

Витяев Е.Е., Методика извлечения знаний из эксперта¹
Ковалерчук Б.Я.

В работе приводится методика извлечения знаний из эксперта, основанная на свойстве монотонности. Эта методика позволяет извлечь из эксперта Булеву функцию принятия решений и переписать её в виде системы правил. Методика использует свойство монотонности, что сильно сокращает количество вопросов, задаваемых эксперту, и тем самым дает возможность извлекать сложные булевы функции знаний за приемлемое время. Даная методика иллюстрируется примером извлечения знаний из эксперта James Ruiz при создании диагностической системы рака груди для Baton Rouge, (Louisiana), Women hospital.

Ключевые слова: извлечение знаний из эксперта, экспертные системы, базы знаний, экспертные оценки.

1. Проблема извлечения знаний из эксперта

В данной работе описывается методика извлечения знаний из эксперта. Эти знания, представленные множеством правил, могут служить ядром компьютерной диагностической системы. Разработанная методика [Kovalerchuk B., Vityaev E., 1997, 2000, 2001] основана на свойстве монотонности. Преимущества методики показаны на примере компьютерной системы диагностики рака груди.

В США рак груди – наиболее часто встречаемый женский рак [Wingo P.A.]. Наиболее эффективный метод борьбы против рака груди – скрининг маммограмм. Однако было обнаружено, что есть значительная интра- и интернаблюдателя вариабельность маммографической интерпретации (до 25 %). Дополнительно, несколько ретроспективных исследований нашли, что ошибка варьируется в пределах от 20 до 43 %. Эти данные ясно демонстрируют потребность улучшить надежность маммографической интерпретации.

Архивы маммографии в больницах во всем мире содержат миллионы результатов биопсии и маммограмм. Несколько университетов и больниц создали базы данных изображений маммографии, которые являются доступными в Интернете. Такие усилия обеспечивают возможность масштабного анализа данных и извлечения знаний в области диагностики рака груди.

Обнаружение полного множества экспертных правил – экспоненциально сложная задача. Полный опрос эксперта может потребовать задания тысячи вопросов эксперту. Это известная проблема при разработке экспертных систем. Например, для 11 бинарных диагностических признаков мы получаем ($2^{11} = 2048$) комбинаций признаков, каждый из которых может дать отдельное правило. Лобовой метод потребовал бы опроса эксперта для каждой из этих 2048 комбинаций.

2. Иерархический подход.

Извлечение знаний из эксперта основано на оригинальном методе восстановления Булевых функций с использованием свойства монотонности [Kovalerchuk B., 1996]. Мы будем иллюстрировать метод на примере диагностической системы рака груди, но специфика задачи практически не будет сказываться на общности метода.

Если попросить эксперта оценить конкретный случай, представленный набором значений признаков, то типичный вопрос будет иметь следующий вид:

«Если признак 1 имеет значение V_1 , признак 2, имеет значение V_2 ..., признак n имеет значение V_n , то соответствует ли упомянутый набор значений признаков случаю подозрительному к раку или нет? ».

Каждый набор признаков (V_1, V_2, \dots, V_n) представляет возможный клинический случай.

Первая задача состоит в том, что бы свести все признаки к бинарным признакам, разбив их значения на два класса – связанных с подозрением на рак и нет.

¹ Работа поддержана грантом РФФИ 08-07-00272-а; интеграционными проектами СО РАН №1, 115, а также Госконтрактом 2007-4-1.4-00-04 и Советом по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-335.2008.1).

Вторая задача состоит в том, что бы построить иерархию признаков, начиная с общих признаков и кончая менее общими признаками. Эта иерархия начинается с определения 11 медицинских первичных бинарных признаков.

Медик-эксперт обнаружил, что первичные 11 бинарных признаков $w_1, w_2, w_3, y_1, y_2, y_3, y_4, y_5, x_3, x_4, x_5$ могут быть представлены иерархией с добавлением двух новых обобщенных признаков x_1 и x_2 :

Уровень 1 (5 признаков)		Уровень 2 (все 11 признаков)
x_1	–	w_1, w_2, w_3
x_2	–	y_1, y_2, y_3, y_4, y_5
x_3	–	x_3
x_4	–	x_4
x_5	–	x_5

Мы рассматриваем пять бинарных признаков x_1, x_2, x_3, x_4 и x_5 , на уровне 1.

Новый обобщенный признак: x_1 – «Количество и объем кальцинозов» со значениями (0 – «доброкачественный» и 1 – «рак») обобщает признаки связанные с количеством и объемом кальцинозов:

- w_1 – количество кальцинозов / см^3 ;
- w_2 – объем кальциноза, см^3 ;
- w_3 – общее количество кальцинозов.

Мы рассматриваем признак x_1 как функцию $v(w_1, w_2, w_3)$, которую надо определить.

Аналогично, новый признак: x_2 – «Форма и плотность кальциноза» со значениями ((1) – «рак» и (0) – «доброкачественная») является обобщением признаков:

- y_1 – «нерегулярность в форме индивидуальных кальцинозов»,
- y_2 – «изменение в форме кальцинозов»,
- y_3 – «изменение в размере кальцинозов»,
- y_4 – «изменение в плотности кальцинозов»,
- y_5 – «плотность кальцинозов».

Мы рассматриваем x_2 как функцию $x_2 = \psi(y_1, y_2, y_3, y_4, y_5)$, которая должна быть определена для диагностики рака.

В результате мы получим декомпозицию задачи определения булевой функции $f(x_1, x_2, x_3, x_4, x_5)$ как это представлено на рис. 1.

Будем предполагать, что следующие признаки имеют бинарные значения 0 – «доброкачественный», 1 – «рак» :

- x_1 – «количество и объем, занятый кальцинозами»;
- x_2 – «форма и плотность кальцинозов»;
- x_3 – «ориентация протоков»;
- x_4 – «сравнение с предыдущей экспертизой»;
- x_5 – «ассоциированные результаты исследования».

3. Свойство монотонности

Чтобы понять, как монотонность может быть использована для диагностики рака груди, рассмотрим оценку кальцинозов в маммограмме. Используя данные выше определения, мы можем представить клинические случаи в терминах бинарных векторов с пятью обобщенными признаками (x_1, x_2, x_3, x_4, x_5). Рассмотрим два клинических случая, которые представлены двумя двоичными последовательностями: (10110) и (10100). Если радиолог правильно диагностировал

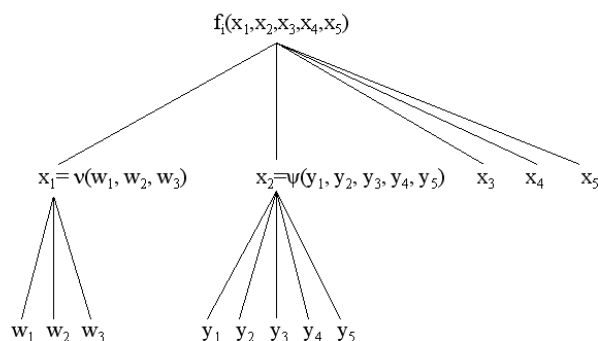


Рис. 1

набор (10100) как злокачественный, то, используя свойство монотонности, мы можем также заключить, что клинический случай (10110) должен также быть злокачественным.

Это заключение основано на кодировании всех признаков «подозрительных на рак» как 1. Заметим, что в (10100) мы имели два показания для рака:

$x_1 = 1$ (количество и объем кальцинозов со значением 1; указание на рак);

$x_3 = 1$ (протоковая ориентация, имеющая значение 1; подозрительна на рак).

Во втором клиническом случае (10110) мы имеем эти два наблюдения для рака и также $x_4 = 1$ (сравнение с предыдущими экспертизами, подозрительными на рак). Аналогично, если мы знаем, что (01010) не подозрительно на рак, то и случай (01000) также нельзя считать подозрительным. Это верно, потому что во втором случае мы имеем меньше признаков, указывающих на наличие рака. Вышеупомянутые соображения – существо того, на чем основан метод.

Медику-эксперту было объяснено свойство монотонности и диалог, который следовал, подтверждал законность предположения. Точно так же функция $x_2 = \psi(u_1, u_2, u_3, u_4, u_5)$ для обобщенного признака x_2 «форма и плотность кальциноза» была подтверждена как монотонная Булева функция.

Булева функция – компактное представление набора диагностических правил. Булева дискриминантная функция может быть представлена в форме множества правил «если–то», но необязательно, чтобы эти правила означали дерево, как в методе решающих деревьев. Булева функция может дать диагностическую дискриминантную функцию, которая не может быть получена методом решающих деревьев.

Таким образом, основными шагами извлечения правил из медика-эксперта являются следующие:

- разработать иерархию понятий и представить их как ряд монотонных Булевых функций;
- восстановить каждую из этих функций минимальной последовательностью вопросов эксперту;
- объединить обнаруженные функции в полную диагностическую функцию;
- представить полную функцию как традиционный набор диагностических правил вида «если–то».

Опишем далее шаг восстановления каждой монотонной Булевой функции с помощью минимального количества вопросов к эксперту. Это предусматривает интервьюирование эксперта с помощью минимальной динамической последовательностью вопросов. Эта последовательность основана на фундаментальной лемме Hansel [Hansel G., 1966, Kovalerchuk B., Talianski V., 1996]. Мы опускаем детальное описание математических шагов. Они могут быть найдены в [Kovalerchuk B., Talianski V., 1996]. Общая идея дается на примере интерактивной процедуры в табл. 1. Минимальная последовательность вопросов означает, что мы достигаем минимума Шенноновской функции, т. е. за минимальное количество вопросов мы можем восстановить самую сложную монотонную Булеву функцию с n аргументами. Последовательность вопросов не написана заранее. Она зависит от предыдущих ответов эксперта, поэтому каждый последующий вопрос определяется динамически. Таблица 1 иллюстрирует эту последовательность.

Столбцы 2 и 3 представляют собой значения определенных выше функций f и ψ . Мы опускаем восстановление функции $V(w_1, w_2, w_3)$, потому что нужно немного вопросов для восстановления этой функции, но общая схема – та же самая, что и для функций f и ψ и начинается с рассмотрения всех бинарных наборов троек (010), (110).

В таблице первый вопрос: «Представляет ли последовательность (01100) случай, подозрительный на рак или нет?» Здесь, $x_1 = 0$ и $(01100) = (x_1, x_2, x_3, x_4, x_5)$. Если ответ «да» (1), то следующий вопрос будет о подозрительности на рак случая (01010). Если ответ «нет» (0), то следующий вопрос будет о подозрительности на рак случая (11100). Эта последовательность вопросов не случайна. Как было упомянуто выше, это выведено из леммы Hansel [Hansel G., 1966]. Все 32 возможных случая для пяти бинарных признаков $(x_1, x_2, x_3, x_4, x_5)$ представлены в столбце 1 табл. 1. Они сгруппированы в группы и называются цепями Hansel [Hansel G., 1966]. Последовательность цепей начинается с самой короткой цепи «цепь 1» (01100) и (11100). Эта цепь состоит из двух случаев $(01100) < (11100)$. Наибольшая «цепь 10» состоит из 6 случаев: $(00000) < (00001) < (00011) < (00111) < (01111) < (11111)$. Случаи упорядочены как векторы в каждой цепи.

Чтобы строить цепи, представленные в табл. 1 (с пятью измерениями, например x_1, x_2, x_3, x_4, x_5 или u_1, u_2, u_3, u_4, u_5), используется следующий процесс.

Каждый шаг порождения цепи состоит в использовании текущей i -размерной цепи и построения $(i + 1)$ -размерной цепи. Поколение цепей для следующего измерения $(i + 1)$ появляется в результате следующего процесса.

- Мы клонируем 1-мерную цепь $(0) < (1)$ и производим ее копию $(0) < (1)$.
- После этого мы наращиваем цепь, добавляя второе измерение:
цепь 1 : $(00) < (01)$;
цепь 2 : $(10) < (11)$.
- Затем мы отделяем главный случай (11) от цепи 2 и добавляем его в качестве головы к цепи 1, создавая две 2-мерные цепи:
новая цепь 1 – $(00) < (01) < (11)$;
новая цепь 2 – (10) .
- Затем снова клонируем цепи и добавляем 0 и 1:
 $(000) < (001) < (011)$;
 $(100) < (101) < (111)$;
 $(010) < (110)$.
- Отделяем главный случай:
 $(000) < (001) < (011) < (111)$;
 $(100) < (101)$;
 $(010) < (110)$.
- Затем снова клонируем цепи и т.д.

В результате получаем цепи из таблицы 1.

Таблица 1. Динамическая последовательность интервью с экспертом

Случай	f – рак	ψ – Форма и плотность кальцинозов	Монотонное продолжение		Цепь	Случай
			$1 \rightarrow 1$	$0 \rightarrow 0$		
1	3	4	5	6	7	8
(01100)	1*	1*	1.2;6.3;7.3	7.1;8.1	Цепь 1	1.1
(11100)	1	1	6.4;7.4	5.1;3.1		1.2
(01010)	0*	1*	2.2;6.3;8.3	6.1;8.1	Цепь 2	2.1
(11010)	1*	1	6.4;8.4	3.1;6.1		2.2
(11000)	1*	1*	3.2	8.1;9.1	Цепь 3	3.1
(11001)	1	1	7.4;8.4	8.2;9.2		3.2
(10010)	0*	1*	4.2;9.3	6.1;9.1	Цепь 4	4.1
(10110)	1*	1	6.4;9.4	6.2;5.1		4.2
(10100)	1*	1*	5.2	7.1;9.1	Цепь 5	5.1
(10101)	1	1	7.4;9.4	7.2;9.2		5.2
(00010)	0	0*	6.2;10.3	10.1	Цепь 6	6.1
(00110)	1*	0*	6.3;10.4	7.1		6.2
(01110)	1	1	6.4;10.5			6.3
(11110)	1	1	10.6			6.4
(00100)	1*	0*	7.2;10.4	10.1	Цепь 7	7.1
(00101)	1	0*	7.3;10.4	10.2		7.2
(01101)	1	1*	7.4;10.5	8.2;10.2		7.3
(11101)	1	1	5.6			7.4
(01000)	0	1*	8.2	10.1	Цепь 8	8.1
(01001)	1*	1	8.3	10.2		8.2
(01011)	1	1	8.4	10.3		8.3
(11011)	1	1	10.6	9.3		8.4
(10000)	0	1*	9.2	10.1	Цепь 9	9.1
(10001)	1*	1	9.3	10.2		9.2
(10011)	1	1	9.4	10.3		9.3
(10111)	1	1	10.6	10.4		9.4
(00000)	0	0	10.2		Цепь 10	10.1
(00001)	0*	0	10.3			10.2
(00011)	1*	0	10.4			10.3
(00111)	1	1	10.5			10.4

(01111)	1	1	10.6			10.5
(11111)	1	1				10.6
Вопросов	13	12				

Цепи пронумерованы от 1 до 10, каждый случай имеет свой номер в цепи. Например, 1.2 означает второй случай в первой цепи. Знак «*» в столбцах 2 и 3 маркируют ответы, полученные от эксперта. Например, 1* для случая (01100) в столбце 2 означает, что эксперт ответил «да». Остальные ответы для той же самой цепи в столбце 2 автоматически получены, используя монотонность. Признак $f(01100) = 1$ для случая 1.1 распространяется на случаи 1.2, 6.3 и 7.3, представленные в колонке $1 \rightarrow 1$ монотонного продолжения. Аналогично, используя таблицу 1, вычисляются значения второй монотонной Булевой функции ψ . Признаки в последовательности (10010) интерпретируются как y_1, y_2, y_3, y_4, y_5 вместо x_1, x_2, x_3, x_4, x_5 , которые использовались для f . Цепи Hansel в этом случае такие же, т.к. количество признаков тоже.

В столбцах 4 и 5 выписаны случаи, распространяющие значения функций по свойству монотонности без опроса эксперта. Столбец 4 предназначен для расширения значений функции со значением 1, столбец 5 для распространения значений функции со значением 0. Если эксперт дал другой ответ $f(01100) = 0$ по сравнению с табл. 1, то значение 0 может быть распространено (в столбце 2) на случаи 7.1 (00100) и 8.1 (01000). Эти случаи перечислены в столбце 5 для случая (01100). Тогда нет необходимости спрашивать у эксперта случаи 7.1 (00100) и 8.1 (01000), т.к. они следуют из монотонности. Отрицательный ответ $f(01100) = 0$ не может быть распространен на случай $f(11100)$, поэтому у эксперта надо спросить относительно значения функции $f(11100)$. Если его/её ответ отрицательный $f(11100) = 0$, то эти значения могут быть распространены на случаи 5.1 и 3.1, перечисленные в столбце 5 для случая 1.2.

Общее количество случаев со знаком «*» в столбце 1 равно 13, для столбцов 2 и 3 они равны соответственно 13 и 12. Эти количества показывают, что 13 вопросов необходимы для восстановления функции f как функций от x_1, x_2, x_3, x_4, x_5 и 12 вопросов необходимы для восстановления функции ψ как функций от y_1, y_2, y_3, y_4, y_5 .

Полное восстановление любой функций f с 11 аргументами без оптимизации процесса интервью потребовало бы до $2^{11} = 2048$ вопросов к медику-эксперту. Согласно лемме Hansel оптимальный (т. е. минимальный) диалог для восстановления монотонной Булевой функции с 11 аргументами потребовал бы не более следующего числа вопросов:

Это число в 2.36 раза меньше, чем 2048 вопросов. Однако даже этот верхний предел 924

$$\binom{11}{5} + \binom{11}{6} = 2 \times 462 = 924$$

можно уменьшить. Введение иерархии уменьшает максимальное количество вопросов для восстановления монотонной Булевой функции с 11 переменными до $12+13+8 = 33$ вопросов.

4. Извлечение правил, используя монотонные Булевы функции

Полученная таблица позволяет явно выписать Булеву функцию экспертного правила принятия решений. Выпишем сначала Булеву функцию для признака $x_2 = \psi(y_1, y_2, y_3, y_4, y_5)$ на основании информации из столбца 3, следуя следующим шагам:

- найти все максимальные нижние единицы для всех цепей в виде элементарных конъюнкций;
- исключить избыточные термины (конъюнкции) из окончательной формулы.

Таким образом, из столбца 3 мы получим

$$x_2 = y_2 y_3 \vee y_2 y_4 \vee y_1 y_2 \vee y_1 y_4 \vee y_1 y_3 \vee y_2 y_3 y_4 \vee y_2 y_3 y_5 \vee y_2 \vee y_1 \vee y_3 y_4 y_5.$$

В цепях 1-5, 8-9 нижними единицами являются минимальные элементы цепей, поэтому соответствующие им конъюнкции состоят из двух элементов для цепей 1-5 и одному элементу для цепей 8-9, а в цепях 6-7, 10 минимальными единицами являются соответственно элементы (01110), (01101), (00111), которым соответствуют конъюнкции $y_2 y_3 y_4$, $y_2 y_3 y_5$, $y_3 y_4 y_5$.

Полученную дизъюнкцию конъюнкций можно упростим до выражения

$$x_2 = \psi(y_1, y_2, y_3, y_4, y_5) = y_2 \vee y_1 \vee y_3 y_4 y_5.$$

Из столбца 2 мы получим Булеву функцию от переменных x_1, x_2, x_3, x_4, x_5 для диагностики рака

$$f(x) = x_2 x_3 \vee x_1 x_2 x_4 \vee x_1 x_2 \vee x_1 x_3 x_4 \vee x_1 x_3 \vee x_3 x_4 \vee x_3 \vee x_2 x_5 \vee x_1 x_5 \vee x_4 x_5.$$

Эту дизъюнкцию можно упростим до выражения

$$f(x) = x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5$$

Для функции $v(w_1, w_2, w_3)$ второго уровня иерархии мы в интерактивном режиме получили следующую функцию

$$x_1 = v(w_1, w_2, w_3) = w_2 \vee w_1w_3.$$

Объединяя все функции, получим Булеву функцию для всех 11 признаков

$$f(x) = x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5 = (w_2 \vee w_1w_3)(y_1 \vee y_2 \vee y_3y_4y_5) \vee x_3 \vee (y_1 \vee y_2 \vee y_3y_4y_5 \vee w_2 \vee w_1w_3 \vee x_4)x_5.$$

5. Правила, извлеченные из эксперта

Из Булевой функции $f(x)$ можно извлечь диагностические правила.

Приведем примеры диагностических правил извлеченных из эксперта.

ЕСЛИ КОЛИЧЕСТВО кальцинозов в см² (w_1) большое

И ОБЩЕЕ КОЛИЧЕСТВО кальцинозов (w_3) большое

И нерегулярность в ФОРМЕ индивидуальных кальцинозов заметная (y_1),

ТО подозрение на рак.

ЕСЛИ КОЛИЧЕСТВО кальцинозов в см² (w_1) большое

И ОБЩЕЕ КОЛИЧЕСТВО кальцинозов большое (w_3)

И изменение в РАЗМЕРЕ кальцинозов (y_3) заметное

И ИЗМЕНЕНИЕ в плотности кальцинозов (y_4) заметное

И ПЛОТНОСТЬ кальциноза (y_5) заметная,

ТО подозрение на рак.

Далее мы представляем некоторые другие извлеченные правила кратко и формально.

ЕСЛИ $w_2 \& y_1$ **ТО** подозрение на рак.

ЕСЛИ $w_2 \& y_2$ **ТО** подозрение на рак.

ЕСЛИ $w_2 \& y_3 \& y_4 \& y_5$ **ТО** подозрение на рак.

ЕСЛИ $w_1 \& w_3 \& y_2$ **ТО** подозрение на рак.

ЕСЛИ $w_1 \& w_3 \& x_5$ **ТО** подозрение на рак.

Данные правила полностью представляют экспертные знания, извлеченные из эксперта в виде Булевой функции $f(x)$, и могут составить базу знаний, экспертную систему или диагностическую систему.

Литература

1. Hansel G. Sur le nombre des fonctions Booleenes monotones den variables, C. R. Acad. Sci. Paris (in French). 1966. Vol. 262, № 20. P. 1088–1090.
2. Kovalerchuk B., Vityaev E., Ruiz J.F. Design of consistent system for radiologists to support breast cancer diagnosis // Joint Conf. of Information Sciences, Duke University, NC, 1997. Vol. 2. P. 118–121.
3. Kovalerchuk, B., Vityaev, E., Ruiz, J. Consistent Knowledge Discovery in Medical Diagnosis. IEEE Engineering in Medicine and Biology Magazine. Special issue: «Medical Data Mining», July / August, 2000. P. 26–37.
4. Kovalerchuk, B., Vityaev, E., Ruiz, J.F. Consistent and Complete Data and «Expert» Mining in Medicine // Medical Data Mining and Knowledge Discovery, Springer. 2001. P. 238–280.
5. Kovalerchuk B., Taliani V. Comparison of empirical and computed fuzzy values of conjunction // Fuzzy Sets and Systems. 1996. Vol. 46. P. 49–53.
6. Kovalerchuk B., Triantaphyllou E., Ruiz J. Monotonicity and logical analysis of data: a mechanism for evaluation of mammographic and clinical data, in Kilcoyne RF, Lear JL, Rowberg AH (eds): Computer applications to assist radiology, Carlsbad, CA, Symposia Foundation. 1996. P. 191–196.
7. Wingo P.A., Tong T., Bolden S. Cancer Statistics, Ca-A Cancer Journal for Clinicians. Vol. 45, № 1. P. 8–30.