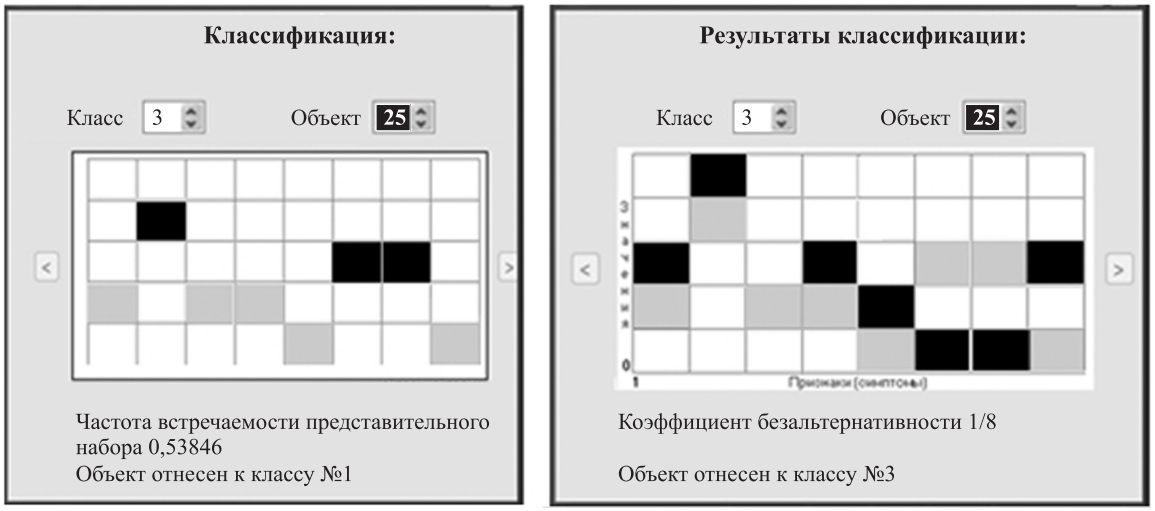

Рис. 1. Ошибочная классификация объекта 3-го класса вероятностным алгоритмом Кора (слева) и безошибочная — возможностным (справа)

выборки, так и те, которые не были использованы при обучении — 11 объектов первого класса и по 12 объектов остальных классов, составляющих контрольную выборку. В результате применения вероятностного алгоритма Кора классификация больных на классы ОА (с точностью 91%) и НД (с точностью 97%) осуществляется только при определенных значениях входных параметров: $r_{\min} = 3$ и $\nu_4 = 0.9$. Следует отметить, что даже при выборе оптимальных параметров наблюдаются случаи ошибочного отнесения больных ОА к классу НД. Классифицировать три формы ОА с помощью вероятностного алгоритма Кора не представляется возможным ни при каких значениях эмпирических параметров.

На этапе обучения возможностного алгоритма Кора найдены нечеткие представительные наборы каждого из четырех классов. На рис. 1 приведены результаты классификации одного из объектов 3-го класса «катаральный аппендицит» контрольной выборки вероятностным (*а*) и возможностным (*б*) алгоритмами Кора. По горизонтали отмечены 8 признаков (симптомов) объектов, по вертикали — значения выделенных признаков. На рис. 1, *а* черным отмечены значения признаков, входящие в представительный набор, по которому принято решение о диагнозе. На рис. 1, *б* все выделенные значения признаков составляют гранулу максимальной возможности третьего класса. Серым отмечены значения признаков объекта, черным — остальные значения признаков, входящие в гранулу. Точность классификации первого класса возможностным алгоритмом составляет 85%. Ранее было показано, что второй и третий классы (флегмонозный и катаральный аппендицит соответственно) не удалось классифицировать с помощью вероятностного алгоритма Кора. Эти классы имеют много общих симптомов, и в результате обучения возможностного алгоритма получено, что их нечеткие представительные похожи, но отличаются от нечетких представительных наборов остальных классов. Поэтому было принято решение объединить 2-й и 3-й классы при диагностике ОА.

На рис. 2 представлены результаты диагностики 2-го и 3-го классов ОА вероятностным алгоритмом

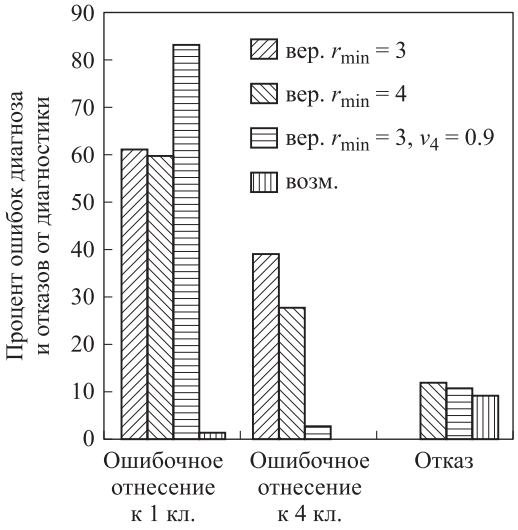

Рис. 2. Результаты диагностики 2-го и 3-го классов ОА вероятностным и возможностным алгоритмами Кора

Кора (вер.) при различных параметрах обучения и возможностным алгоритмом (возм.). Важно заметить, что у возможностного алгоритма нет ошибочных отнесений объектов второго и третьего классов к первому, который является самым опасным проявлением ОА, и ни один больной ОА не отнесен возможностным алгоритмом к группе НД. Такой результат является принципиально важным, поскольку в медицинской практике чрезвычайно опасно отнести пациента, страдающего острым аппендицитом, к группе НД, т. е. исключить у него ОА.

Заключение

В результате проведенных исследований построена вероятностная модель медицинского объекта и вероятностная модель диагностики заболеваний. Разработан алгоритм диагностики заболеваний типа Кора, включающий алгоритм обучения системы компьютерной диагностики по обучающей выборке объектов и алгоритм распознавания — постановку предварительного диагноза пациенту на основании решающего правила,

минимизирующего вероятность потерь, сопутствующих ошибочной классификации.

Построена возможностная модель медицинского объекта, характеризующая нечеткую связь между зарегистрированными у пациента симптомами заболевания и его реальным состоянием (диагнозом), а также возможностная модель диагностики заболеваний. Разработан возможностный (нечеткий) алгоритм Кора. На этапе обучения алгоритма определяются нечеткие представительные наборы значений признаков каждого заболевания. Решающее правило минимизирует возможность сопутствующих постановке данного диагноза потерь, оценивая последствия для здоровья больного разных вариантов ошибочного диагноза.

Результаты вычислительного эксперимента показывают, что применение возможностного алгоритма Кора позволяет провести более детальную классификацию разновидностей острого аппендицита. В вероятностном алгоритме используются такие входные параметры, как пороги по частоте встречаемости представительных наборов, пороговое значение, в пределах которого признаки считаются «неразличимыми», длина представительного набора, и результаты классификации сильно различаются в зависимости от значений этих параметров. Наилучшие результаты диагностики ОА вероятностным алгоритмом Кора получены при длине представительного набора, равного трем признакам ($r_{\min} = 3$), т. е. для оптимальной классификации используются всего три признака, в то время как при постановке диагноза в возможностном алгоритме учитываются значения всех восьми признаков. Таким образом, в результате сравнения вероятностного алгоритма Кора и его нечеткого аналога выявлены и обоснованы теоретические и практические преимущества нечеткого алгоритма при постановке диагноза компьютерной системой.

Работа выполнена при финансовой поддержке РФФИ (гранты 11-07-00338-а, 14-07-00409-а).

Список литературы

1. *Котов Ю.Б.* Новые математические подходы к задачам медицинской диагностики. М., 2011.
2. *Газарян В.А., Иваницкая Н.В., Пытьев Ю.П., Шаховская А.К.* // Вестн. Моск. ун-та. Физ. Астрон. 2003. № 2. С. 12.
3. *Газарян В.А., Илюшин В.Л., Пытьев Ю.П., Шаховская А.К.* // Вестн. Моск. ун-та. Физ. Астрон. 2005. № 4. С. 3.
4. *Газарян В.А., Иваницкая Н.В., Пытьев Ю.П., Шаховская А.К.* // Вестн. Моск. ун-та. Физ. Астрон. 2006. № 6. С. 15.
5. *Газарян В.А., Нагорный Ю.М., Пытьев Ю.П. Шаховская А.К.* // Интеллектуальные системы. 2008. **12**, № 1–4. С. 65.
6. *Пытьев Ю.П.* Возможность как альтернатива вероятности. М., 2007.
7. *Dempster A.P.* // Intern. J. of Approximate Reasoning. 2008. **48**. P. 365.
8. *Dubois D., Prade H.* // Ann. of Math. and Artifical Intelligence. 2001. **32**. P. 35.
9. *Газарян В.А., Матвеева Т.В., Чехонина Ю.Г., Шаховская А.К.* // Интеллектуальные системы. 2010. **14**. № 1–4. С. 107.
10. *Журавлев Ю.И.* // Журн. вычисл. матем. и матем. физ. 2002. **42**, № 9. С. 1425.
11. *Doria A.S.* // Pediatr. Radiol. 2009. **39**, N 2. P. 144.
12. *Williams R.F., Blakely M.L., Fischer P.E.* et al. // J. Am. Coll. Surg. 2009. **208**, N 5. P. 819.

A comparative analysis of the efficiency of probabilistic and possibilistic algorithms for medical diagnostics

Yu. P. Pyt'ev[1], **V. A. Gazaryan**[1,2,a], **P. B. Rosnitskiy**[3]

[1] *Department of Computer Methods of Physics, Faculty of Physics, M. V. Lomonosov Moscow State University, Moscow 119991, Russia.*
[2] *Department «Theory of Probability and Mathematical Statistics», Faculty of Applied Mathematics and Information Technologies, Financial University under the Government of the Russian Federation, Moscow 125993, Russia.*
[3] *Department of Acoustics, Faculty of Physics, M. V. Lomonosov Moscow State University, Moscow 119991, Russia.*
E-mail: [a] *varvaragazaryan@yandex.ru.*

Mathematical methods for pattern recognition and algorithms for the classification of diseases based on them are widely used to solve problems of medical diagnostics [1]. In [2], in order to classify functional disorders of the gastrointestinal tract, an algebraic model of the Kora algorithm was applied. In [3–5] it was shown that to solve many problems of medical diagnostics possibilistic methods for making a medical diagnosis are much more efficient. The present work considers a comparative analysis of probabilistic and possibilistic models of diagnostics, as well as Kora algorithms and the results of their application to solving problems of acute appendicitis diagnostics.

Keywords: pattern recognition, identification problem, probabilistic model of diagnostics, possibilistic model of diagnostics, granulation, Kora classification algorithm, acute appendicitis.
PACS: 02.70.–c, 02.50.Le.
Received 23 January 2014.

English version: *Moscow University Physics Bulletin* 3(2014).

Сведения об авторах

1. Пытьев Юрий Петрович — доктор физ.-мат. наук, зав. кафедрой, профессор; тел.: (495) 939-13-32; e-mail: yuri.pytyev@gmail.com.
2. Газарян Варвара Арамовна — канд. физ.-мат. наук, мл. науч. сотрудник; тел.: (495) 939-41-78; e-mail: varvaragazaryan@yandex.ru.
3. Росницкий Павел Борисович — студент; e-mail: pavrosni@yandex.ru.