ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1978 год Выпуск 76

УДК 519.95:681.3.06

ОБ ОДНОМ МЕТОДЕ РАСПОЗНАВАНИЯ ОБРАЗОВ "СОВОКУПНЫЙ АНТИСИНДРОМ"

С.И.Гольдберг

Синдром в медицине обозначает сочетание признаков, характерных для данной болезни. Множество синдромов является главным инструментом для постановки диагноза. Особенно важны в этом случае патогномонические синдромы, а именно синдромы, встречаемые только при определенной болезни. Поиск этих синдромов использовался в получении распознающих, решающих правил на ЭВМ [1-3]. Однако с точки зрения распознавания и в плане изучения класса объектов представляют интерес и "антисиндромы" - наборы признаков, указывающих на то, что объект не принадлежит данному классу. Формализации этого понятия, изучению множества минимальных в некотором смысле антисиндромов и алгоритму построения всех минимальных антисиндромов посвящена данная статья.

Предполагается, что информация об объекте задается в виде строки из символов 0 и 1, где наличие на i-м месте 1 означает присутствие у объекта i-го признака, а 0 - отсутствие у объекта i-го признака.

Пусть L - множество всех n-мерных строк, заполненных символами 0 и 1, $A_1 \subset L$ - обучающая выборка класса B_1, $A_2 \subset L$ - обучающая выборка класса B_2. Требуется найти правило, классифицирующее произвольную строку из L относительно классов B_1 и B_2.

Введем ряд определений для L и $A \subset L$ - таблицы из m строк и n колонок, заполненных символами 0 и 1.

Строка $\gamma = (\gamma_1, \ldots, \gamma_n)$, $\gamma \in L$, $\gamma = \alpha \wedge \beta$, называется пересечением строк $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha \in L$ и $\beta = (\beta_1, \ldots, \beta_n)$, $\beta \in L$, если $\gamma_i = \alpha_i \wedge \beta_i (\forall i)$.

Строка $\gamma = (\gamma_1,\dots,\gamma_n)$, $\gamma \in L$, $\gamma = \alpha \vee \beta$, называется объединением строк $\alpha = (\alpha_1,\dots,\alpha_n)$, $\alpha \in L$ и $\beta = (\beta_1,\dots,\beta_n)$, $\beta \in L$, если $\gamma_i = \alpha_i \vee \beta_i (\forall_i)$.

Строка $\bar{\alpha} = (\bar{\alpha}_1,\dots,\bar{\alpha}_n)$, $\bar{\alpha} \in L$, называется дополнением строки $\alpha = (\alpha_1,\dots,\alpha_n)$, $\alpha \in L$, если $\bar{\alpha}_i = 1 - \alpha_i$.

Строка $\alpha \in L$ содержится в строке $\beta \in L$ $(\alpha \le \beta)$, если $\alpha \wedge \beta = \alpha$, и строго содержится $(\alpha < \beta)$, если $\alpha \wedge \beta = \alpha$ и $\alpha \neq \beta$.

Очевидно, L – булева алгебра.

Антисиндромом таблицы A называется строка $\alpha \in L$, которая не поглощается ни в одной строке таблицы A. Очевидно, что строка $\alpha \in L$ тогда и только тогда является антисиндромом таблицы A, когда $\alpha \wedge \bar{\beta} \neq 0$, где β – любая строка из таблицы A. Минимальный элемент множества антисиндромов таблицы A будем называть минимальным антисиндромом таблицы A. Множество всех минимальных антисиндромов A' назовем совокупным антисиндромом.

Пусть $\bar{A}$ – таблица, получившаяся из A заменой каждой строки на ее дополнение $\bar{\alpha}$.

Рассмотрим выборки A_1 и A_2 как таблицы A_1 и A_2. Каждому антисиндрому таблицы A_1 приписываем вес $\frac{e}{m_2}$, где e – число строк таблицы A_2, содержащих данный антисиндром, а m_2 – число строк в таблице A_2. Аналогично определим вес каждого антисиндрома таблицы A_2.

Возьмем произвольную строку γ из L. Отнесем ее к классу B_1, если

$$\sum_{i \in Q_1} p_i - \sum_{i \in Q_2} p_i < 0,$$

где Q_1 – множество минимальных антисиндромов таблицы A_1, которые содержатся в строке γ; Q_2 – множество минимальных антисиндромов таблицы A_2, которые содержатся в строке γ, p_i – веса соответствующих минимальных антисиндромов.

Строка γ относится к классу B_2, если

$$\sum_{i \in Q_1} p_i - \sum_{i \in Q_2} p_i > 0.$$

Так как строка α из таблицы A_1 не может быть антисиндромом таблицы A_1, то для нее множество Q_1 пусто и предложенное правило не может отнести строку из A_1 к классу B_2.

В случае же выполнения естественного условия: "для любой строки $\alpha \in A_1$ не существует строки $\beta \in A_2$ такой, что $\alpha \le \beta$", – строка $\alpha \in A_1$ является антисиндромом для таблицы A_2, т.е. содержит некото-

рые минимальные антисиндромы, причем их вес не равен нулю. Значит, строка $\alpha \in A_1$ будет отнесена к классу B_1.

Аналогично классифицируются строки из A_2.

Показателем информативности e-го признака может служить число $\sum_{i \in I_1} p_i - \sum_{i \in I_2} p_i$, где I_1 – множество минимальных антисиндромов таблицы A_1, у которых на e-м месте находится символ I; I_2 – множество минимальных антисиндромов таблицы A_2, у которых на e-м месте находится символ I; p_i – веса соответствующих минимальных антисиндромов.

Модуль $\left| \sum_{i \in I_1} p_i - \sum_{i \in I_2} p_i \right|$ показывает абсолютную значимость e-го признака, а $\operatorname{sign}\left(\sum_{i \in I_1} p_i - \sum_{i \in I_2} p_i \right)$ указывает номер класса.

ТЕОРЕМА I. Если в таблице A не существует строки, поглощающейся другой строкой таблицы A, то справедливо утверждение $(\bar{A}') = \bar{A}$.

ДОКАЗАТЕЛЬСТВО. По определению антисиндрома таблицы A, $\alpha \wedge \beta \neq 0$ для любой строки β из $\bar{A}$. Откуда, также по определению антисиндрома, следует, что любая строка β из $\bar{A}$ является антисиндромом таблицы $\bar{A}'$. Пусть $\gamma \notin \bar{A}$ и $\gamma \in (\bar{A}')$. Из определения минимального антисиндрома следует, что $\bar{\gamma} \wedge \beta \neq 0$, где β – любой антисиндром таблицы $\bar{A}'$, не равный γ, в частности, это верно для любого β из $\bar{A}$. Тогда γ – антисиндром таблицы A, т.е. существует $\gamma_0 \leq \bar{\gamma}$ и $\gamma_0 \in A' = \bar{A}'$, но $\gamma \wedge \gamma_0 = 0$, значит, $\gamma \notin (\bar{A}')'$. Противоречие показывает, что $(\bar{A}') \subset \bar{A}$. Отсюда и из того, что в таблице $\bar{A}$, в силу условия теоремы, не существует строки, содержащейся в другой строке таблицы $\bar{A}$, следует, что строка β из таблицы $\bar{A}$ может быть только минимальным антисиндромом таблицы $\bar{A}'$, т.е. $\bar{A} \subset (\bar{A}')'$. Из того, что $(\bar{A}') \subset \bar{A}$ и $\bar{A} \subset (\bar{A}')'$, следует $(\bar{A}')' = \bar{A}$. Теорема доказана.

ТЕОРЕМА 2. Пусть A и B – таблицы по n столбцов каждая. $A \cup B$ – таблица, составленная из всех таблиц A и B. Положим $A' + B' = \{\alpha = \beta \vee \gamma \mid \forall \beta \in A', \forall \gamma \in B'\}$. Тогда $(A \cup B)'$ – набор всех таких строк из $A' + B'$, что они не содержат никаких строк из $A' + B'$.

ДОКАЗАТЕЛЬСТВО. Пусть $\alpha \in (A \cup B)'$. Тогда α – антисиндром таблиц A и B, и, значит, существует $\beta \in A'$, такое, что $\beta \le \alpha$, и существует $\gamma \in B'$, такое, что $\gamma \le \alpha$. Отсюда $\beta \vee \gamma \le \alpha$, однако, очевидно, $\beta \vee \gamma$ – антисиндром таблицы $A \cup B$, следовательно, $\alpha = \beta \vee \gamma$. Понятно, что α не содержит никакой не равной α строки из $A' + B'$.

Теперь пусть $\beta \in A'$, $\gamma \in B'$ и $\beta \vee \gamma$ не содержит никакой другой строки из таблицы $A' + B'$. Если $\beta \vee \gamma \notin (A \cup B)'$, то, так как $\beta \vee \gamma$ – антисиндром таблицы $A \cup B$, существует $\alpha \in (A \cup B)'$, такое $\alpha < \beta \vee \gamma$. Но, как доказано, $\alpha \in A' + B'$. Противоречие с тем, что строка $\beta \vee \gamma$ не содержит никакой другой строки из таблицы $A' + B'$. Значит, $\beta \vee \gamma \in (A \cup B)'$.

Рассмотрим процедуру поиска всех минимальных антисиндромов. Прежде всего отметим, что если исключить из таблицы A строки, которые содержатся в других строках таблицы A, то совокупный антисиндром получившейся таблицы, очевидно, равен A'. Поэтому исключим такие строки заранее.

Для построения антисиндрома $\alpha = (\alpha_1, \ldots, \alpha_n)$ таблицы A достаточно знать номера столбцов, на которых в антисиндроме стоят символы I. Для каждого такого номера введем обозначение $k(\beta, i)$. Здесь k – порядковый номер антисиндрома, i – номер столбца в антисиндроме с символом I, β – строка таблицы $\bar{A}$, которая использовалась для получения этого номера i.

Сформируем первый антисиндром: $1(\beta_1, i_1)$ – минимальный номер столбца, в котором расположен символ I из первой строки таблицы A. Из всех строк таблицы $\bar{A}$, у которых в столбце i_1 стоит символ 0, выберем строку β_2 с наименьшим номером. Пусть i_2 – минимальный номер столбца с символом I в строке β_2, тогда строим $1(\beta_2, i_2)$. Далее ищем строку из A с нулями в столбцах i_1 и i_2, имеющую минимальный номер. Получим $1(\beta_3, i_3)$, где i_3 – наименьший номер столбца с символом I в строке β_3, и так далее, пока будут находиться нужные строки. Первый антисиндром образуется из символов I, стоящих в столбцах с номерами $i_1, i_2, \ldots, i_n$, и символов 0 – во всех других столбцах.

Пусть построено k антисиндромов таблицы A. Построим (k+1)-й антисиндром. Рассмотрим множество строк $\{\beta_j\}$, участвующих в построении k-го антисиндрома и содержащих столбец с символом I такой, что номер этого столбца больше $k(\beta_j, i_j)$. Выберем из множества $\{\beta_j\}$ строку β_1 с наибольшим номером. Составим набор для (k+1)-го антисиндрома: $[k+1](\beta_1, i_1) = k(\beta_1, I_1)$; $[k+1](\beta_2, i_2) =$

$=k(\beta_2, i_2)$; ... ; $[k+1](\beta_{l-1}, i_{l-1}) = k(\beta_{l-1}, i_{l-1})$.

Член $[k+1](\beta_1, i_1)$ будет равен минимальному номеру $N>k(\beta_1,i_1)$ столбца с символом I в строке β_1, для которого выполняются ограничения:

I) либо $N> k(\beta_1, i_1)$, либо в первой строке в столбце с номером N стоит 0;

2) либо $N > k(\beta_2, i_2)$, либо в строке β_2 в столбце с номером N стоит 0 и т.д. до

l-1) либо $N > k(\beta_{l-1},i_{l-1})$, либо в строке β_{l-1} в столбце с номером N стоит 0.

Дальнейшее построение производится, как при составлении первого антисиндрома, однако из поступающих на рассмотрение строк берутся номера, удовлетворяющие всем перечисленным выше ограничениям, а также условию:

1) либо $N> k(\beta_1, i_1)$, либо в строке β_1 в столбце с номером N стоит 0.

Если же таких номеров нет, то набор дальше не продолжаем, и (k+1) -й антисиндром строим вновь, используя вместо строки β_1 строку β_{l-1}.

Рассмотрим предложенный алгоритм на примере. Пусть

A:		$\overline{A}$:
	OIIOIOI	IOOIOIO
	IOOIIOO	OIIOOII
	OOIIOOI	IIOOIIO
	OOIOOII	IIOIIOO

Получаем множество строк:

IIOOOOO
IOIOOOO
IOOOOIO
IOOOOOI
OIOIOOO
OOIIIOO
OOIIOIO
OOOIOIO
OOOIIOI
OIOOOIO
OOOOIIO

Любая получившаяся в результате процедуры строка α является антисиндромом таблицы A, так как, по построению, $\alpha \wedge \beta \neq 0$, где β -

любая строка из таблицы $\bar{A}$. Покажем, что все минимальные антисиндромы таблицы A попали в построенное множество антисиндромов.

Пусть γ – произвольный антисиндром таблицы A, β_1 – первая строка таблицы $\bar{A}$. Возьмем минимальный по номеру столбец с символом I в строке β_1 $\wedge\gamma \neq 0$. Затем возьмем минимальный по номеру столбец с символом I в строке $\beta_2 \wedge \gamma \neq 0$, где β_2 – минимальная по номеру строка таблицы $\bar{A}$ с символом 0 в уже отобранном столбце, и т.д.

Таким образом, получили набор $i = (i_1, \ldots, i_e)$. Каждый элемент этого набора i_k является номером столбца с символом I в строке k с символами 0 в столбцах с номерами $i_1, \ldots, i_{k-1}$.

Причем в соответствии с методом составления набора:

1) либо $i_k > i_1$, либо в первой строке в столбце с номером i_k стоит 0;

2) либо $i_k > i_2$, либо в строке β_2 в столбце с номером i_k стоит 0 и т.д. до

k-1) либо $i_k > i_{k-1}$, либо в строке β_{k-1} в столбце с номером i_k стоит 0.

Такой набор i построится и в результате работы процедуры, что следует прямо из ее описания.

Соответствующая набору i строка α является антисиндромом таблицы A и в то же время содержится в антисиндроме γ. Так как вместо γ можно взять минимальные антисиндромы, то ясно, что в результате процедуры получаются все минимальные антисиндромы таблицы A.

Однако на примере видно, что в результате работы процедуры могут получаться и не минимальные антисиндромы. Поэтому для получения только минимальных антисиндромов нужно вставить в алгоритм блок, проверяющий, содержится ли вновь полученный антисиндром в уже имеющихся антисиндромах.

Задача поиска всех минимальных антисиндромов аналогична задаче нахождения всех простых покрытий множества строк, состоящих из символов 0 и I [4]. Однако в рассматриваемых примерах предложенный алгоритм оказался более эффективным, чем рассматриваемый в [4].

ПРИМЕР. Даны таблицы:

A_1:		A_2:	
	0I0I00		00I0I0
	00II0I		I00II0
	I000II		0II00I
	0I00II		I0I00I

Стоит вопрос: куда отнести строчку III000? Строим обратные таблицы:

$\bar{A}_1$:

IOIOIO
IIOOIO
OIIIOO
IOIIOO

$\bar{A}_2$:

IIOIOI
OIIOOI
IOOIIO
OIOIIO

В результате применения процедуры получим

для A_1:

IIOOOO
IOIOOO
IOOIOO
OIIOOO
OOIOIO
OIOIIO
OOOIIO
OIOIOI

для A_2:

IIOOOO
IOIIOO
IOIOIO
IOOIII
IOOOII
OIOIOO
OIOOIO
OOIIOO
OOOIII
OOOOII

A_1':

IIOOOO	$p_1 = 0$
IOIOOO	$p_2 = I/5$
IOOIOO	$p_3 = 2/5$
OIIOOO	$p_4 = I/5$
OOIOIO	$p_5 = I/5$
OOOIIO	$p_6 = I/5$
OIOIOI	$p_7 = 0$

A_2':

IIOOOO	$p_1 = 0$
IOIOIO	$p_2 = 0$
OIOIOO	$p_3 = I/4$
OIOOIO	$p_4 = I/4$
OOIIOO	$p_5 = I/4$
OOOOII	$p_6 = 2/4$

Информативность признаков следующая:

1 признака I2/20
2 признака I/20
3 признака 7/20
4 признака 2/20
5 признака 7/20
6 признака I0/20

Предложенное в статье решающее правило придает строке III000 вес 2/5 и относит ее к классу B_2 с обучающей выборкой A_2.

Литература

1. БОНГАРД М.М. Проблемы распознавания.М.,"Наука",1967.

2. ЛБОВ Г.С., КОТЮКОВ В.И., МАШАРОВ Ю.П. Метод обнаружения логических закономерностей на эмпирических таблицах. - В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы,вып.67.) Новосибирск,1967,с.26-42.

3. ЗИТЯЕВ Е.Е. Метод обнаружения закономерностей и метод предсказания. - В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы, вып.67.) Новосибирск,1976,с.54-69.

4. ЗАКРЕЦКИЙ А.Д. Логические уравнения. Киев,"Наука и техника", 1975.

Поступила в ред.-изд.отд.
25 августа 1978 года