УДК 004.435 + 004.4'423

ПОСТРОЕНИЕ РЕШАЮЩИХ ПРАВИЛ ДЛЯ СИСТЕМ АВТОМАТИЗИРОВАННОГО СКРИНИНГА

В. В. Афанасьева,
заведующая отделением медицинской профилактики
Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова
А. Б. Кубайчук,
начальник отдела медицинских информационных систем
В. В. Шаповалов,
доктор техн. наук, профессор, директор
Федеральное государственное научное учреждение «Научно-исследовательский конструкторско-технологический институт биотехнических систем»

В статье рассматриваются подходы к построению врачебных решающих правил для автоматизированных систем скринирующей диагностики. Основное внимание уделено применению метода дискриминантных функций и методов нечеткой логики в алгоритмах анализа информации.

Methods of medical decision rules building for automated screening systems are considered in this article. The main consideration has been given to using the discriminant functions method and fuzzy logic methods in information analysis algorithms

Решение большого класса диагностических задач сводится к скринингу текущего или прогнозируемого состояния некоторого объекта (системы). Скрининг (в пер. – просеивание) представляет собой быстроосуществимый тест, который обычно имеет характер экспресс-анализа принадлежности объекта скрининга к некоторой группе объектов [1].

Автоматизация процессов скринирующей диагностики на базе новых информационных технологий (в основе которых лежит применение средств вычислительной техники для частичного или полного осуществления процессов сбора, хранения, преобразования и передачи информации) предполагает создание и внедрение автоматизированных систем скринирующей диагностики (АССД).

В АССД медицинского назначения важнейшую роль играют врачебные решающие правила, на основании которых производится оценивание принадлежности состояния объекта скрининга некоторой области пространства его возможных состояний [2].

Для построения врачебных решающих правил при создании АССД могут быть использованы подходы, основанные на применении метода дискриминантных функций и методов нечеткой логики в алгоритмах анализа информации.

Суть первого подхода, математический аппарат которого подробно описан, в частности, в работах [3–5], сводится к построению детерминированной функции $f(\mathbf{x})$, где $\mathbf{x}$ – вектор признаков, принимающий различные значения. В зависимости от значений $f(\mathbf{x})$ принимается та или иная гипотеза.

Различают линейные и нелинейные дискриминантные функции. В линейном случае функция $f(\mathbf{x})$ для n признаков имеет вид

$$f(\mathbf{x}) = \sum_{i=1}^{n} C_i X_i + C_0$$

либо является полиномом более высокого порядка.

В случае построения линейных дискриминантных функций существует довольно большое число различных методов для нахождения коэффициентов ($C_0, C_1, \ldots, C_n$). Простейшим из них является метод наименьших квадратов для нахождения коэффициентов линейной регрессии по обучающей выборке – набору статистических данных о пациентах с известным диагнозом. При этом в случае двухальтернативной ситуации (ставится один из возможных диагнозов – имеется или отсутствует рассматриваемая патология) при наличии одного диагноза соответствующему вектору из обучающей выборки приписывается значение «+1», а в случае другого – «–1». Тогда, если объем обучающей выборки обозначить как m, то коэффици-

енты (C_0, C_1, …, C_n) будут находиться из условия минимизации функционала

$$\sum_{k=1}^{m}\left[(-1)^{x(x^k)} - C_0 - C_1\mathbf{X}_1^k - \ldots - C_n\mathbf{X}_n^k\right]^2,$$

где $\mathbf{X}(\mathbf{X}^k) = 0$, если вектор $\mathbf{X}$ соответствует первому диагнозу, и 1 в противном случае. При этом x_i^k равно 1, если у k-го обследуемого имеется $(i-1)$-й признак, и 0, если его нет.

Если $m > +1$, то условие минимальности функционала сводится к невырожденной системе из $n+1$ линейных уравнений относительно (C_0, C_1, …, C_n), что позволяет легко найти последние. В дальнейшем при использовании построенной таким образом дискриминантной функции будет выноситься тот или иной диагноз в зависимости от знака $f(\mathbf{x})$.

Возможны и иные способы нахождения коэффициентов (C_0, C_1, …, C_n), когда условия оптимальности и соответствующий функционал, их определяющий, выбираются из других соображений: максимизация средневзвешенного расстояния векторов обучающей выборки до разделяющей гиперплоскости (вектор, оказавшийся по «чужую» сторону гиперплоскости, соответственно, уменьшает это расстояние, т. е. входит в функционал, определяющий средневзвешенное расстояние, с обратным знаком) и т. д.

Привлекательность описанного выше подхода в значительной мере объясняется простотой соответствующих вычислительных процедур. Он был применен, в частности, при создании программного комплекса АСПОН-Д и подтвердил свою высокую эффективность на практике, поскольку тщательный анализ обучающих выборок на большом объеме информации позволил достаточно точно оценить значимости симптомов [6].

Однако построение более сложной системы АСПОН-РВ (для детей раннего возраста до 3 лет) потребовало, в связи со значительно более сложными отношениями между параметрами и большим объемом входной информации (более 4000 данных на ребенка), применения более сложных методов анализа информации. В этих условиях был успешно использован описанный далее подход, основанный на применении методов нечеткой логики в алгоритмах анализа информации.

Одно из упрощающих предположений, принимаемых с самого начала, состоит в том, что заранее фиксируется обсуждаемый применительно к данному обследуемому диагноз (профиль патологии), а затем дается ответ на вопрос, в какой мере показатели, характеризующие обследуемого, позволяют сделать вывод о наличии у него данного профиля патологии. В простейшем случае ответ может носить бинарный характер, т. е. «нет» или «да», а в общем случае степень уверенности в наличии у данного обследуемого фиксированного профиля патологии может характеризоваться количественно. При такой схеме заключение по каждому профилю патологии выносится в результате применения специфической для данного профиля процедуры обработки данных обследования без учета возможных корреляций с другими профилями, а полная обработка результатов обследования представляет собой перебор всех профилей патологии с применением соответствующих этим профилям процедур.

Стоит заметить, что такой подход, приемлемый для предварительных профилактических обследований, ориентированных на ограниченное число профилей патологии, не подходит для диагностической системы с большим количеством возможных диагнозов. В этой ситуации система должна сама отбирать в ходе анализа данных обследования ограниченное число возможных диагнозов (профилей патологии), причем не посредством простого перебора всех мыслимых диагнозов (что имело бы неприемлемо высокую трудоемкость), а путем целенаправленного поиска по определенному алгоритму.

Простейший вариант такого алгоритма требует задания древовидной структуры на совокупности профилей (диагнозов) и состоит в последовательном уточнении, т. е. в переходе от некоторой группы диагнозов (профилей патологии) к некоторой подгруппе. При этом для детализации заключения привлекаются данные обследования, которые могли быть не использованы на предыдущих этапах при более грубом анализе. Разумеется, разработка таких непереборных методов целесообразна только в том случае, когда необходимо сократить время компьютерной обработки результатов обследования.

Ограничимся, однако, раздельным рассмотрением различных профилей патологии и в дальнейшем будем считать, что речь идет о некотором фиксированном профиле.

Один из наиболее естественных подходов к решению рассматриваемой задачи состоит в том, что каждому профилю патологии и каждому возможному уровню обобщенного медицинского показателя (ОМП) сопоставляется балльная оценка, отражающая значимость данного уровня ОМП для вынесения заключения о наличии и степени выраженности данного профиля патологии. Затем балльные оценки по всем ОМП, принимаемым в рассмотрение применительно к данному профилю патологии, суммируются, и окончательно классификация обследуемых производится в зависимости от соотношения значений этой суммы и нескольких установленных заранее пороговых значений. Балльные оценки и пороги определяются экспертами и отражают их субъективные представления о связях между ОМП и профилями патологии.

Следует обратить внимание на некоторые проблемы, связанные с описанным подходом.

Первой проблемой можно считать выбор балльных оценок для уровней ОМП применительно к

данному профилю патологии. Сюда же относится вопрос о выборе самих уровней конкретных ОМП и о числе этих уровней.

Вторая проблема связана с вопросом о том, каким образом должна выражаться балльная оценка профиля патологии через балльные оценки отдельных ОМП. Дело в том, что аддитивное выражение, применяющееся в работе [3], не в состоянии передать все представления врача-эксперта о взаимосвязи значений ОМП и данного профиля патологии. Как, например, отразить представление о том, что, скажем, одновременное обнаружение значений некоторой совокупности ОМП уже достаточно для вынесения заключения о наличии данного профиля патологии независимо от значений остальных ОМП?

В обычной аддитивной формуле низкие значения остальных ОМП могут скомпенсировать высокие значения ОМП из указанной совокупности, что не дает возможности по суммарному баллу сделать вывод о наличии данного профиля патологии. Этого можно попытаться избежать посредством перехода от аддитивной модели к линейной, включающей масштабирующие множители при каждом из баллов отдельных ОМП. Для выбора таких множителей необходимо произвести разделение ОМП применительно к данному профилю патологии на группы (незначимые, неспецифические, полуспецифические, специфические), для каждой из которых имеется свой характерный масштаб вклада в результирующую балльную оценку. Заметим также, что при таком подходе трудно выразить представление эксперта о том, что достаточная выраженность какого-то ОМП (или нескольких ОМП) свидетельствует об отсутствии данного профиля патологии. Это приводит к мысли о полезности включения в число функций, характеризующих выраженность данного профиля патологии в зависимости от совокупности значений ОМП, неаддитивных и даже нелинейных функций. Такое расширение выразительных возможностей своей оборотной стороной имеет проблему выбора вида подобной функции (будем называть ее критерием). Эту задачу в большинстве случаев возлагают на эксперта.

Следует отметить, что для линейных критериев при определенных условиях вид критерия может быть определен с помощью методов математической статистики (в первую очередь, регрессионного и факторного анализа). Для этого необходимы достаточно репрезентативные сведения о наблюдавшихся у обследуемых ОМП и о наличии у них заболеваний, относящихся к данному профилю патологии (сведения по профилю патологии должны быть получены в результате дополнительного специализированного обследования).

Кроме того, существует проблема интерпретации значений критериев, т. е. определения критических порогов и заключений о результатах обследования, соотносимых с интервалами между соседними порогами.

Далее в общих чертах описан подход, который можно использовать для построения диагностических критериев и который определенным образом позволяет решить указанные выше проблемы.

В качестве средства описания критериев выбрана теория нечетких множеств [6]. Характерными особенностями этого аппарата являются, с одной стороны, почти полный параллелизм с аппаратом классической двузначной логики, а с другой стороны – возможность формализовать представления о степени выраженности того или иного качества (признака). Первое свойство делает этот аппарат удобным для формализации высказываний, сделанных с помощью обычного языка, например представлений экспертов, второе же дает возможность введения достаточно тонких градаций в высказываемые суждения. Данное положение можно проиллюстрировать следующим примером.

Заключение о наличии некоего признака по значениям ряда других признаков может быть выражено в терминах классической логики высказываний в виде формул типа

$$A = (B \wedge \hat{C}) \vee (D \wedge E).$$

Эту формулу можно интерпретировать следующим образом: профиль патологии A приписывается обследуемому тогда, когда у него имеется признак B и отсутствует признак C или имеются признаки D и E. При этом A, B, C, D и E нужно понимать как высказывания об обследуемом. Выделяется некоторое количество элементарных высказываний, а остальные образуются из элементарных с помощью логических связок.

Следует помнить, что каждый признак либо отсутствует, либо присутствует, причем какие-либо промежуточные градации исключаются. Двузначность становится наглядной на так называемом модельном уровне, когда фиксируется некоторое множество S (называемое универсальным), а высказывания понимаются как его подмножества.

В интересующих нас ситуациях S можно понимать как совокупность всевозможных наборов данных обследования. Если при обследовании проверяются N бинарных признаков, то результат одного обследования может быть представлен вектором длины N из нулей и единиц, при этом показатель, соответствующий признаку A, будет стоять на i-м месте ($N \geq i \geq 1$). Объединение всех значений признака A, имеющихся в S, образует множество A, являющееся подмножеством множества S.

Все операции над обычными множествами имеют свои аналоги среди операций над нечеткими множествами, но у одной операции может оказаться несколько аналогов, и ни при каком выборе этих аналогов нельзя добиться того, чтобы набор операций над нечеткими множествами обладал всеми свойствами операций над обычными. Поэтому не-

обходима осторожность при экстраполяции обычных представлений.

Наиболее совершенной в этом плане является формальная система, включающая операции $\cap$, $\cup$, $-$, константы $\varnothing$ и S [7]. Результаты этих операций над функциями принадлежности выглядят так:

$$f_{(A\cup B)} = f_A \cup f_B;$$
$$f_{(A\cap B)} = f_A \cap f_B;$$
$$\hat{}f_A = f_S - f_A,$$

где $\cap$, $\cup$ обозначают взятие минимума и максимума соответственно, а функция f_S тождественно равна 1. С учетом этих определений операции над нечеткими множествами обладают следующими свойствами: коммуникативностью, ассоциативностью, идемпотентностью, дистрибутивностью, инволютивностью и справедливостью теоремы де Моргана.

Формулы, построенные из некоторых базисных нечетких множеств с помощью введенных операций, наиболее понятны с точки зрения классической логики. Свойства, перечисленные выше, позволяют преобразовать формулы к виду, наиболее удобному для вычислений. Функции принадлежности можно задавать аналитически, например дробно-линейной функцией, показательной и пр. Альтернативный способ задания – перечислительный. Например, можно описать функцию принадлежности нечеткого множества формулой

$$f = (\text{ГП}_1{:}0),\ (\text{ГП}_2{:}0{,}25),\ \ldots,\ (\text{ГП}_n{:}0{,}95),$$

где ГП_n – функция принадлежности для n-й группы признаков, а число, стоящее за ГП_n, есть соответствующее значение функции принадлежности.

Введенные ранее выразительные возможности при конструировании формул, выражающих решающие правила, недостаточны, и их можно расширить введением дополнительных операций. Однако при этом структура формул и их смысл становятся менее ясными и у эксперта, составляющего формулу, появляется чрезмерное число степеней свободы, которыми он затрудняется распорядиться. Например, можно ввести две дополнительные операции: алгебраическое произведение $A \otimes B$ и алгебраическую сумму $A \oplus B$:

$$f_{A\otimes B} = f_A \otimes f_B;\ f_{A\oplus B} = f_A + f_B - f_A \otimes f_B,\ f_A^{\wedge} = 1 - f_A.$$

Эти операции удовлетворяют равенствам де Моргана.

Каждая из этих операций, кроме того, ассоциативна. Свойства дистрибутивности для них не имеют места. Для введенных операций выполняются тождества

$$A \otimes \varnothing = 0,\ A \oplus \varnothing = A,\ A \otimes S = A,\ A \oplus S = S.$$

С помощью операции $\oplus$ можно, например, учитывать уровень «фона», определяемого ОМП, не являющимися максимальными.

Для нечетких множеств нет естественного аналога операции импликации, имеющей теоретико-множественное представление $A \to B \equiv \hat{}A \cup B$. Однако для связанного с этой операцией отношения мажорирования $A \le B$ или $f_A \le f_B$ существует нечеткий аналог. Таким образом, на нечетких подмножествах данного универсального множества имеется структура дистрибутивной решетки.

Для применения рассмотренного подхода в интересах обработки данных необходимо задание некоторой функции принадлежности, соответствующей заданному профилю патологии, на множестве возможных наборов результатов обследования. Поскольку такой набор состоит из множества различных компонентов (ответы на вопросы анкеты родителей, данные врачебного осмотра, инструментальные и лабораторные данные и т. д.), непосредственное задание функции принадлежности для всех таких наборов невозможно.

Для этого нужно сначала связать некоторую функцию принадлежности с каждым компонентом данных обследования. Такая функция сопоставляет каждому возможному значению некоторого фиксированного элемента данных обследования числовую характеристику, лежащую в пределах от 0 до 1. Указанные функции принадлежности (которые можно назвать первичными) играют роль атомов в формулах нечеткой логики, выражающих решающие правила диагностики. Однако с их введением появляются только средства, отражающие степень выраженности отдельных диагностических признаков, но нет средств, дающих количественные характеристики парам типа «признак – диагноз» с точки зрения их влияния на выносимое заключение. В линейной статистике подобную роль играют, например, коэффициенты корреляции (регрессии).

В подходах, основанных на формулах логического типа, не всегда оказывается возможным найти выразительные средства для нетривиального отражения подобных отношений. Описанный выше набор операций нечеткой логики оказывается достаточным для решения подобных задач.

Предполагается, что всякий критерий может быть представлен в виде формулы, записанной с помощью операций нечеткой логики. Такая формула выражает функцию принадлежности данного профиля патологии (диагноза) через функции принадлежности ОМП. При этом заранее накладываются некоторые структурные ограничения на вид рассматриваемых формул (свои – для задачи определения профиля патологии и свои – для постановки диагноза). Выбор конкретной формулы, построенной с учетом структурных ограничений и применением только операций «и», «или» и «не», производится врачом-экспертом. Далее эксперт-

ным путем определяются дополнительные связи между профилями патологии (диагнозами) и признаками (симптомами) типа отношения специфичности и т. д. С учетом этой дополнительной информации диагностическая формула преобразуется, для чего используется операция алгебраического произведения ⊗ (см. выше). Это преобразование производится чисто механически, без участия врача-эксперта. В результате формула приобретает свой окончательный вид. Для ее применения необходимо сопоставить каждому ОМП также некоторую соответствующую формулу принадлежности.

В заключение следует отметить, что описанные подходы позволяют сократить усилия, необходимые для построения решающих правил, более целесообразным способом организовать процесс создания АССД. В то же время в силу наличия эвристической составляющей в создании врачебных правил роль врача-эксперта остается крайне важной – в конечном итоге именно от результатов его деятельности будет зависеть медицинская эффективность созданной АССД.

Литература

1. **Шаповалов В. В., Шерстюк Ю. М.** Автоматизированный скрининг – проблема экспертных знаний // Инновации. 2003. № 10 (67). С. 89–91.
2. **Шаповалов В. В., Шерстюк Ю. М.** Формальная модель автоматизированной системы скринирующей диагностики здоровья населения // Информационные технологии в здравоохранении. 2001. № 8–9. С. 8–10.
3. **Вапник В. Н., Червоненкис А. Я.** Теория распознавания образов. М.: Наука, 1974. 415 с.
4. **Ким Дт. О., Мьюлер Ч. У.** Факторный, дискриминантный и кластерный анализ: Пер. с англ. М.: Финансы и статистика, 1989. 215 с.
5. **Афифи А., Эйзен С.** Статистический анализ; подход с использованием ЭВМ: Пер. с англ. М.: Мир, 1982. 488 с.
6. **Бернштейн Л. С., Коровин С. Я., Мелихов А. Н.** Ситуационные советующие системы с нечеткой логикой. М.: Наука, 1990. 272 с.
7. **Кофман А.** Введение в теорию нечетких множеств: Пер. с англ. М.: Радио и связь, 1982. 432 с.

УДК 621.391
ББК 32.811
В74

В74 Вопросы передачи и защиты информации: Сборник статей / СПбГУАП. СПб., 2006. 226 с.: ил.
ISBN 5-8088-0168-0

Предлагаемый сборник статей посвящен вопросам создания безопасных информационных технологий. Само понятие «безопасные технологии» рассматривается здесь в самом широком смысле: технологии обеспечения надежной передачи и хранения информации, защиты информации от несанкционированного доступа, построения эффективных сетевых протоколов.

Темы статей фокусируются, в основном, на двух направлениях исследования: методов повышения достоверности передачи информации и систем защиты информации на основе открытых (публичных) ключей. Большинство статей объединяет использование идей и методов теории помехоустойчивого кодирования.

Сборник будет полезен для специалистов и студентов, интересующихся практикой использования кодов, исправляющих ошибки.

По вопросам приобретения книги обращаться по адресу:
190000, Санкт-Петербург, Б. Морская ул., д. 67, ГУАП,
Фундаментальная библиотека
Телефон: (812) 710-66-42
Факс: (812) 313-70-18
E-mail: ius@aanet.ru