УДК 681.3

РАЗРАБОТКА ЛОГИЧЕСКИХ МОДЕЛЕЙ ПОСТАНОВКИ ДИАГНОЗА ЗАБОЛЕВАНИЙ МОЛОЧНЫХ ЖЕЛЕЗ С ПОМОЩЬЮ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ И ДЕРЕВА РЕШЕНИЙ

Ю.Е. Сумина, И.Я. Львович

Статья отражает разработанные модели рационального принятия решения для диагностики заболеваний молочных желез с целью повышения эффективности процесса диагностики данных патологических состояний

Ключевые слова: заболевания молочных желез, алгоритм, модель, диагностика, дерево решений, сеть Петри

В настоящее время проблема выявления заболеваний молочных желез является достаточно актуальной. На сегодняшний день рак молочной железы (МЖ) занимает одно из первых мест в списке распространенных опухолевых заболеваний. И важно не просто диагностировать патологический процесс, но и определить его на ранней стадии. Общепризнанно, что рак молочной железы встречается в 3-5 раз чаще на фоне доброкачественных заболеваний молочных желез и в 30-40 раз чаще при узловых формах мастопатии с явлениями пролиферации эпителия молочных желез.

Процесс управления постановкой диагноза заболеваний МЖ в современных условиях невозможен без привлечения математического моделирования. При этом одним из важнейших направлений является имитационное моделирование (ИМ).

Увеличение числа методов исследования заболеваний МЖ выдвигает требование системного подхода к диагностике заболеваний МЖ. Алгоритм процесса диагностики состоит из взаимосвязанных событий – опроса пациента (выявление жалоб, данных анамнеза заболевания), процедур осмотра для выявления общего и локального статуса, проведения общих и специальных методик.

В связи с этим предлагается имитационная сетевая модель рассматриваемой задачи диагностики заболеваний МЖ, в которой причинно-следственная связь описывается при помощи сети Петри (СП).

Рассмотрим подробно структуру данной сети. Структура СП процесса диагностики заболеваний МЖ может быть описана формально в виде пятерки $<B, D, I, O, M>$, где B – множество условий (позиций) – методов, используемых в ходе исследования, D – множество классов правил – правил применения той или иной методики, I – множество входных функций, O – множество выходных функций, M – маркировка СП.

$$B = <b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}, b_{11}, b_{12}, b_{13}>,$$

$$D = <d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, d_{17}, d_{18}, d_{19}, d_{20}>.$$

Под переходами будем понимать события, происходящие в лечебно-диагностической системе (выполнение диагностических, аналитических и вспомогательных операций), а также события, обладающие фиксированной продолжительностью, соответственно, позиции будут представлять собой условия, выполнение которых влечет за собой реализацию переходов.

СП рассматриваемой задачи диагностики заболеваний МЖ представлена на рис. 1. Функциональные назначения позиций указаны в табл. 1.

Таблица 1

Функциональные назначения позиций

Наименование позиции	Функциональное значение позиции
b0	Начало процесса диагностики
b1	Бимануальное обследование признаков заболеваний
b2	Лабораторно-клинические исследования признаков заболеваний
b3	Ультразвуковое исследование
b4	Предварительный анализ результатов
b5	Рентгеномаммография
b6	Дуктография
b7	Пневмокистография
b8	Компьютерная томография
b9	Термография
b10	СВЧ-радиометрия
b11	Пункциональная биопсия
b12	Цитологический и морфологический методы
b13	Анализ симптоматики и постановка диагноза

Сумина Юлия Евгеньевна – ВГТУ, аспирант, тел. (4732) 46-76-99

Львович Игорь Яковлевич – ВГТУ, д-р техн. наук, профессор, тел. (4732) 46-76-99

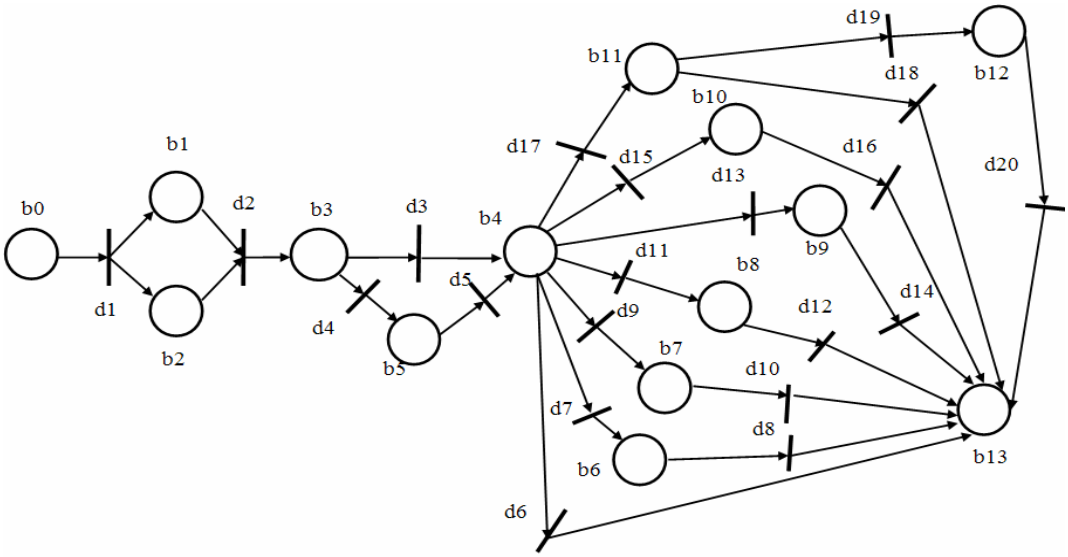

Рис. 1. Сетевая имитационная модель диагностики заболеваний МЖ

Для представления динамических свойств объекта вводится функция маркировки М. При начальной разметке схемы M_1 = {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0} единственным готовым к срабатыванию является переход d_1, срабатывание которого ведет к смене разметки $M_1 \vdash^{d1} M_2$, где M_2 = {0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}. Далее срабатывает переход d_2, что приводит к новой маркировке M_3 = {0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}. Построение сети организовано таким образом, что какие бы переходы не срабатывали, в итоге получится маркировка M_n = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1}, то есть последним будет событие b_{13} – постановка диагноза заболевания молочной железы.

Разработанная модификация сети Петри и правила ее функционирования и позволяют: проводить формирование функциональной модели системы диагностики; отслеживать текущее состояние системы диагностики; проводить генерацию вариантов управления путем имитации.

Помимо управления процессом постановки диагноза в современных условиях важным аспектом является эффективная диагностика заболеваний МЖ и определение диагноза на ранних стадиях заболевания. Применение систем анализа на основе деревьев решений дает эффективный и качественный результат для диагностики заболеваний МЖ.

Системы анализа на основе деревьев решений предоставляют широкие возможности, которые позволяют свести исходную матрицу данных X к набору простых правил, представленных в виде иерархической структуры – дерева. Этот метод моделирования сочетает мощный аналитический аппарат генерации решений с простотой использования технологии и интуитивно понятными конечными результатами.

Деревья решений представляют собой последовательные иерархические структуры, состоящие из узлов, которые содержат правила, т.е. логические конструкции вида "если … то …". Конечными узлами дерева являются "листья", соответствующие найденным решениям и объединяющие некоторое количество объектов классифицируемой выборки.

На сегодняшний день существует значительное число алгоритмов, реализующих построение деревьев решений, среди которых распространение и популярность получил алгоритм С4.5 – алгоритм построения дерева решений с неограниченным количеством потомков у узла, разработанный Р. Куинленом. Этот алгоритм не умеет работать с непрерывным целевым полем, поэтому решает только задачи классификации.

Пусть в некотором узле дерева сконцентрировано некоторое множество примеров X^*, $X^* \subset X$. Тогда существуют три возможные ситуации.

1. Множество X^* содержит один или более примеров, относящихся к одному классу y_k. Тогда дерево решений для X^* – это "лист", определяющий класс y_k .

2. Множество X^* не содержит ни одного примера, т.е. представляет пустое множество. Тогда это снова "лист", и класс, ассоциированный с "листом", выбирается из другого множества, отличного от X^* (скажем, из множества, ассоциированного с родителем).

3. Множество X^* содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество X^* на некоторые

подмножества. Для этого выбирается один из признаков *j*, имеющий два и более отличных друг от друга значений и X^* разбивается на новые подмножества, где каждое подмножество содержит все примеры, имеющие определенный диапазон значений выбранного признака. Это процедура будет рекурсивно продолжаться до тех пор, пока любое подмножество X^* не будет состоять из примеров, относящихся к одному и тому же классу.

Для построения дерева с одномерным ветвлением, находясь на каждом внутреннем узле, необходимо найти такое условие проверки, связанное с одной из переменных *j*, которое бы разбивало множество, ассоциированное с этим узлом на подмножества. Общее правило для выбора опорного признака можно сформулировать следующим образом: “выбранный признак должен разбить множество X^* так, чтобы получаемые в итоге подмножества X^*_k, $k = 1, 2, ..., p$, состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, т.е. количество чужеродных объектов из других классов в каждом из этих множеств было как можно меньше”. Р. Куинленом был предложен теоретико-информационный критерий:

$$T(j) = H(X^*) - \sum_{k=1}^{p} \frac{\left|X^*_k\right|}{\left|X^*\right|} * H\left(X^*_k\right) \Rightarrow \max \forall j = 1,2,...,m,$$

где $H(X^*)$ и $H(X^*_k)$ – энтропия подмножеств, разбитых на классы, рассчитанная по формуле Шеннона.

Очень часто алгоритмы построения деревьев решений дают сложные деревья, которые имеют много узлов и ветвей. Такие "ветвистые" деревья очень трудно понять, а ценность правила, справедливого скажем для 1-3 объектов, крайне низка и в целях анализа данных практически непригодно. Гораздо предпочтительнее иметь дерево, состоящее из малого количества узлов, не вполне идеально классифицирующее обучающую выборку, но обладающее способностью столь же хорошо прогнозировать результат для тестовой выборки. Для решения вышеописанной проблемы часто применяется так называемое "отсечение ветвей", которое происходит снизу вверх, двигаясь с листьев дерева, отмечая узлы как листья, либо заменяя их поддеревом. Если под точностью дерева решений понимается отношение правильно классифицированных объектов, то нужно отсечь или заменить поддеревом те ветви, которые не приведут к возрастанию ошибки.

После индукции дерева решений его можно использовать для распознавания класса нового объекта. Обход дерева решений начинается с корня дерева. На каждом внутреннем узле проверяется значение объекта X_m по атрибуту, который соответствует алгоритму проверки в данном узле, и, в зависимости от полученного ответа, находится соответствующее ветвление, и по этой дуге осуществляется движение к узлу, находящемуся на уровень ниже и т.д. Обход дерева заканчивается, как только встретится узел решения, который и дает название класса объекта X_m.

Целью описанных ниже исследований явилось построение и анализ дерева решений для диагностики мастопатии и фиброаденомы молочных желез на основании 57 историй больных заболеваниями молочных желез и неподтвержденным диагнозом. В матрице обучающей информации содержатся признаки наличия или отсутствия 8 основных симптомов заболеваний молочных желез и группировочный признак, указывающий к какой группе относится больная (соответственно – мастопатия, фиброаденома или норма). Для построения дерева решений использовался алгоритм C4.5. Дерево решений было построено в системе Deductor.

Полученная древовидная схема, представленная на рис. 2, очень наглядна и удобна для анализа.

С помощью визуализатора определяется насколько сильно выходное поле зависит от каждого из входных факторов.

Каждому входному атрибуту соответствует значимость (таблица)- степень зависимости выходного поля от этого атрибута, представленная в табл. 2. Параметр значимость тем больше, чем больший вклад вносит конкретный входной атрибут при классификации выходного поля. Фактически данный визуализатор показывает степень нелинейной зависимости между выходным и входными полями.

Таблица 2

Значимость признаков

Атрибут	Значимость, %
Возраст	26,514
Утолщение слоя железистой ткани	19,798
Плотная консистенция образования	18,432
Фиброзные изменения	9,025
Наличие участков затемнения	7,261
Наличие уплотнения	6,595
Наличие одн//множественных теней	6,467
Болезненность при пальпации	5,736
Боли в МЖ	0,173

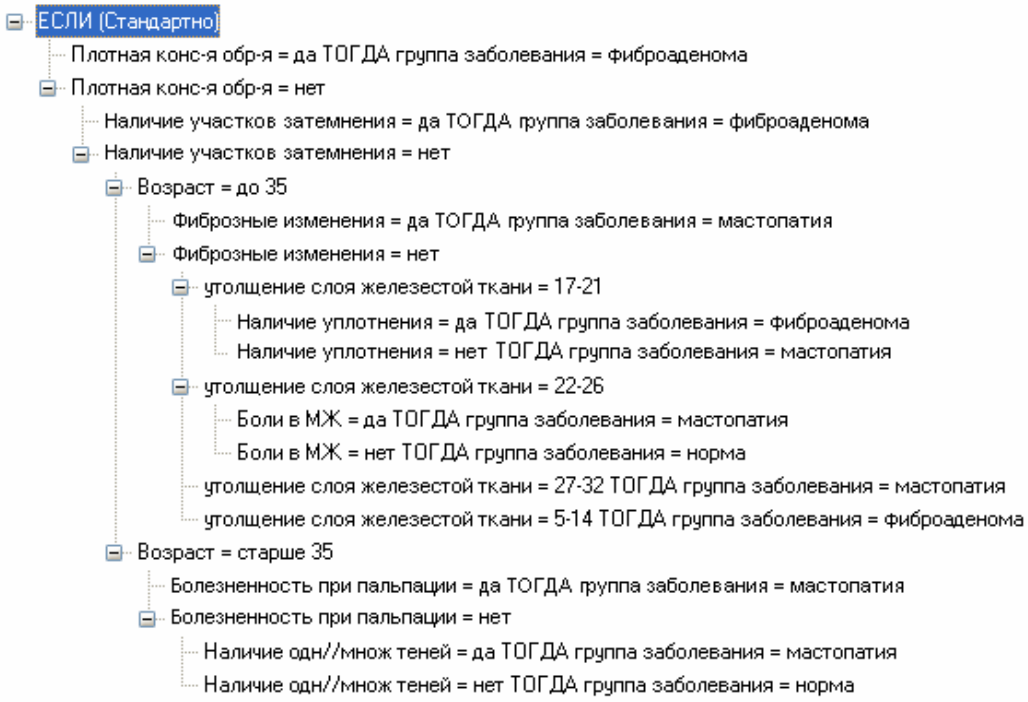

Рис. 2. Дерево решений

Достоверность данного метода анализа составляет 85 %, которая была подтверждена путем проверки тестовой выборки из 19 историй больных с заболеваниями мастопатией, фиброаденомой и неподтвержденным диагнозом.

Литература

1. Нейштадт Э.Л. Патология молочной железы / Э.Л. Нейтадт, О.А. Воробьева. – СПб.: Фолиант, 2003. – 208 с.

2. Запорожцева Ю.Е., Львович И.Я., Новикова Е.И. Клинические характеристики больных с заболеваниями молочных желез // Моделирование и управление процессами в здравоохранении: межвуз. сб. науч. тр. Воронеж: ВГТУ, 2009. С. 19-22.

3. Запорожцева Ю.Е., Львович И.Я., Новикова Е.И. Анализ современных методов диагностики мастопатии / Ю.Е. Запорожцева, И.Я. Львович, Е.И. Новикова // Управление процессами диагностики и лечения: межвуз. сб. науч. тр. Воронеж: ВГТУ, 2008. С. 34-37.

Воронежский государственный технический университет

DEVELOPMENT OF LOGICAL MODELS DIAGNOSIS BREAST DISEASES WITH SIMULATION AND DECISION TREE

J.E. Sumina, I.Yj. Lyvovich

The article reflects the developed model of rational decision making for diagnosis of diseases of dairy same-climbing to improve the efficiency of the diagnostic information of pathological conditions

Key words: diseases mammary glands, algorithms, models, diagnostics, decision tree, the Petrinet