МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ПЕДАГОГИЧЕСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Е.В. Дюкова

ДИСКРЕТНЫЕ (ЛОГИЧЕСКИЕ) ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ: ПРИНЦИПЫ КОНСТРУИРОВАНИЯ, СЛОЖНОСТЬ РЕАЛИЗАЦИИ И ОСНОВНЫЕ МОДЕЛИ

УЧЕБНОЕ ПОСОБИЕ ДЛЯ СТУДЕНТОВ
МАТЕМАТИЧЕСКИХ ФАКУЛЬТЕТОВ ПЕДВУЗОВ

Допущено
Учебно-методическим объединением
по специальностям педагогического образования
в качестве учебного пособия
для студентов высших учебных заведений,
обучающихся по специальности 030100 - информатика

Москва
2003

ББК 22.12я73
Д 95
УДК 519.710

Печатается по решению редакционно-издательского совета
Московского педагогического государственного университета

Научный редактор академик РАО, член-корр. РАН,
Доктор физико-математических наук, профессор В.Л. Матросов

Рецензенты:

Д95 **Дюкова Е.В.**
Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели. Учебное пособие для студентов математических факультетов педвузов. - М.: Прометей, 2003. – 29 с.

ISBN 5-70420-1092-9

В пособии изложены общие принципы, лежащие в основе дискретного подхода к задачам распознавания, центральной проблемой которого является поиск информативных фрагментов признаковых описаний объектов. При поиске информативных фрагментов используется аппарат логических функций, в частности методы преобразования нормальных форм булевых функций, а также теория покрытий булевых и целочисленных матриц. Рассматриваются основные модели дискретных (логических) процедур распознавания и изучаются вопросы, связанные со сложностью их реализации.

Пособие предназначено для студентов педагогических вузов по дисциплине "Теоретические основы информатики", обучающихся по специальности 030100 - "Информатика".

На сегодняшний день другая литература по рассматриваемым в пособии вопросам, которая могла бы быть рекомендована в качестве учебного пособия для студентов, отсутствует.

ISBN 5-70420-1092-9

Введение

В самых общих чертах задача распознавания состоит в следующем. Исследуется некоторое множество объектов M. Известно, что M представимо в виде объединения l подмножеств $K_1, \ldots, K_l$, называемых классами. Объекты из M описываются некоторой системой признаков $\{x_1, \ldots, x_n\}$. Имеется конечный набор $S_1, \ldots, S_m$ объектов из M, о которых известно, каким классам они принадлежат. Это прецеденты или обучающие объекты. Пусть их описания имеют вид $S_1 = (a_{11}, \ldots, a_{1n})$, $S_2 = (a_{21}, \ldots, a_{2n})$, …, $S_m = (a_{m1}, \ldots, a_{mn})$, здесь a_{ij} - значение признака x_j для объекта S_i. Требуется по предъявленному набору значений признаков $(b_1, \ldots, b_n)$, описывающему некоторый объект из M, о котором, вообще говоря, не известно какому классу он принадлежит, определить этот класс.

Обычно при постановке практических задач распознавания первоначальные описания объектов содержат все доступные наблюдению или измерению характеристики или параметры. В результате объекты оказываются описанными несколькими десятками, а иногда и сотнями величин. Подобная ситуация характерна, в частности, для задач медицинской диагностики, геологического, технического и социологического прогнозирования и. т.д. Первоначально для решения таких задач применялись в основном статистические методы. Считалось, что данное направление является частью математической статистики. Для анализа сложных описаний с помощью статистических методов необходимо принимать на веру дополнительные гипотезы вероятностного характера, т.е. предъявлять достаточно сильные требования к пространствам исследуемых объектов. Кроме того, для получения надежных результатов на основе статистического подхода требуются чрезвычайно большие массивы прецедентов, т.е. обучающая выборка должна быть достаточно представительной. Оказалось, что набор большого числа прецедентов требует, как правило, дорогостоящих и трудоемких работ, а в некоторых случаях вообще невозможен. Например, такая ситуация имеет место в задачах прогнозирования редких металлов. Для решения подобных задач не было адекватных математических методов и их пришлось создавать на основе совершенно новых идей. В настоящее время интенсивно развиваются методы построения оптимальных (корректных) алгоритмов распознавания (алгебраический подход [11, 14, 16]) и дискретный подход к проблеме синтеза эффективных алгоритмов распознавания, основанный на комбинаторном анализе исходной информации [1-10, 20].

Центральной проблемой дискретного подхода является поиск наиболее информативных подописаний (или фрагментов описаний)

обучающих объектов. Например, информативными считаются такие подописания, которые позволяют различать объекты из разных классов или отличать данный объект от всех объектов, не принадлежащих тому же классу, что и рассматриваемый.

Поиск информативных фрагментов основан на использовании аппарата дискретной математики, в частности булевой алгебры, теории д.н.ф., теории покрытий булевых и целочисленных матриц. Основополагающими работами являются работы академика РАН Ю.И. Журавлева. Одной из первых работ в этом направлении была статья [2], в которой рассматривалась задача прогнозирования золотоносных месторождений и для построения распознающего алгоритма было использовано хорошо известное в дискретной математике понятие теста. Это понятие было введено член-корреспондентом РАН С.В. Яблонским и первоначально применялось в задачах контроля управляющих систем [18]. В задачах контроля тест – это совокупность наборов значений переменных, которая позволяет различать исправное состояние системы от неисправного и находить возможную ошибку.

Упомянутая выше статья Дмитриева, Журавлева и Кренделева, а также статьи Вайнвайга и Бонгарда, в которых описывалась модель распознающего алгоритма под названием “Кора”, положили начало широкому применению методов дискретного анализа в задачах распознавания, классификации и прогнозирования.

В настоящее время дискретная техника анализа признаковых описаний используется и в методах оптимизации многопараметрических моделей распознавания и в методах коррекции (как при построении базовых алгоритмов, так и при построении корректирующих операций) и т.д. Выявление информативных фрагментов описаний обучающих объектов дает возможность проводить качественный анализ исходной информации, например с целью информационной классификации признаков и уменьшения их числа.

Дискретные методы привели также к появлению целого класса сложно устроенных эвристик, называемых дискретными или логическими процедурами распознавания (при их конструировании, как мы увидим дальше, может быть использован аппарат логических функций).

Таким образом, использование аппарата и методов дискретной математики для решения задач распознавания имеет ряд достоинств, к числу которых, прежде всего, следует отнести возможность получения результата при отсутствии сведений о функциях распределения и при наличии малых обучающих выборок. Однако применение дискретного подхода оказывается во многих случаях сложным в силу чисто вычислительных трудностей переборного характера, возникающих на этапе поиска информативных фрагментов описаний объектов. Следует отметить, что в силу экспоненциального роста числа фрагментов при

возрастании размерности описаний, решение проблемы только за счет повышения производительности вычислительной техники нереально.

Наиболее полное изложение методов дискретного анализа информации в задачах распознавания можно найти в [8].

1. Общие принципы построения дискретных процедур распознавания

Предполагается, что исходные описания объектов даны в виде наборов значений целочисленных признаков (при этом желательно не очень высокой значности). Кроме того, предполагается, что описания объектов из разных классов различаются.

Пусть H – набор из r различных признаков, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, и пусть S - некоторый объект из M, $S = (a_1, \ldots, a_n)$, здесь a_j – значение признака x_j, $j = 1, 2, \ldots, n$.

Набор признаков H выделяет в описании объекта S фрагмент $(a_{j_1}, \ldots, a_{j_r})$. В случае $S \in \{S_1, \ldots, S_m\}$ фрагмент $(a_{j_1}, \ldots, a_{j_r})$ будем называть *элементарным классификатором* и обозначать через (S, H).

Введем обозначения: N - множество всех элементарных классификаторов, т.е $N = \{(S, H) \mid S \in \{S_1, \ldots, S_m\}, H \subseteq \{x_1, \ldots, x_n\}\}$; $N(K)$, $K \in \{K_1, \ldots, K_l\}$, - множество всех элементарных классификаторов из N, порождаемых обучающими объектами из класса K, т.е. $N(K) = \{(S, H) \mid S \in \{S_1, \ldots, S_m\}, H \subseteq \{x_1, \ldots, x_n\}, S \in K\}$.

Пусть даны два объекта S' и S'' из M, $S' = (a'_1, \ldots, a'_n)$, $S'' = (a''_1, \ldots, a''_n)$. Близость объектов S' и S'' по набору признаков H, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, будем оценивать величиной

$$B(S', S'', H) = \begin{cases} 1, & \textit{если} \quad a'_{j_t} = a''_{j_t} \quad \textit{при} \quad t = 1,2,...,r, \\ 0, & \textit{в противном случае.} \end{cases}$$

Таким образом, объекты S' и S'' близки по набору признаков H ($B(S', S'', H) = 1$), в том и только в том случае, если совпадают фрагменты (S', H) и (S'', H).

Опишем общую схему работы распознающего алгоритма A.

На первом этапе (этапе обучения) для каждого класса K строится некоторое множество элементарных классификаторов с заданными свойствами, т.е. строится некоторое подмножество $N_A(K)$ из $N(K)$. Элементарные классификаторы из $N_A(K_1), \ldots, N_A(K_l)$ и только они считаются информативными.

Рассмотрим наиболее типичные примеры. Один из таких примеров - модели алгоритмов голосования по представительным наборам (алгоритмы типа "Кора").

Определение. Элементарный классификатор (S, H) из $N(K)$ назовем *представительным набором* для K, если для любого обучающего объекта $S' \notin K$ имеет место $B(S, S', H) = 0$ (т.е. элементарные

классификаторы (S, H) и (S', H) не совпадают).

Таким образом, представительный набор для K – фрагмент описания некоторого обучающего объекта из K, позволяющий отличать этот объект от любого другого обучающего объекта, не вошедшего в класс K.

В моделях с представительными наборами в качестве $N_A(K)$ рассматриваются представительные наборы для K. Более информативными считаются короткие представительные наборы. Поэтому в таких моделях, как правило, строят не все представительные наборы, а только те, длины которых не превосходят некоторой наперед заданной величины.

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$, $H^{(t)} = \{x_{j_1}, \ldots, x_{j_{t-1}}, x_{j_{t+1}}, \ldots, x_{j_r}\}$, $t \in \{1, 2, \ldots, r\}$.

Определение. Представительный набор (S, H) для K назовем *тупиковым*, если для любого t, $t \in \{1, 2, \ldots, r\}$, элементарный классификатор $(S, H^{(t)})$ не является представительным набором для K.

Содержательно тупиковость означает, что представительный набор является неизбыточным (при сжатии он теряет способность отличать порождающий его объект от некоторых объектов из других классов). В качестве $N_A(K)$ можно рассматривать тупиковые представительные наборы.

Впервые алгоритм голосования по представительным наборам описан в [1].

Определение. Набор признаков H назовем *тестом*, если для любого $K \in \{K_1, \ldots, K_l\}$ и любого обучающего объекта $S \in K$ элементарный классификатор (S, H) является представительным набором для K.

Очевидно, набор признаков $\{x_{j_1}, \ldots, x_{j_r}\}$ является тестом, если для любых двух обучающих объектов S_i и S_p, принадлежащих разным классам, можно указать t, $t \in \{1, 2, \ldots, r\}$, такое что $B(S_i, S_p, \{x_t\}) = 0$. Содержательно тест – набор признаков, позволяющих безошибочно разделять обучающий материал на классы.

Определение. Тест назовем *тупиковым*, если никакое его собственное подмножество тестом не является.

В качестве $N_A(K)$ можно рассматривать подмножество $N(K)$, порождаемое тестами. Так же, как и в случае с представительными наборами, ограничивают длину теста или берут тупиковые тесты.

Первая модель тестового алгоритма описана в [2].

Замечание. Материал обучения обычно представляется в виде

таблицы – таблицы обучения T, столбцы которой соответствуют отдельным признакам, а каждая строка есть набор значений признаков, описывающий один из обучающих объектов.

Дадим стандартное, точнее более принятое определение теста.

Определение. Набор столбцов H таблицы T назовем *тестом*, если в подтаблице, образованной этим набором, любые две строки, описывающие объекты из разных классов, различны.

Итак, на первом этапе (этапе обучения) строится множество информативных для данного алгоритма фрагментов описаний обучающих объектов, т.е. строится множество $N_A = N_A(K_1) \cup \ldots \cup N_A(K_l)$. После построения N_A осуществляется процедура голосования. В простейшей модификации считается, что элементарный классификатор (S', H) из $N_A(K)$ подает "голос" за принадлежность опознаваемого объекта S к классу K, если $B(S', S, H) = 1$, т.е. соответствующие фрагменты описаний объектов S и S' совпадают.

Число "голосов" $\Gamma(S, K_j)$, поданных парами из $N_A(K_j)$ за принадлежность объекта S к классу K_j, в простейшей модификации вычисляется по формуле

$$\Gamma(S, K_j) = \frac{1}{\left|N_A(K_j)\right|} \sum_{(S',H) \in N_A(K_j)} B(S, S', H),$$

где $\left|N_A(K_j)\right|$ - мощность множества $N_A(K_j)$, $\Gamma(S, K_j)$ – оценка за принадлежность S к классу K_j.

Вычисляются оценки $\Gamma(S, K_1), \ldots, \Gamma(S, K_l)$. Тогда, если

$$\Gamma(S, K_t) = \max_{1 \le j \le l} \Gamma(S, K_j)$$

и $\Gamma(S, K_t) \ne \Gamma(S, K_u)$ при $u \ne t$, то алгоритм A относит объект S к классу K_t. Если же

$$\Gamma(S, K_t) = \Gamma(S, K_u) = \max_{1 \le j \le l} \Gamma(S, K_j),$$

то алгоритм A отказывается от распознавания объекта S.

Для оценки качества работы распознающего алгоритма A часто используется процедура скользящего контроля, которая заключается в следующем. Для каждого i из $\{1, 2, \ldots, m\}$ по обучающей подвыборке $\{S_1, \ldots, S_m\} \setminus \{S_i\}$ вычисляются оценки $\Gamma(S_i, K_j)$, $j = 1, 2, \ldots, l$. Пусть q - число правильно распознанных объектов S_i, $i = 1, 2, \ldots, m$. Тогда качество работы алгоритма A оценивается функционалом

$$\varphi_{ск}(A) = q / m.$$

Для больших задач описанная процедура требует существенных

вычислительных затрат. В случае, когда A - алгоритм голосования по представительным наборам, для процедуры скользящего контроля существует "быстрый" метод вычисления оценок $\Gamma(S_i, K_j)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, l$ [20]. Данный метод позволяет сократить время счета примерно в m раз.

Для оценки качества работы алгоритма A может быть использован и независимый контроль – набор из t объектов, не вошедших в обучающую выборку, про которые также известно, каким классам они принадлежат. Такой набор объектов называется контрольной выборкой. Пусть q - число правильно распознанных объектов из контрольной выборки. В данном случае качество работы алгоритма A оценивается функционалом

$$\varphi_{контр}(A) = q / t .$$

В [10, 12, 20] приведены теоретические и экспериментальные результаты, касающиеся построения эффективных (в смысле качества работы) алгоритмов распознавания, базирующихся на понятии представительного набора.

На практике применяются более сложные модели с дополнительными параметрами, характеризующими представительность (типичность) обучающих объектов и их подописаний по отношению к своим классам, а также информативность признаков.

В моделях с представительными наборами отбираются наиболее “весомые” представительные наборы, например те, которые по данному набору признаков встречаются в классе у достаточно большого числа объектов [6, 10, 20]. Рассматриваются также модели с почти представительными наборами. В этих моделях информативными для класса K считаются такие классификаторы из $N(K)$, которые по данному набору признаков встречаются достаточно часто в описаниях объектов из класса K и достаточно редко в описаниях объектов из других классов.

В тестовых моделях для увеличения быстродействия применяют стохастические алгоритмы, в которых построение множества всех тупиковых тестов заменено построением достаточно представительной случайной выборки из него. В данном случае оценки $\Gamma(S, K_j)$, $j \in \{1, 2, \ldots, l\}$, вычисляются приближенно и оценивается возможная при этом ошибка [4, 13].

В [9, 10, 20] предлагается ряд новых эвристик дискретного характера, в которых информативными для класса K считаются наборы из допустимых значений признаков, отсутствующие в описаниях всех обучающих объектов класса K.

Рассматриваемые модели могут быть модифицированы на случай веществennозначной информации [17]. Для каждого признака x_j,

$j = 1, 2, \ldots, n$, задается некоторый вещественный параметр E_j – точность измерения признака x_j. Опять же в простейшей модификации величина $B(S', S'', H)$ определяется следующим образом. Пусть $S' = (a'_1, \ldots, a'_n)$, $S'' = (a''_1, \ldots, a''_n)$, $H = \{x_{j_1}, \ldots, x_{j_r}\}$. Тогда

$$B(S', S'', H) = \begin{cases} 1, \textit{ если } \left|a'_{j_t} - a''_{j_t}\right| \le E_{j_t},\ t = 1, 2, \ldots, r, \\ 0, \textit{ в противном случае.} \end{cases}$$

Другой подход к обработке вещественнозначной информации (предложен Ю.И. Журавлевым) заключается в том, что с самого начала исходная информация перекодируется в целочисленную (модели с перекодировками).

Замечание. Приведенная схема построения распознающего алгоритма дает представление о том, как устроены модели типа вычисления оценок (или алгоритмы голосования по опорным множествам) [11]. Среди этих моделей исторически первыми появились тестовые модели (точнее алгоритмы голосования по множеству тупиковых тестов). Тестовые алгоритмы оказались очень трудоемкими в смысле вычислительных затрат. Теоретические исследования статистических свойств тупиковых тестов показали, что почти всегда тупиковые тесты имеют примерно одну и ту же длину. Эта длина, вообще говоря, зависит от соотношения между параметрами m и n. Например, в случае, когда информация бинарная, $l = 2$ и $(m_1 m_2)^{\alpha} \le n, \alpha > 1$, для почти всех таблиц при $n \to \infty$ почти все тупиковые тесты имеют длину порядка $\frac{1}{2}\log(m_1 m_2 n)$. Указанное обстоятельство послужило одним из обоснований для построения алгоритмов вычисления оценок, в которых в качестве множества $N_A(K)$ берутся всевозможные элементарные классификаторы (S, H) из $N(K)$, где H имеет фиксированную мощность r. Значение r может задаваться путем некоторого предварительного анализа обучающей выборки. Модель предназначена в основном для обработки вещественнозначной информации. Поэтому задаются параметры E_j и P_j, $j = 1, 2, \ldots, n$, характеризующие соответственно точность измерения каждого признака x_j и его информативность, а также параметры $\gamma(S')$, $S' \in \{S_1, \ldots, S_m\}$, характеризующие представительность (типичность) обучающих объектов по отношению к своим классам. Для этой модели существует формула эффективного вычисления оценки $\Gamma(S, K_j)$, $S = (a_1, \ldots, a_n)$, $K_j \in \{K_1, \ldots, K_l\}$, которая имеет вид

$$\Gamma(S, K_j) = \frac{1}{m_j} \sum_{S' \in K_j} \left[\gamma(S') \left(\sum_{i \in I(S', S)} P_i \right) C^{r-1}_{d(S', S)} \right],$$

где $S' = (a'_1, \ldots, a'_n)$, $I(S', S) = \{i \mid |a'_i - a_i| \le E_i, i \in \{1, 2, \ldots, n\}\}$, $d(S', S) = |I(S', S)|$ - мощность множества $I(S', S)$.

Упражнения

Задача 1. Пусть $K \in \{K_1, \ldots, K_l\}$ и пусть множество $N_A(K)$ порождается

а) множеством тестов;

б) множеством представительных наборов.

Пусть S - обучающий объект из класса K. Можно ли утверждать, что в каждом из перечисленных случаев алгоритм A отнесет объект S к классу K.

Задача 2. Пусть $(S, H) \in N(K)$, $N_A(K) = N(K)$.

В каком случае при голосовании по фрагменту (S, H) на скользящем контроле объект S будет отнесен к классу K и в каком случае этот объект будет отнесен к другому классу?

Задача 3. Пусть $K \in \{K_1, \ldots, K_l\}$,

$$N(\overline{K}) = \bigcup_{j=1}^{l} N(K_j) \setminus N(K).$$

Элементарный классификатор (S, H) из $N(\overline{K})$ назовем антипредставительным набором для K, если для любого обучающего объекта $S' \in K$ имеет место $B(S, S', H) = 0$. Пусть $N_A(\overline{K})$ порождается множеством антипредставительных наборов для K. Принадлежность распознаваемого объекта S к классу K будем оценивать величиной

$$\Gamma(S, K) = \frac{1}{|N_A(\overline{K})|} \sum_{(S', H) \in N_A(\overline{K})} B(S, S', H),$$

где $|N_A(\overline{K})|$ - мощность множества $N_A(\overline{K})$.

Показать, что в случае $l = 2$ алгоритм голосования по представительным наборам и алгоритм голосования по антипредставительным наборам отнесут объект S к одному и тому же классу. Показать, что в общем случае (при $l > 2$) утверждение неверно.

2. Методы построения элементарных классификаторов в дискретных процедурах распознавания

2.1. Построение элементарных классификаторов на основе преобразования нормальных форм логических функций

При конструировании дискретных процедур распознавания часто используется аппарат логических функций.

Все неопределяемые ниже понятия можно найти в [15,19,21].

Пусть $f_K(x_1, \ldots, x_n)$ – частичная (не всюду определенная) двузначная функция, которая принимает значение 1 на наборах, являющихся описаниями объектов из класса K, и 0 – на наборах, описывающих остальные обучающие объекты, т.е. f_K – характеристическая функция класса K. Покажем, что элементарным классификаторам, порождаемым парами из $N(K)$, соответствуют специальные конъюнкции функции f_K и, следовательно, каждому распознающему алгоритму соответствует некоторое множество таких конъюнкций (каждый алгоритм основан на построении некоторого множества таких конъюнкций).

Для простоты будем рассматривать случай, когда объекты описаны бинарными признаками.

Пусть E^n – множество наборов вида $(\alpha_1, \ldots, \alpha_n)$, где $\alpha_i \in \{0, 1\}$ при $i = 1, 2, \ldots, n$; A_K и B_K обозначают, соответственно, множества наборов из E^n, на которых функция f_K равна 1 и 0; B - элементарная конъюнкция (э.к.) над переменными $x_1, \ldots, x_n$; N_B - интервал истинности э.к. B.

Введем ряд определений.

Определение. Э.к. B назовем *почти допустимой* для f_K, если $N_B \bigcap A_K \neq \emptyset$.

Очевидным является

Утверждение 1. *Э.к.* $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ *является почти допустимой для* f_K *тогда и только тогда, когда* $(\sigma_1, \ldots, \sigma_r)$ *является элементарным классификатором для* K*, порождаемым набором признаков* $\{x_{j_1}, \ldots, x_{j_r}\}$.

Таким образом, каждой почти допустимой конъюнкции B функции f_K соответствует некоторое множество элементарных классификаторов из $N(K)$. Их число равно $|N_B \bigcap A_K|$.

Определение. Э.к. B назовем *допустимой* для f_K, если $N_B \bigcap A_K \neq \emptyset$ и $N_B \bigcap B_K = \emptyset$.

Очевидным является

Утверждение 2. *Э.к.* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является допустимой для* f_K *тогда и только тогда, когда* $(\sigma_1, \dots, \sigma_r)$ *- представительный набор для* K*, порождаемый набором признаков* $\{x_{j_1}, \dots, x_{j_r}\}$.

Определение. Э.к. B назовем *неприводимой* для f_K, если не существует элементарной конъюнкции B' такой, что $N_{B'} \supset N_B$ и $N_{B'} \bigcap B_K = N_B \bigcap B_K$.

Э.к. B является неприводимой, если она в определенном смысле неизбыточна. При сжатии такой конъюнкции увеличивается пересечение ее интервала истинности N_B с множеством нулей функции f_K. В частности, если B - неприводимая и допустимая конъюнкция, то при сжатии она перестает быть допустимой.

Определение. Э.к. B назовем *максимальной* для f_K, если она является допустимой и не существует допустимой конъюнкции B' такой, что $N_{B'} \supset N_B$.

Из приведенных определений следует, что э.к является максимальной для f_K тогда и только тогда, когда она является допустимой и неприводимой.

Очевидным является

Утверждение 3. *Набор признаков* $\{x_{j_1}, \dots, x_{j_r}\}$ *порождает тупиковый представительный набор для* K *вида* $(\sigma_1, \dots, \sigma_r)$ *тогда и только тогда, когда конъюнкция* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является максимальной для* f_K.

Очевидным является

Утверждение 4. *Набор признаков* $\{x_{j_1}, \dots, x_{j_r}\}$ *является тестом тогда и только тогда, когда для каждой функции* f_{K_t}, $t \in \{1, 2, \dots, l\}$, *каждая почти допустимая конъюнкция вида* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является допустимой*.

Замечание. Приведенные выше определения почти допустимой, допустимой, неприводимой и максимальной конъюнкции частичной булевой функции полностью переносится на случай, когда f_K является всюду определенной булевой функцией, т.е. когда $A_K = E^n \setminus B_K$.

Таким образом, при построении распознающих алгоритмов с использованием аппарата логических функций возникают задачи построения почти допустимых, допустимых и максимальных конъюнкций частичной булевой функции. Наибольшую сложность представляет поиск максимальных конъюнкций. Одним из наиболее известных способов

построения максимальных конъюнкций частичной булевой функции является следующий.

Рассматривается всюду определенная булева функция $F_K(x_1, \ldots, x_n)$, которая совпадает с f_K на множестве нулей и единиц, на остальных наборах из E^n функция $F_K(x_1, \ldots, x_n)$ равна 1. Задание множества нулей всюду определенной булевой функции равносильно заданию совершенной КНФ этой функции. Пусть B_K состоит из наборов $(\beta_{11}, \ldots, \beta_{1n})$, $(\beta_{21}, \ldots, \beta_{2n})$, …, $(\beta_{u1}, \ldots, \beta_{un})$. Тогда, очевидно, функция F_K реализуется КНФ вида $D_1 \,\&\, \ldots \,\&\, D_u$, где

$D_i = x_1^{\overline{\beta}_{i1}} \vee \ldots \vee x_n^{\overline{\beta}_{in}}$, $i = 1, 2, \ldots, u$.

Положим $N_{F_K} = E^n \setminus B_K$, $N_{\overline{F}_K} = B_K$.

Утверждение 5. *Э.к. B является допустимой для F_K тогда и только тогда, когда каждая дизъюнкция D_i, $i \in \{1, 2, \ldots, u\}$, содержит хотя бы один множитель из B.*

Доказательство. 1. Пусть B - допустимая конъюнкция для F_K. Покажем, что каждая дизъюнкция D_i, $i \in \{1, 2, \ldots, r\}$, содержит хотя бы один множитель из B.

Предположим противное. Пусть некоторая дизъюнкция D_i не содержит ни одного множителя из B. Не ограничивая общности можно считать, что $B = x_1^{\sigma_1} \ldots x_r^{\sigma_r}$. Тогда D_i имеет вид

$$D_i = x_1^{\overline{\sigma}_1} \vee \ldots \vee x_r^{\overline{\sigma}_r} \vee x_{r+1}^{\sigma_{r+1}} \vee \ldots \vee x_n^{\sigma_n}.$$

Рассмотрим набор $\alpha = (\sigma_1, \ldots, \sigma_r, \overline{\sigma}_{r+1}, \ldots, \overline{\sigma}_n)$. По построению $\alpha \in N_B \subseteq N_{F_K}$. С другой стороны, набор α обращает D_i в 0 и, следовательно, $\alpha \in N_{\overline{F}_K}$. Противоречие.

2. Пусть каждая дизъюнкция D_i, $i \in \{1, 2, \ldots, u\}$, содержит некоторый множитель из B. Тогда, очевидно, любой набор из N_B обращает КНФ $D_1 \,\&\, \ldots \,\&\, D_u$ в 1, т.е. принадлежит N_{F_K}.

Утверждение доказано.

Утверждение 6. *Э.к. B ранга r является неприводимой для F_K тогда и только тогда, когда в КНФ $D_1 \,\&\, \ldots \,\&\, D_u$ можно указать r дизъюнкций $D_{i_1}, \ldots, D_{i_r}$ таких, что каждая дизъюнкция содержит в точности один множитель из B и, если $r > 1$, $p, q \in \{i_1, \ldots, i_r\}$, $p \neq q$, то дизъюнкции D_p и D_q содержат разные множители из B.*

Доказательство. 1. Пусть э.к. B является неприводимой для F_K и пусть условие утверждения не выполнено. Это означает, что в B можно

указать переменную $x_t^{\sigma_t}$ такую, что для каждой дизъюнкции D_i в исходной КНФ выполнено одно из двух следующих условий:

1) D_i не содержит $x_t^{\sigma_t}$;

2) D_i содержит $x_t^{\sigma_t}$ и D_i содержит некоторую другую переменную из B. Удалим $x_t^{\sigma_t}$ из B. Получим э.к. B'. Обозначим через $M(B)$ множество дизъюнкций в исходной КНФ, не содержащих переменные из B, через $M(B')$ аналогичное множество для B'. Очевидно, $N_{B'} \supset N_B$ и $M(B') = M(B)$. Отсюда следует, что $N_{B'} \bigcap N_{\overline{F}_K} = N_B \bigcap N_{\overline{F}_K}$. Противоречие.

2. Пусть для э.к. $B = x_1^{\sigma_1} \ldots x_r^{\sigma_r}$ выполнено условие утверждения. Покажем, что B - неприводимая конъюнкция.

Предположим противное. Тогда можно указать э.к. B' такую, что $N_{B'} \supset N_B$ и $N_{B'} \bigcap N_{\overline{F}_K} = N_B \bigcap N_{\overline{F}_K}$. Э.к. B' получается из B удалением хотя бы одного множителя $x_t^{\sigma_t}$. В исходной КНФ есть дизъюнкция, содержащая $x_t^{\sigma_t}$ и не содержащая ни одного другого множителя из B. Таким образом, $M(B') \supset M(B)$ и, значит, $N_{B'} \bigcap N_{\overline{F}_K} \neq N_B \bigcap N_{\overline{F}_K}$. Противоречие.

Утверждение доказано.

Замечание. Из доказательства утверждений 5 и 6 следует, что они справедливы и в случае, когда КНФ имеет произвольный вид, т.е. не обязательно является совершенной.

Из утверждения 5 следует, что если произвести перемножение логических скобок и в получившейся ДНФ сделать упрощения, пользуясь тем, что $x\bar{x}=0$, $xx = x$, $x \vee x = x$, то в результате получим ДНФ, состоящую из всех допустимых конъюнкций функции F_K. Теперь удалим допустимые конъюнкции, не являющиеся неприводимыми, пользуясь тождеством $x \vee xx' = x$. Получим ДНФ, состоящую из всех максимальных конъюнкций функции F_K (или сокращенную ДНФ). Для того, чтобы получить множество максимальных конъюнкций для f_K, нужно из построенного множества максимальных конъюнкций для F_K, отобрать те, которые являются допустимыми для f_K.

Для дискретного подхода наибольший интерес представляет случай, когда число признаков существенно превосходит число объектов. В этом случае число скобок будет мало по сравнению с числом переменных. И можно показать, что почти всегда (для почти всех КНФ рассматриваемого вида) число допустимых конъюнкций по порядку при $n \to \infty$ больше числа максимальных конъюнкций. Следовательно, алгоритм построения

максимальных конъюнкций, о котором было сказано ранее, не является эффективным. Теоретические и экспериментальные исследования показывают, что в рассматриваемом случае более разумно начинать с построения неприводимых конъюнкций функции F_K, а затем уже для каждой построенной неприводимой конъюнкции проверять, является ли она допустимой. В [5] был доказан следующий факт. Если $u \le n^{1-\varepsilon}$, $\varepsilon > 0$, то число неприводимых конъюнкций почти всегда (для почти всех КНФ рассматриваемого вида) асимптотически при $n \to \infty$ совпадает с числом максимальных конъюнкций функции F_K. На основе указанного факта в той же работе был построен алгоритм поиска максимальных конъюнкций функции F_K, являющийся в определенном смысле асимптотически оптимальным с позиции вычислительной сложности. Исходная задача построения всех максимальных конъюнкций функции F_K заменялась более простой задачей построения всех неприводимых конъюнкций функции F_K, т.е. задача решалась приближенно. Сложность приближенного решения оценивалась числом конъюнктивных умножений. Для случая, когда $u \le n^{1-\varepsilon}$, $\varepsilon > 0$, был предложен алгоритм поиска всех неприводимых конъюнкций функции F_K, для которого число конъюнктивных умножений почти всегда при $n \to \infty$ асимптотически совпадает с числом максимальных конъюнкций функции F_K. В данном алгоритме одно конъюнктивное умножение требует просмотра не более $O(un)$ переменных в заданной КНФ.

Таким образом, было показано, что если число нулей u функции F_K достаточно мало по сравнению с числом переменных n, то почти всегда удается построить ДНФ, содержащую все максимальные конъюнкции функции F_K и почти не отличающуюся от искомой ДНФ по сложности (длине), затратив на её построение в некотором смысле минимальное число операций “&”.

В [5, 6, 8] было показано, что приведенные результаты верны и для более общего случая, когда исходная КНФ имеет произвольный вид, т.е. не обязательно является совершенной, и при этом функция F_K, задаваемая КНФ, является двузначной функцией, определенной на k-ичных n-мерных наборах, $k \ge 2$. В указанных работах особо был выделен случай, когда исходная КНФ реализует монотонную булеву функцию, т.е. не содержит отрицаний переменных (этот случай имеет наибольшее значение для практики).

2.2. Построение элементарных классификаторов на основе поиска покрытий булевых и целочисленных матриц

При реализации алгоритмов поиска (тупиковых) представительных

наборов и (тупиковых) тестов чаще используются построения, в основе которых лежит поиск неприводимых покрытий булевых матриц.

Пусть L – булева матрица.

Определение. Набор столбцов H матрицы L назовем *покрытием*, если каждая строка матрицы L в пересечении хотя бы с одним из столбцов, входящих в H, даёт 1.

Определение. Покрытие назовем *неприводимым* (*тупиковым*), если никакое его собственное подмножество не является покрытием.

При построении неприводимых покрытий обычно используется следующий критерий. Набор столбцов H матрицы L является неприводимым покрытием тогда и только тогда, когда выполнены следующие два условия: 1) подматрица L^H матрицы L, образованная столбцами набора H, не содержит строки вида (0, 0, …, 0); 2) L^H содержит каждую из строк вида (1, 0, …, 0), (0, 1, …, 0), … , (0, 0, …, 1), т.е. L^H содержит единичную подматрицу.

Как возникает задача построения покрытий и неприводимых покрытий при построении (тупиковых) тестов и (тупиковых) представительных наборов? Для нахождения искомого множества элементарных классификаторов строится специальная булева матрица (матрица сравнения таблицы T). Обозначим её L_T. Каждая строка этой матрицы образуется в результате сравнения пары строк таблицы T, описывающих объекты из разных классов. При этом в столбце с номером j ставится 1, если сравниваемые строки различаются в разряде с номером j, и 0 в противном случае. Итак, если сравниваются объекты S_i и S_u и при этом $B(S_i, S_u, \{x_j\}) = 1$, то в строке матрицы L_T, определяемой парой (S_i, S_u) ставится 0, в противном случае 1. Через $L_T^{(i)}$ обозначим подматрицу матрицы L_T, которая образована сравнением обучающего объекта S_i со всеми обучающими объектами, не принадлежащими тому же классу, что и объект S_i.

Очевидными являются следующие два утверждения.

Утверждение 7. *Набор столбцов таблицы T с номерами $j_1, \ldots, j_r$ является (тупиковым) тестом тогда и только тогда, когда набор столбцов матрицы L_T с номерами $j_1, \ldots, j_r$ является (неприводимым) покрытием.*

Утверждение 8. *Элементарный классификатор $(S_i, \{x_{j_1}, \ldots, x_{j_r}\})$ из $N(K)$ является (тупиковым) представительным набором для K тогда и только тогда, когда набор столбцов матрицы $L_T^{(i)}$ с номерами $j_1, \ldots, j_r$ является (неприводимым) покрытием.*

Таким образом, при конструировании основных моделей дискретных процедур распознавания поиск информативных элементарных классификаторов сводится к построению покрытий булевых матриц.

Задача построения множества всех неприводимых покрытий булевой матрицы L размера $u \times n$ может быть сформулирована как задача преобразования КНФ монотонной булевой функции в сокращенную ДНФ. Действительно, строке с номером i поставим в соответствие дизъюнкцию $D_i = x_{p_1} \vee \ldots \vee x_{p_q}$, где $p_1, \ldots, p_q$ – номера тех столбцов, которые в пересечении с этой строкой дают 1. Пусть f – монотонная булева функция, реализуемая КНФ $D_1 \,\&\, \ldots \,\&\, D_u$.

Пользуясь утверждением 5 нетрудно доказать

Утверждение 9. *Э.к. $x_{j_1} \ldots x_{j_r}$ является допустимой для f тогда и только тогда, когда набор столбцов H матрицы L с номерами $j_1, \ldots, j_r$ является покрытием.*

Пользуясь утверждением 6 нетрудно доказать

Утверждение 10. *Э.к. $x_{j_1} \ldots x_{j_r}$ является неприводимой для f тогда и только тогда, когда набор столбцов матрицы L с номерами $j_1, \ldots, j_r$ содержит единичную подматрицу.*

Из утверждений 9 и 10 следует

Утверждение 11. *Э.к. $x_{j_1} \ldots x_{j_r}$ является максимальной для f тогда и только тогда, когда набор столбцов матрицы L с номерами $j_1, \ldots, j_r$ является неприводимым покрытием.*

Из последних трех утверждений следует, что алгоритмы построения максимальных конъюнкций монотонной булевой функции нетрудно модифицировать для построения неприводимых покрытий булевой матрицы. Модификация упомянутого выше асимптотически оптимального алгоритма построения максимальных конъюнкций монотонной булевой функции, заданной КНФ, приведена ниже в разд. 3.

Можно избежать построения вспомогательной булевой матрицы L_T, если ввести понятие покрытия более общего вида [4, 5, 7, 8].

Действительно, пусть L - целочисленная матрица размера $u \times n$ с элементами из $\{0,1,\ldots,k-1\}$, $k \geq 2$; E_k^r, $r \leq n$, - множество наборов вида $(\sigma_1,\ldots,\sigma_r)$, где $\sigma_i \in \{0,1,\ldots,k-1\}$.

Пусть далее $\sigma \in E_k^r$, $\sigma = (\sigma_1,\ldots,\sigma_r)$.

Определение. Набор столбцов H матрицы L назовем *σ-покрытием*, если в подматрице L^H матрицы L, образованной столбцами набора H, нет строки $(\sigma_1,\ldots,\sigma_r)$.

Определение. Набор столбцов H матрицы L, являющийся

σ-покрытием, назовем *тупиковым σ-покрытием*, если L^H содержит подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{pmatrix} \beta_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_{r-1} & \sigma_r \\ \sigma_1 & \beta_2 & \sigma_3 & \cdots & \sigma_{r-1} & \sigma_r \\ . & . & . & . & . & . \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_{r-1} & \beta_r \end{pmatrix},$$

где $\beta_p \neq \sigma_p$ при $p = 1, 2, \ldots, r$. Такую подматрицу будем называть *σ-подматрицей*.

Нетрудно видеть, что если $k = 2$ и $\sigma = (0, \ldots, 0)$, то понятие (тупикового) σ-покрытия совпадает с понятием (неприводимого) покрытия. Аналогом единичной подматрицы в этом случае является σ-подматрица.

Таблицу обучения T можно рассматривать как пару матриц L_1 и L_2, где L_1 - матрица, состоящая из описаний обучающих объектов из класса K, L_2 - матрица, состоящая из описаний остальных обучающих объектов. Тогда, очевидно, что элементарный классификатор вида $(\sigma_1, \ldots, \sigma_r)$ будет (тупиковым) представительным набором для K, порождаемым набором признаков $\{x_{j_1}, \ldots, x_{j_r}\}$, тогда и только тогда, когда набор столбцов матрицы L_1 с номерами $j_1, \ldots, j_r$ не является $(\sigma_1, \ldots, \sigma_r)$-покрытием, а набор столбцов матрицы L_2 с номерами $j_1, \ldots, j_r$ является (тупиковым) $(\sigma_1, \ldots, \sigma_r)$-покрытием.

Через $R(\sigma)$ обозначим множество наборов $(\beta_1, \ldots, \beta_r)$ в E_k^r таких, что $\beta_j \neq \sigma_j$ при $j = 1, 2, \ldots, r$.

Определение. Набор столбцов H матрицы L назовем *$R(\sigma)$-покрытием*, если в подматрице L^H матрицы L, образованной столбцами набора H, нет ни одной строки из $R(\sigma)$.

Определение. Набор столбцов H матрицы L, являющийся $R(\sigma)$-покрытием, назовем *тупиковым $R(\sigma)$-покрытием*, если L^H содержит подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{pmatrix} \sigma_1 & \beta_{12} & \beta_{13} & \cdots & \beta_{1r-1} & \beta_{1r} \\ \beta_{21} & \sigma_2 & \beta_{23} & \cdots & \beta_{2r-1} & \beta_{2r} \\ . & . & . & . & . & . \\ \beta_{r1} & \beta_{r2} & \beta_{r3} & \cdots & \beta_{rr-1} & \sigma_r \end{pmatrix},$$

где $\beta_{i\,p} \neq \sigma_p$ при $i, p = 1,2,\ldots,r$, $i \neq p$. Такую подматрицу будем называть $R(\sigma)$-подматрицей.

Нетрудно видеть, что если $k = 2$ и $\sigma = (1,\ldots,1)$, то понятие (тупикового) $R(\sigma)$-покрытия совпадает с понятием (неприводимого) покрытия. Аналогом единичной подматрицы является $R(\sigma)$-подматрица.

Алгоритмы построения неприводимых покрытий булевой матрицы естественным образом модифицируется на случай построения тупиковых σ-покрытий и тупиковых $R(\sigma)$-покрытий целочисленной матрицы.

Пусть L - булева матрица, $\sigma \in E_2^r$.

Через $C(L,\sigma)$ и $B(L,\sigma)$ обозначим соответственно совокупность всех σ-покрытий и всех тупиковых σ-покрытий матрицы L. Положим

$$C(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_2^r} C(L,\sigma),\ B(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_2^r} B(L,\sigma).$$

Связь между задачами построения множеств $C(L)$ и $B(L)$ булевой матрицы L размера $u \times n$ и задачей преобразования нормальных форм булевой функции устанавливается следующим образом.

Пусть $(\sigma_{i1},\ldots,\sigma_{in})$ - строка матрицы L с номером i, $i \in \{1,2,\ldots,u\}$. Этой строке ставится в соответствие дизъюнкция $\tilde{D}_i = x_1^{\bar{\sigma}_{i1}} \vee \ldots \vee x_n^{\bar{\sigma}_{in}}$. Пусть $\tilde{f}$ - булева функция, реализуемая КНФ $\tilde{D}_1 \,\&\ldots\&\, \tilde{D}_u$. Используя утверждения 5 и 6 нетрудно доказать приведенные ниже утверждения 12-14.

Утверждение 12. *Э.к.* $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ *является допустимой для* $\tilde{f}$ *тогда и только тогда, когда набор столбцов матрицы* L *с номерами* $j_1, \ldots, j_r$ *является* $(\sigma_1,\ldots,\sigma_r)$*-покрытием.*

Утверждение 13. *Э.к.* $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ *является неприводимой для* $\tilde{f}$ *тогда и только тогда, когда набор столбцов матрицы* L *с номерами* $j_1, \ldots, j_r$ *содержит* $(\sigma_1,\ldots,\sigma_r)$*-подматрицу.*

Утверждение 14. *Э.к.* $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ *является максимальной для* $\tilde{f}$ *тогда и только тогда, когда набор столбцов матрицы* L *с номерами* $j_1, \ldots, j_r$ *является тупиковым* $(\sigma_1,\ldots,\sigma_r)$*-покрытием.*

Если строке матрицы L с номером i, $i \in \{1,2,\ldots,u\}$, поставить в соответствие дизъюнкцию $\tilde{\tilde{D}}_i = x_1^{\sigma_{i1}} \vee \ldots \vee x_n^{\sigma_{in}}$ и рассмотреть булеву функцию $\tilde{\tilde{f}}$, реализуемую КНФ $\tilde{\tilde{D}}_1 \,\&\ldots\&\, \tilde{\tilde{D}}_u$, то задачи построения допустимых и максимальных конъюнкций функции $\tilde{\tilde{f}}$ могут быть

сформулированы соответственно как задачи построения $R(\sigma)$-покрытий и тупиковых $R(\sigma)$-покрытий матрицы L.

Упражнения

Задача 1. Пусть две строки булевой матрицы L имеют вид $(b_1,\ldots,b_n)$ и $(c_1,\ldots,c_n)$, где $c_j \geq b_j$ при $j=1,2,\ldots,n$. Будем говорить, что вторая строка охватывает первую. Показать, что при удалении охватывающих строк из L множество ее $(0,0,\ldots,0)$-покрытий не меняется.

Задача 2. Пусть функция F задана множеством $N_F=\{(1,1,1),(0,1,1),(1,0,1)\}$. Заполнить таблицу по образцу, указанному в первой строке.

Конъюнкция	Почти допустимая	Допустимая	Неприводимая	Максимальная
$x_1x_2x_3$	да	да	нет	нет
x_2x_3	?	?	?	?
$\bar{x}_2x_3$	?	?	?	?
$\bar{x}_1\bar{x}_2\bar{x}_3$	?	?	?	?
x_1	?	?	?	?

Для функции F построить неприводимую конъюнкцию ранга $r>1$, которая не являлась бы максимальной.

Задача 3. Заполнить таблицу

Конъюнкция			
почти допустимая	допустимая	неприводимая	максимальная
да	?	нет	?
да	?	да	?
?	да	да	?
?	?	да	нет
?	?	нет	нет

Задача 4. Пусть булева функция $F(x_1,x_2,x_3)$ задана КНФ $(x_1\vee x_2)\&(x_2\vee x_3)\&(x_1\vee x_3)$.

Построить сокращенную ДНФ функции F указанными ниже способами:

а) перемножением логических скобок;

б) сведением задачи к задаче построения неприводимых покрытий булевой матрицы.

Задача 5. Пусть булева функция $F(x_1, x_2, x_3)$ реализуется КНФ $(\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3)\&(\bar{x}_1 \vee x_2 \vee x_3)\&(x_1 \vee \bar{x}_2 \vee \bar{x}_3)\&(\bar{x}_1 \vee \bar{x}_2 \vee x_3)\&(x_1 \vee x_2 \vee \bar{x}_3)$. Построить сокращенную ДНФ функции F указанными ниже способами:

а) перемножением логических скобок;

б) сведением задачи к задаче построения тупиковых σ-покрытий булевой матрицы;

в) сведением задачи к задаче построения тупиковых $R(\sigma)$-покрытий булевой матрицы.

Задача 6. Пусть f - частичная булева функция. Показать, что задачу построения множества всех максимальных конъюнкций функции f можно решить на основе

а) построения неприводимых покрытий для ряда булевых матриц;

б) построения максимальных конъюнкций для ряда монотонных булевых функций.

3. Алгоритм построения неприводимых покрытий булевой матрицы в случае большого числа столбцов

Описываемый ниже алгоритм построения неприводимых покрытий булевой матрицы фактически основан на переборе всех ее единичных подматриц, что позволяет “быстро” обрабатывать матрицы, у которых большое число столбцов и относительно небольшое число строк.

Пусть $L = (a_{ij})$, $i = 1, 2, \ldots, u$, $j = 1, 2, \ldots, n$, - булева матрица.

Два различных единичных элемента $a_{i_1 j_1}$ и $a_{i_2 j_2}$ матрицы L назовем совместимыми, если $a_{i_1 j_2} = a_{i_2 j_1} = 0$. Набор Q из r единичных элементов матрицы L назовем совместимым, если выполнено одно из двух следующих условий:

1) $r = 1$;

2) $r > 1$ и любые два элемента в Q совместимы.

Обозначим через $S(L)$ совокупность всех совместимых наборов из единичных элементов матрицы L, через $P(L)$ – множество всех неприводимых покрытий матрицы L. Каждый набор из $S(L)$ определяет некоторую единичную подматрицу матрицы L.

Элементу a_{ij} в L присвоим номер $N[i, j] = (j-1)u + i$.

Пусть $R(L)$ – множество всех единичных элементов в L. Если $R \subseteq R(L)$, то через $e_1(R)$ и $e_2(R)$ будем обозначать соответственно элементы с наименьшим и наибольшим номерами в R.

Упорядочим $S(L)$. Для каждого Q из $S(L)$ укажем следующий за ним элемент $\succ Q$.

Пусть $Q = \{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$, где $N[i_{t+1}, j_{t+1}] > N[i_t, j_t]$ при $t = 1, 2, \ldots, r-1$.

Обозначим через R_t, $t \in \{1, 2, \ldots, r\}$, множество элементов в $R(L)$, номера которых больше $N[i_t, j_t]$.

Обозначим через G_t, $t \in \{1, 2, \ldots, r\}$, совокупность всех единичных элементов матрицы L, построенную следующим образом. Для каждого $v \in \{1, 2, \ldots, t\}$ из L вычеркнем строку с номером i_v и столбец с номером j_v, а также строки, дающие 1 в пересечении со столбцом с номером j_v, и столбцы, дающие 1 в пересечении со строкой с номером i_v. Тогда G_t - совокупность всех не вычеркнутых единичных элементов матрицы L. Число вычеркнутых строк обозначим через $\lambda_t(Q)$. Положим $\lambda(Q) = \lambda_r(Q)$.

Нетрудно видеть, что если $a_{ij} \in G_t$, то $\{a_{i_1 j_1}, \ldots, a_{i_t j_t}, a_{ij}\}$ – совместимый набор единичных элементов в L.

Если $G_r = \emptyset$, то Q является в некотором смысле максимальным

совместимым набором. В случае $\lambda(Q)=u$ набор столбцов с номерами $j_1, \ldots, j_r$ является неприводимым покрытием.

Перечислим возможные случаи и в каждом из них укажем $\circ Q$:

1) $G_r \cap R_r \neq \emptyset$; тогда $\succ Q = Q \cup \{e_1(G_r \cap R_r)\}$;

2) $G_r \cap R_r = \emptyset$;

а) $r = 1$; тогда $\succ Q = \{e_1(R_r)\}$;

б) $r > 1$ и $G_{r-1} \cap R_r \neq \emptyset$; тогда $\succ Q = (Q \setminus \{a_{i_r j_r}\}) \cup \{e_1(G_{r-1} \cap R_r)\}$;

в) $r > 1$ и $G_{r-1} \cap R_r = \emptyset$; тогда, если $r = 2$, то $\succ Q = \{e_1(R_1)\}$, и если $r > 2$, то $\succ Q = (Q \setminus \{a_{i_{r-1} j_{r-1}}, a_{i_r j_r}\}) \cup \{e_1(G_{r-2} \cap R_{r-1})\}$.

Заметим, что $G_{r-2} \cap R_{r-1} \neq \emptyset$ при $r > 2$, так как $a_{i_r j_r} \in G_{r-2} \cap R_{r-1}$.

Набор элементов $\{a_{i_1 j_1}, \ldots, a_{i_r j_r}\}$, принадлежащий $S(L)$, назовем верхним, если из того, что $\{a_{p_1 j_1}, \ldots, a_{p_r j_r}\} \in S(L)$ следует $i_t \leq p_t$ при $t = 1, 2, \ldots, r$.

Обозначим через $\widetilde{S}(L)$ множество всех верхних наборов в $S(L)$.

Алгоритм строит $P(L)$ за $|S(L)|$ шагов. На шаге i алгоритм выбирает в L совместимый набор $Q[i, L]$ и проверяет условие: $\lambda(Q[i, L]) = u$. Если это условие выполнено, то для исключения повторений при $i \geq 2$ проверяется еще условие: $Q[i, L] \in \widetilde{S}(L)$. Если и второе условие выполнено, то набор столбцов, в которых расположены элементы множества $Q[i, L]$, заносится в искомое множество неприводимых покрытий. В противном случае множество неприводимых покрытий, построенное на предыдущих шагах, не пополняется.

Выбор совместимых наборов происходит по следующим правилам:

1) $Q[1, L] = \{e_1(R(L))\}$;

2) если $Q[i, L] \neq \{e_2(R(L))\}$, то $Q[i+1, L] = \succ Q[i, L]$;

3) если $Q[i, L] = \{e_2(R(L))\}$, то алгоритм заканчивает работу.

Из описания работы алгоритма видно, что в его основе лежит процесс ветвления, который удобно представить в виде дерева.

Строится дерево решений $Д_L$, вершинам которого соответствуют элементы из $S(L)$. Исключение составляет корневая вершина, которая остается свободной. Висячим вершинам соответствуют максимальные совместимые наборы, и неприводимые покрытия порождаются теми из них, для которых выполнено условие: $\lambda(Q[i, L]) = u$. Формулируются правила, позволяющие при построении дерева $Д_L$ переходить от одной вершины к другой в таком порядке, который соответствует последовательному просмотру его ветвей. Переход от одной вершины к следующей линеен относительно размера матрицы L (требует просмотра

не более $3un$ элементов матрицы L). Число вершин в $Д_L$ равно $|S(L)|+1$.

Данный алгоритм опубликован в [3-6]. При обосновании его эффективности получены асимптотики типичных значений числа неприводимых покрытий и длины неприводимого покрытия (т.е. изучены метрические свойства множества неприводимых покрытий) для случая $u \le n^{1-\varepsilon}$, $\varepsilon > 0$. Показано, что в указанном случае при $n \to \infty$ величина $|P(L)|$ почти всегда (для почти всех матриц L размера $u \times n$) асимптотически совпадает с величиной $|S(L)|$. Отметим, что технические основы указанных оценок впервые были разработаны в работах В.А. Слепян и В.Н. Носкова при исследовании метрических свойств множества тупиковых тестов.

Асимптотически оптимальный алгоритм построения неприводимых покрытий булевой матрицы можно модифицировать на случай построения тупиковых σ-покрытий булевой и целочисленной матрицы. При обосновании асимптотической оптимальности данного алгоритма были получены в [3-6] асимптотики типичных значений числа тупиковых σ-покрытий и длины тупикового σ-покрытия.

Упражнения

Задача 1. Модифицировать описанный в данном разделе алгоритм для построения множества

$$B(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} B(L,\sigma)$$

в случае, когда матрица L является

а) булевой матрицей ($k = 2$);

б) целочисленной матрицей с элементами из $\{0,1,\ldots,k-1\}$, $k > 2$.

Задача 2. Решить задачу, аналогичную задаче 1, для случая $R(\sigma)$-покрытий.

Задача 3. Модифицировать описанный в данном разделе алгоритм для решения задачи преобразования КНФ, не содержащей отрицаний переменных (реализующей монотонную булеву функцию), в ДНФ.

Литература

1. Баскакова Л.В., Журавлёв Ю.И. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств//Ж. вычисл. матем. и матем. физ. 1981. Т. 21. №5. С. 1264-1275.
2. Дмитриев А.И., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов или явлений // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1966. Вып. 7. С. 3-17.
3. Дюкова Е.В. Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1977. 233. № 4. С. 527-530.
4. Дюкова Е.В. Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Пробл. кибернетики. М.: Наука, 1982. Вып. 39. С. 165-199.
5. Дюкова Е.В. О сложности реализации некоторых процедур распознавания // Ж. вычисл. матем. и матем. физ. 1987. Т27, №1. С.114-127.
6. Дюкова Е.В. Алгоритмы распознавания типа Кора: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1989. Вып. 2. С. 99-125.
7. Дюкова Е.В. Асимптотические оценки некоторых характеристик множества представительных наборов и задача об устойчивости //Ж. вычисл. матем. и матем. физ. 1995. Т. 35. № 1. С. 122-134.
8. Дюкова Е.В., Журавлёв Ю.И. Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. 2000. Т. 40. №8. С. 1264-1278.
9. Дюкова Е.В., Инякин А.С. Задача таксономии и тупиковые покрытия целочисленной матрицы // Сообщения по прикладной математике. М.: ВЦ РАН, 2001. 28с.
10. Дюкова Е.В., Песков Н.В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и матем. физ. 2002. Том 42, № 5, С. 741-753.
11. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. М.: Наука, 1978. Вып. 33. С. 5-68.
12. Журавлев Ю.И. Об алгоритмах распознавания с представительными наборами (о логических алгоритмах) // Ж. вычисл. матем. и матем. физ. 2002. Т. 42. № 9. С. 1425-1435.
13. Кузнецов В.Е. Об одном стохастическом алгоритме вычисления информативных характеристик таблиц по методу тестов // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1973. Вып. 23. С. 8-23.
14. Матросов В.Л. Синтез оптимальных алгоритмов в алгебраических

замыканиях моделей алгоритмов распознавания // Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1988. Вып. 1. С. 149-176.
15. Матросов В.Л., Стеценко В.А. Лекции по дискретной математике. Учебное пособие для магистрантов математических факультетов педагогических университетов. М.: Прометей, 1997, 220 с.
16. Рудаков К.В. Об алгебраической теории универсальных и локальных ограничений для задач классификации // Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1988. Вып. 1. С. 176-200.
17. Рязанов В.В. О построении оптимальных алгоритмов распознавания и таксономии (классификации) при решении прикладных задач // Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1988. Вып. 1. С. 229-279.
18. Чегис И.А., Яблонский С.В. Логические способы контроля электрических схем // Тр. МИАН СССР, М., 1958.
19. Яблонский С.В. Введение в дискретную математику // М.: Наука, 1986. 384 с.
20. Djukova E.V., Peskov N.V. Selection of Typical Objects in Classes for Recognition Problems // J. Pattern Recognition and Image Analysis. 2002. V. 12. No. 3. P. 243-249.
21. Дискретная математика и математические вопросы кибернетики. Т. 1. / Под ред. С.В. Яблонского и О.Б. Лупанова. М.: Наука, 1974. 312 с.

Темы курсовых и дипломных работ

1. Разработка и реализация на ЭВМ алгоритмов построения неприводимых покрытий булевой матрицы. Тестирование алгоритмов на случайных матрицах.
2. Разработка и реализация на ЭВМ алгоритмов построения тупиковых покрытий целочисленной матрицы. Тестирование алгоритмов на случайных матрицах.
3. Разработка и реализация на ЭВМ алгоритмов построения минимальных покрытий булевой матрицы. Тестирование алгоритмов на случайных матрицах.
4. Решение прикладной задачи распознавания алгоритмами голосования: а) по тестам; б) по представительным наборам. Проведение сравнительного анализа.
5. Решение прикладной задачи распознавания алгоритмами голосования: а) по множеству антипредставительных наборов; б) по покрытиям класса. Проведение сравнительного анализа.

Оглавление