Управление в биологических системах и медицине

УДК 62-506.001:616.07

РАЗМЫТЫЕ ПРАВИЛА РАСПОЗНАВАНИЯ В ЗАДАЧАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ И ПРОГНОЗА

СТАДНИК О. Е.

(Обнинск)

Описывается постановка задачи распознавания образов с использованием размытых разбиений для медицинской диагностики и прогноза. Приводятся результаты решения задачи распознавания формы течения и прогнозирования исхода лечения злокачественного новообразования. Показывается эффективность размытых решающих правил.

1. Введение

Важными задачами при управлении процессом лечения больных являются диагностика степени тяжести заболевания и установление индивидуального прогноза исхода лечения. Решив эти задачи, можно максимально индивидуализировать терапевтическую программу еще на стадии ее планирования.

Для решения указанных задач при помощи ЭВМ обычно используют методы распознавания образов [1, 2].

В течение последнего десятилетия в биомедицинских исследованиях, связанных с задачами подобного типа, формируется новый подход, основанный на применении алгоритмов размытой классификации и распознавания [3—5].

Ошибки в медицинской диагностике и прогнозировании обычно связывают с недостаточностью информации о пациенте. Другим, не менее существенным ограничением является стандартное предположение о наличии четких границ между группами. Практически врачи при установлении диагноза или прогноза исходят из того, что группы не имеют четко очерченных границ. В этой ситуации использование методов, связанных со статистической оценкой параметров, оказывается малоэффективным, особенно при работе с выборками малого объема (что нередко бывает при решении биомедицинских задач). Использование же алгоритмов размытой классификации и распознавания позволяет учесть сложность структуры классов объектов биомедицинских систем, нестатистическую неопределенность принадлежности объектов к определенным типам, наличие объектов «промежуточного» характера.

В данной работе представлена попытка разработки метода решения задач диагностики и прогнозирования состояния больных, основанного на использовании алгоритма построения размытого решающего правила. Во втором разделе дается постановка задачи и приводится кратное описание алгоритма. Третий раздел содержит описание результатов его использования в задаче распознавания формы течения и прогнозирования исхода лечения злокачественного новообразования.

2. Постановка задачи

Пусть каждый объект описывается набором из m параметров. Введем в рассмотрение m-мерное пространство параметров R^m, тогда каждому объекту будет соответствовать некоторая точка $x \in R^m$. Введем в рассмотрение перечень k непересекающихся классов $P_1, \ldots, P_k$. При этом пред-

полагается, что классы могут быть упорядочены на некотором содержательно интерпретируемом направлении, если $k>2$. Пусть далее имеется обучающая выборка n объектов $X=\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, о каждом из которых известно, к какому классу он принадлежит, т. е. для каждого $\mathbf{x}\in X$ известна характеристическая функция

$$\omega_l(x)=\begin{cases}1, & \text{если } \mathbf{x}\in P_l,\ l=1,k;\ \mathbf{x}\in X,\\ 0, & \text{если } \mathbf{x}\overline{\in} P_l,\end{cases}$$

При этом обучающая выборка построена так, что в разные классы входят объекты, значительно отличающиеся один от другого в некотором содержательном смысле.

Введем в рассмотрение линейную разделяющую функцию (ЛРФ)

(1) $$F(\alpha, \mathbf{x})=(\alpha, \mathbf{x}),$$

где α — вектор весовых коэффициентов.

Используемый в настоящей работе алгоритм построения размытого решающего правила распознавания [6] осуществляет поиск такого направления (1), которое соответствует наилучшему в некотором смысле разделению объектов множества X на k размытых классов $\tilde{P}_i$, $i=1, k$. Размытые классы образуют размытое k-разбиение множества X, т. е.

$$\sum_{l=1}^{k}\mu_l(x)=1, \quad 1\geqslant\mu_l(x)\geqslant 0; \quad l=1,k; \quad \mathbf{x}\in X.$$

Для получения размытого разбиения в пространстве R^m отыскивается k точек, называемых центрами классов V_i. Величина, обратная расстоянию от данной точки x до центра V_i, отнесенная к сумме величин, обратных расстояниям от этой точки до каждого центра, выбирается за функцию принадлежности точки x к классу i. Центры классов при этом лежат на некотором направлении и выбираются как функция, минимизирующая функционал, описанный в [6]. Там же доказана монотонность итерационного процесса, лежащего в основе алгоритма распознавания на конечном множестве точек.

Алгоритм состоит из двух частей. Первая осуществляет поиск «компактных» групп объектов множества X при каждом фиксированном положении направления (заданного значением вектора α). Эта часть является модификацией алгоритма Fuzzy ISODATA [7], ориентированной на построение размытой классификации одномерных данных. Соответствие полученного размытого разбиения $\tilde{P}_1, \ldots, \tilde{P}_k$ разбиению $P_1, \ldots, P_k$ множества X, заданного характеристической функцией $\omega_l(x)$, $l=1, k$, оценивается при помощи функционала, имеющего смысл суммарной ошибки распознавания объектов обучающей выборки

(2) $$\Phi(\mu, \omega)=\sum_{x\in X}\sum_{l=1}^{k}(1-\mu_l(x))\,\omega_l(x)=\sum_{\mathbf{x}\in X}\Delta\mu(x),$$

где $\Delta\mu(x)$ — ошибка распознавания объекта $\mathbf{x}\in X$.

Во второй части алгоритма производится поиск направления (т. е. вектора α), минимизирующего Φ. Для этого при каждом варьировании вектора α первой частью алгоритма строится размытое разбиение множества X на полученном направлении и оценивается ошибка распознавания по этому направлению. Минимизация Φ и поиск оптимального α^* осуществляется при помощи метода локальных вариаций.

Полученные по алгоритму решающие правила могут использоваться затем для распознавания объектов, не входящих в обучающую выборку X. С этой целью для новой точки $\mathbf{y}\in X$ вычисляется по формуле (1) значение ЛРФ $F(\alpha^*, \mathbf{y})$, а затем определяются веса принадлежности $\mu_l(y)$, $l=1, k$ по формулам, приведенным в [6].

В описанной постановке задачи входная информация для используемого алгоритма — это обучающая выборка заданного размера и принадлеж-

ности объектов заданы значениями характеристической функции, т. е. так же, как в детерминистской постановке задачи распознавания образов. Напротив, результат применения алгоритма для произвольной точки y исследуемого пространства признаков есть набор чисел $\mu_1(y), \ldots, \mu_k(y)$, имеющий смысл оценки степени принадлежности точки к соответствующему классу.

Следует подчеркнуть, что в большинстве алгоритмов распознавания в детерминистской постановке для построения решающего правила используются вспомогательные конструкции — функции $\varphi_1(x), \ldots, \varphi_k(x)$ принадлежности точки x к классам. Окончательное же детерминированное решающее правило — это, в таких алгоритмах, некая логическая процедура принятия решений на базе информации о функциях принадлежности.

Содержательное различие между постановкой задачи размытого распознавания и «неразмытого» (к ним можно отнести как детерминистскую постановку, так и вероятностную) заключается также и в выборе специального критерия экстремизации (2).

Если об объектах из обучающей выборки известна информация $(x, \mu_l(x))$ (т. е. каждый объект описан не только набором m параметров, но и априорным значением функций принадлежности, взамен характеристической функции $\omega_l(x)$), то задачу построения размытого решающего правила можно решать как аппроксимационную, в постановке, аналогичной приведенной в [8]. В этом случае функционал (2) примет вид

$$\Phi(\mu^A, \mu^B) = \sum_{x \in X} \sum_{l=1}^{k} | \mu_l^A(x) - \mu_l^B(x) |,$$

где $\mu_l^A(x)$ — априорное значение принадлежности объекта x к классу l, $\mu_l^B(x)$ — вычисленное значение принадлежности. Алгоритм же при этом не изменится.

3. Распознавание подострой формы лимфогранулематоза

Установление формы течения лимфогранулематоза (ЛГМ) имеет принципиальное значение для управления процессом лечения. Лимфогранулематоз — это злокачественное новообразование лимфатической ткани, протекающее с поражением лимфатических узлов и внутренних органов. Относится к группе гемобластов, занимающих пятое место после основных локализаций рака по частоте встречаемости [9].

Несмотря на существование общепринятых методик лечения, инструкций и рекомендаций, в основном определяющих лечебную тактику, лимфогранулематоз по своим проявлениям, течению, ответу на проводимую терапию является настолько сложным заболеванием, что практически в каждом случае необходимо индивидуализировать лечебные планы, выбирая оптимальную программу (схемы комбинированной терапии, дозы лучевой терапии и т. д.) [10].

Почти все исследователи, занимавшиеся лимфогранулематозом, указывали на различную продолжительность заболевания. Наряду с обычной формой заболевания (при которой срок жизни составляет 5 лет и более) выделяют также подострую форму (срок жизни больных — до 12 месяцев) [11]. Некоторые авторы обычной называли форму, при которой длительность заболевания от первых симптомов до смертельного исхода колеблется в пределах от 2 до 3 лет, свыше 3 лет и т. д. [12, 13]. До сих пор в медицинской литературе нет единого мнения о соответствии длительности заболевания различным формам. Одной из причин трудности получения данных о продолжительности течения ЛГМ является возможность наличия длительного бессимптомного латентного периода. Начало заболевания устанавливается из жалоб больного и является в значительной мере субъективным.

Несмотря на большое количество работ, посвященных прогнозированию течения ЛГМ, в настоящее время диагностика подострого варианта осуществляется, как правило, после констатации неэффективности лечения и летального исхода. Отсутствуют надежные критерии прижизненной диагностики подострого варианта лимфогранулематоза.

Вместе с тем известно, что показатели периферической крови могут указывать на степень поражения организма основным процессом [14]. Поскольку показатели периферической крови косвенно позволяют оценить злокачественность процесса и в то же время указывают на состояние компенсаторных возможностей организма, то они и были выбраны в качестве параметров, описывающих больного для построения решающих правил.

Итак, задача заключалась в том, чтобы для каждого вновь поступающего больного лимфогранулематозом определить по набору исходных гематологических показателей форму заболевания: подострую или обычную. Соответственно нужно было разработать решающие правила для этой задачи. При этом в клинической практике обычно считают, что у больного, прожившего менее 12 месяцев, считая с момента обнаружения заболевания, форма течения ЛГМ является подострой, более 3 лет — обычной. Больных же, проживших более 12, но менее 36 месяцев, вообще «четко» не относят ни к той, ни к другой форме, подразумевая, что это некоторая переходная форма. Таким образом, высказывание о форме течения лимфогранулематоза является нечетким в смысле Л. Заде.

Для выработки критериев распознавания подострой формы ЛГМ был проведен ретроспективный анализ исходных показателей периферической крови у больных III—IV стадией ЛГМ, которые измерялись у больных при поступлении (до начала обследования и лечения). Таким образом, каждый больной описывался 10 исходными гематологическими показателями (содержание гемоглобина, количество эритроцитов, лейкоцитов, тромбоцитов, лейкограмма, скорость оседания эритроцитов).

Данные получены в клинико-диагностической лаборатории Научно-исследовательского института медицинской радиологии Академии медицинских наук СССР [1].

В обучающую выборку вошло 24 больных из числа пациентов, подвергнутых в период с 1972 по 1981 г. диагностической лапаратомии со спленкэктомией с целью морфологической верификации заболевания. В обучающую выборку были включены больные двух классов, существенно отличающихся по продолжительности жизни. В первый класс было включено восемь больных с клинической картиной заболевания, соответствующей подострому варианту с развитием летального исхода на протяжении одного года. Второй класс составили 16 больных, продолжительность жизни которых была не менее пяти лет с момента обнаружения заболевания до летального исхода.

Таким образом, больных из обучающей выборки можно отнести к так называемым «клинически чистым типам» [3]. Поэтому постановка задачи размытого распознавания с введением априорных значений функций принадлежности как характеристических функций $\omega_l(x)$, $l=1, k$, принимающих значения 0 и 1, является содержательно оправданной.

По описанному алгоритму минимизации суммарной ошибки распознавания были получены размытые линейные решающие правила.

Качество распознавания подострого варианта ЛГМ оценивалось как по обучающей выборке (по процедуре скользящего контроля), так и по контрольной группе больных, тремя способами. По первому из них ошибка распознавания фиксировалась в том случае, когда принадлежность $\mu(x)$ объекта x к «своему» классу была ниже порогового значения $\mu^*=0{,}5$. По второму объект считался правильно распознанным, если принадлежность $\mu(x)$ его к «своему» классу была не менее 0,6, неправильно распознанным, если принадлежность к «не своему» классу была не менее 0,6. При невыполнении обоих условий фиксировался отказ от распознавания.

[1] В интерпретации результатов принимал участие кандидат медицинских наук В. В. Павлов.

Третий способ заключался в вычислении относительной ошибки распознавания по формуле

$$S_1 = \frac{\sum_{x \in X} \Delta\mu(x)}{n} 100\%.$$

По обучающей выборке больных, на основании процедуры скользящего контроля, количество ошибок было равно: по первому способу вычисления ошибок — одна ошибка, по второму — без ошибок и три отказа от распознавания, по третьему способу $S_1=12\%$.

Для проверки эффективности полученных размытых решающих правил были проанализированы показатели периферической крови больных из контрольной выборки. Ее составило 63 больных, наблюдавшихся в клинике НИИМР АМН СССР с 1972 по 1982 год. Из них у трех срок жизни был не более 12 месяцев, у 11 — от 12 до 36 месяцев, у 18 — от трех до пяти лет, у остальных — свыше пяти лет. Больные из контрольной выборки со сроком жизни не более 12 месяцев и более пяти лет не были включены в обучающую выборку, так как им не проводилась диагностическая лапаратомия со спленкэктомией. Больным, включенным в обучающую выборку, такая операция проводилась. На основании размытого решающего правила были вычислены принадлежности больных из контрольной выборки к классам «подострая форма лимфогранулематоза» и «обычная форма лимфогранулематоза». При этом было получено, что по первому способу подсчета ошибок неправильно распознанными были шесть пациентов, по второму способу было четыре ошибки и семь отказов от распознавания, по третьему способу $S_1=17\%$.

Для прогнозирования срока жизни больных можно было бы использовать уравнение множественной регрессии. Для построения уравнения регрессии использовалась обучающая выборка из 24 пациентов, тех же, по данным о которых строилось размытое решающее правило. Описывались они тем же набором параметров. Как показали проведенные эксперименты на ЭВМ по построению регрессионной модели, использование этой модели для прогноза срока жизни пациентов из контрольной выборки (тех же, что описаны выше), приводит к средней абсолютной ошибке 29 месяцев. Если принять значение 12 месяцев в качестве порогового при определении формы течения ЛГМ (подострая или обычная), то процент ошибок по контрольной выборке равен 20,6%. Применение модели линейного дискриминантного анализа в этом случае не дало улучшения результатов по сравнению с регрессионным анализом. Таким образом, использование размытых решающих правил позволило получить наиболее точный прогноз формы течения заболевания.

По описанному алгоритму построения размытого решающего правила проводилось также прогнозирование формы течения лимфогранулематоза в постановке, описанной в конце второго раздела настоящей статьи. Переход к заданию априорных функций принадлежности на основании данных о сроке жизни пациентов (не равных характеристическим функциям) позволил включить в обучающую выборку восемь пациентов со сроком жизни от 12 до 60 месяцев, у которых имелась морфологическая верификация заболевания (т. е. им проводилась диагностическая лапаратомия со спленкэктомией).

Априорные значения принадлежности объектов определялись по длительности жизни от момента обнаружения заболевания до летального исхода. Центр класса «подострая форма лимфогранулематоза» в этом случае соответствовал 12 месяцам, центр класса «обычная форма лимфогранулематоза» соответствовал 36 месяцам. Значения априорных функций принадлежности вычислялись по формулам, приведенным в [6].

По описанному алгоритму минимизации суммарной ошибки распознавания были получены размытые решающие правила и по ним вычислены значения принадлежности к двум классам 55 объектов из контрольной

Длительность заболевания, месяцев	0—8	9—12	13—20	21—36	37—60	свыше 60
Число больных	3	8	5	6	18	47
$\overline{\mu_1}$	0,96	0,84	0,65	0,31	0,17	0,11
σ_{μ_1}	0,08	0,21	0,23	0,16	0,14	0,09

выборки. Ошибка распознавания в этом случае вычислялась по формуле

$$S_2 = \frac{\sum_{x \in X} \sum_{l=1}^{2} | \mu_l^A(x) - \mu_l^B(x) |}{2n} 100\%,$$

где $\mu_l{}^A(x)$ — априорное значение принадлежности, $\mu_l{}^B(x)$ — вычисленное по размытому решающему правилу значение принадлежности. Ошибка распознавания объектов из обучающей выборки была равна 11%, а из контрольной выборки — 12,8%. При этом по первому способу подсчета ошибок было пять, по второму их было три и четыре отказа от распознавания.

В таблице приведена гистограмма распределения всех больных по срокам их жизни, а также соответствующие каждому сроку средние значения принадлежности к классу «подострая форма лимфогранулематоза» (вычисленные по размытому решающему правилу) и среднеквадратическое отклонение.

Использование регрессионной модели, построенной по обучающей выборке из 32 больных, дало для всех 87 больных (из обучающей и контрольной выборок) среднюю абсолютную ошибку 23 месяца. При введении порогового значения 12 месяцев для определения формы течения заболевания ЛГМ (подострая или обычная) ошибка была равна 18,3%. Решающие правила, полученные по алгоритмам дискриминантного анализа, дали ошибку 17,2%, что также хуже, чем по размытому решающему правилу.

Таким образом, применение разработанной методики позволяет по значениям исходных гематологических показателей диагностировать подострую форму лимфогранулематоза и прогнозировать выживаемость больных при лечении их по стандартной терапевтической программе.

Разработанная методика может быть рекомендована в качестве дополнительного теста при обследовании больных лимфогранулематозом. На основании полученных данных можно, по-видимому, выделить группу больных с подострым вариантом течения процесса и применить к ним более интенсивную терапевтическую программу, адекватную данной форме течения заболевания.

Полученные результаты показывают эффективность алгоритма построения размытых решающих правил при решении задач диагностики и прогноза.

ЛИТЕРАТУРА

1. Распознавание образов и медицинская диагностика/Под ред. Неймарка Ю. И. М.: Наука, 1972.
2. *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. М.: Наука, 1983.
3. *Woodbury M. A., Clive J.* Clinical Pure Types as a Fuzzy Partition.— J. Cybernetics, 1974, v. 4, № 3, p. 111—121.
4. *Bezdek J. C.* Pattern Recognition with Fuzzy Objective Functions Algorithms. N. Y.— London: Plenum Press, 1981.
5. *Заде Л. А.* Размытые множества и их применение в распознавании образов и кластер-анализе.— В кн.: Классификация и кластер. М.: Мир, 1980, с. 208—247.

6. *Бородкин Л. И., Стадник О. Е.* Алгоритм построения решающего правила в задаче распознавания образов с использованием размытых множеств.– АиТ, 1983, № 10, с. 128–134.
7. *Dunn J. C.* A Fuzzy Relative of the ISODATA Process and it's Use in Detecting Compact Well-Separated Clusters.– J. Cybernetics, 1974, v. 3, № 3, p. 32–57.
8. *Борисов А. Н., Кокле Э. А.* Распознавание размытых образов по признакам.– В кн.: Кибернетика и диагностика. Рига: Зинатне, 1969, вып. 4, с. 135–147.
9. *Переслегин И. А., Филькова Е. М.* Лимфогранулематоз. М.: Медицина, 1980.
10. *Колыгин Б. А.* Лимфогранулематоз у детей. М.: Медицина, 1983.
11. *Павлов В. В., Байсоголов Г. Д.* Эволюция морфологического варианта и прогноз при лимфогранулематозе.– Медицинская радиология, 1983, № 9, с. 6–9.
12. *Успенский А. Е.* Лимфогранулематоз. М.: Медгиз, 1958.
13. *Манкин З. В.* Лимфогранулематоз. М.– Л.: Медгиз, 1938.
14. *Байсоголов Г. Д., Хмелевская З. И., Шишкин И. П.* Прогностическое значение клинических показателей при лимфогранулематозе.– Вестн. Академии медицинских наук СССР, 1978, № 1, с. 54–59.

Поступила в редакцию
22.III.1985

FUZZY RECOGNITION RULES IN MEDICAL DIAGNOSIS AND PREDICTION

STADNIK O. Ye.

For pattern recongnition in medical diagnosis with the use of various decompositions the results in recognition of the course and prediction of the outcome of treating malignant formations are reported. Effectiveness of fuzzy decision rules is demonstrated.